The Landscape and Perspectives of the Human Gut Metaproteomics

Authors

Zhongzhi Sun, Zhibin Ning, and Daniel Figeys

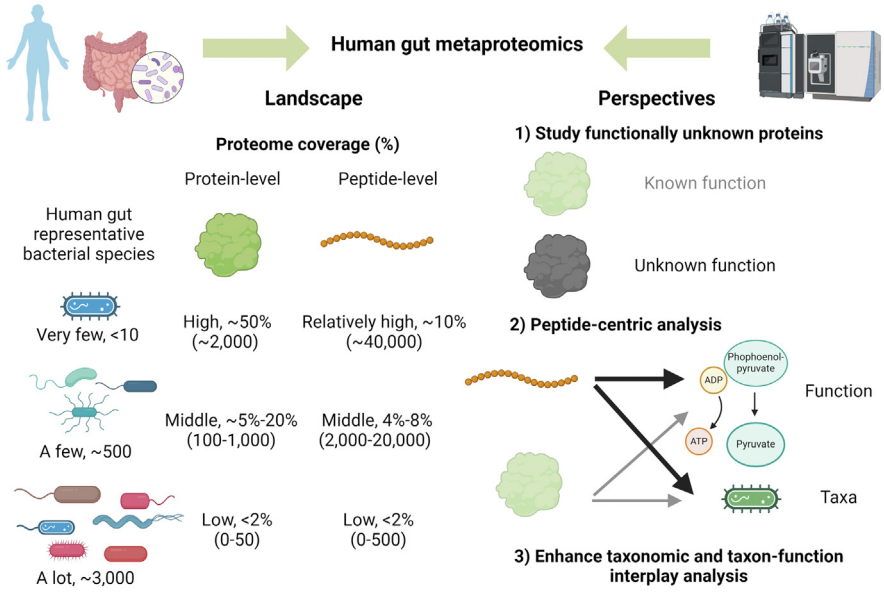
Correspondence

dfigeys@uottawa.ca

In Brief

Metaproteomics emerges as a valuable tool for studying the human gut microbiome. This article explores proteome and protein annotation coverage in published human gut metaproteomics datasets. It emphasizes the importance of improving proteome coverage and enhancing functional annotation. The analysis advocates the adoption of peptide-centric analysis and underscores the necessity to improve the integration of taxonomic and functional data in metaproteomic research.

Graphical Abstract



Highlights

- Evaluate the proteome and protein annotation coverage in human gut metaproteomics.
- Compare the taxonomy of metaproteomics datasets with the metagenomics database.
- Advocate the use of metaproteomics for studying functionally unknown proteins.
- Assess the feasibility of applying peptide-centric analysis in metaproteomics.
- Propose refining the taxon scope and conducting taxon-function cross-analysis.

2024, Mol Cell Proteomics *23*(*5*), 100763

https://doi.org/10.1016/j.mcpro.2024.100763

The Landscape and Perspectives of the Human Gut Metaproteomics

Zhongzhi Sun[1,2], Zhibin Ning[1], and Daniel Figeys[1,2,*]

The human gut microbiome is closely associated with human health and diseases. Metaproteomics has emerged as a valuable tool for studying the functionality of the gut microbiome by analyzing the entire proteins present in microbial communities. Recent advancements in liquid chromatography and tandem mass spectrometry (LC-MS/MS) techniques have expanded the detection range of metaproteomics. However, the overall coverage of the proteome in metaproteomics is still limited. While metagenomics studies have revealed substantial microbial diversity and functional potential of the human gut microbiome, few studies have summarized and studied the human gut microbiome landscape revealed with metaproteomics. In this article, we present the current landscape of human gut metaproteomics studies by reanalyzing the identification results from 15 published studies. We quantified the limited proteome coverage in metaproteomics and revealed a high proportion of annotation coverage of metaproteomics-identified proteins. We conducted a preliminary comparison between the metaproteomics view and the metagenomics view of the human gut microbiome, identifying key areas of consistency and divergence. Based on the current landscape of human gut metaproteomics, we discuss the feasibility of using metaproteomics to study functionally unknown proteins and propose a whole workflow peptide-centric analysis. Additionally, we suggest enhancing metaproteomics analysis by refining taxonomic classification and calculating confidence scores, as well as developing tools for analyzing the interaction between taxonomy and function.

Applying Metaproteomics to Study the Human Gut Microbiome

The human gut microbiome has complex interactions with the host, and the taxonomic composition and functional activity of the gut microbiome have been closely associated with human health and diseases (1). Proteins are the foundations of most biological processes, comprising about 50% of the dry mass of a cell across different species (2). Their quantification is instrumental in assessing the biomass contributions of different bacterial species within a community (3). Consequently, a comprehensive analysis of the entire protein complement in the microbiome is critical for unraveling host-microbiome interactions.

Metaproteomics was first proposed by Wilmes and Bond in 2004, defined as the "Large-scale characterization of the entire protein complement of environmental microbiota at a given point in time" (4). Initially, metaproteomics involved separating proteins on a 2D gel and manually selecting individual protein spots for mass spectrometric analyses, which was experimentally demanding and low throughput (5). During this early stage of metaproteomics, approximately 2000 proteins could be detected in a microbial community (6). Although it is still in its early stages, the liquid chromatography and tandem mass spectrometry (LC-MS/MS)-based bottom-up metaproteomics is now able to detect approximately 50,000 ~ 70,000 protein groups in a single study using different techniques (7, 8), showcasing the rapid technological advancement over the past 2 decades.

The Advantage of Metaproteomics Compared to Other Omics

Metaproteomics measures the presence and abundance of proteins, thereby revealing gene expression dynamics within microbial communities (9). Additionally, by assigning proteins to individual species or higher taxa, metaproteomics offers insights into the taxonomic composition of the microbiota (9, 10). However, metaproteomics encompasses much more than measuring gene expression and species biomass within microbial communities (10). When compared with other high-throughput omics techniques—including amplicon sequencing, metagenomics, metatranscriptomics, and metabolomics—metaproteomics has several unique advantages.

Firstly, metaproteomics is more likely to reveal the real functionality of the microbial community. This contrasts with DNA-based metagenomics and metatranscriptomics, which only suggest the potential functional capabilities of microbial communities. The reason for this disparity lies in the fact that

From the [1]School of Pharmaceutical Sciences, and [2]Department of Biochemistry, Microbiology and Immunology, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada
*For correspondence: Daniel Figeys, dfigeys@uottawa.ca.

https://doi.org/10.1016/j.mcpro.2024.100763

not all DNA (genes) are transcribed into RNA, and not all RNA transcripts are subsequently translated into proteins. Secondly, metaproteomics enables the study of post-translation modifications (PTMs) in microbes. PTMs play a crucial role in modulating protein activity, structural conformation, and interactions, significantly influencing bacterial behavior within the microbiome and in interactions with the host (11). Such modifications, critical for understanding complex biological processes, remain elusive in other high-throughput omics analyses.

In addition, metaproteomics has advantages for studying host-microbe interactions, as both human proteins and microbial proteins are able to be identified and quantified. Human proteins, which constitute approximately 15% of the biomass in metaproteomic samples (12) are pivotal in mediating these interactions. Moreover, metaproteomics can also analyze isotope content, which helps determine carbon sources and provides a deeper understanding of metabolism in microbial communities (13). Overall, metaproteomics distinctively enhances our comprehension of the human gut microbiome, offering insights that are not readily obtainable through other omics technologies.

Discoveries Achieved Through Metaproteomics

Metaproteomics has emerged as a powerful tool for unraveling the pathogenesis of various diseases and for identifying potential biomarkers. Its application ranges from unraveling the microbial contribution to oxidative stress in inflammatory bowel disease (14) to uncovering the interplay between gut microbiota and the development of type 1 diabetes (15). Metaproteomics has also been applied to exploring host-microbiome interactions underlying other diseases, such as cancer (16, 17), obesity (18, 19), and COVID-19 (20, 21). A comprehensive overview of the clinical applications of metaproteomics has been provided by Wolf *et al.* (22).

Despite these strides, metaproteomics confronts ongoing challenges. While numerous reviews have offered comprehensive insights into research methods (9, 23), accomplishments (22, 24), challenges (25, 26), and future directions (12, 27), few studies offer guidance for advancing metaproteomics based on data from previous studies. Although the protein landscape of human gut bacterial species identified in metaproteomics has been preliminarily explored (28), certain important areas, such as proteome coverage demand further exploration.

In this article, we systematically reanalyzed the identified peptides and proteins from 15 human gut metaproteomics studies compiled in the MetaPep (29) database. Notably, all raw files in our dataset utilized HCD-FTMS for MS2 scans, offering superior resolution and accuracy. All the peptide and protein identification results were acquired from MetaLab-MAG (30), a user-friendly publicly accessible metaproteomics data analysis platform, ensuring the reproducibility of our findings and paving the way for integrating additional datasets in subsequent research.

CURRENT LANDSCAPE OF HUMAN GUT METAPROTEOMICS

Proteome Coverage in Metaproteomics

In the metaproteomics community, it is widely acknowledged that there is still room for enhancing the depth of proteome coverage (31–33). Although ultra-deep proteomics methods have shown promise in detecting nearly the entire proteome of single-species bacteria, identifying up to 75 to 77% of open-reading frames (34, 35), the issue of limited proteome coverage becomes more significant when dealing with complex microbial communities in metaproteomics. Previous research indicated that increased species diversity reduced the number of identified protein groups, and with the current state of metaproteomics technology, the estimated species and proteome coverage in a complex sample containing around 300 bacterial species is about 20% and 5%, respectively (31). From a more intuitive standpoint, the current stage ultra-deep metaproteomics techniques detected an average of 69,280 peptides, and 30,686 protein groups per microbiome sample (7), these numbers represent a mere 2.34% of the theoretical microbiome proteome, based on an estimated 1,310,000 coding sequences (CDS) (12). These studies suggest that a significant portion of proteins and species remain undetected in metaproteomic analyses, aligning with our previous observations that bacteria with less than 0.5% biomass are difficult to detect using current metaproteomics workflows (36).

While full quantification of all proteins is not necessary to study responses in microbiome networks (37), achieving more comprehensive proteome coverage remains crucial, given its current limitations. To date, total coverage of metaproteomics across human gut bacterial species and functions has not been thoroughly investigated. Recently, MetaPep (29) compiled identified peptides from over 2000 human gut metaproteomics raw files from 15 published studies, allowing us to comprehensively evaluate the extent of proteome coverage and the scope of detectable proteins within the metaproteomics field. These peptides were identified by searching raw files from each study with MetaLab-MAG. During the search, MetaLab-MAG integrated a human proteome fasta file from UniProt into the search space. Carbamidomethyl[C] was set as a fixed modification, with Oxidation[M] and Acetyl [ProteinN-term] as variable modifications. All other parameters were kept at their default settings. To refine our focus on microbial peptides, peptides exclusively found in the human proteome were excluded. Out of the compiled 1,163,940 peptides in MetaPep, only 2837 were found in both bacterial and human proteomes.

Peptide-Level Proteome Coverage—We first directly investigate the proteome coverage at the peptide level. To be specific, peptides identified in human gut metaproteomic

studies were mapped to the species or lowest common ancestor (LCA) of human gut bacteria from the UHGG (Unified Human Gastrointestinal Genome) dataset as described in the original MetaPep publication (29). These peptides were further mapped to the phylogenetic tree of representative bacterial species from the UHGG dataset (38) and visualized with iTOL (39). Our analysis revealed several interesting patterns.

First, we noted an uneven phylogenetic distribution of identified peptides across human gut bacteria (Fig. 1). Out of the total 1,163,940 peptides from MetaPep, 293,181 (25.2%) could be assigned to 4110 of the 4744 representative prokaryotic species (4716 bacterial and 28 archaeal) in the UHGG database. These peptides that were found exclusively in one prokaryotic species are referred to as genome-distinct peptides. In contrast to the high proportion (84.5%) of genome-distinct peptides among all in-silico digested peptides of UHGG representative genomes (29), the ratio of genome-distinct peptides in MetaPep is much lower. This difference is reasonable because genome-distinct peptides, unlike peptides shared by multiple species, are expected to have a lower

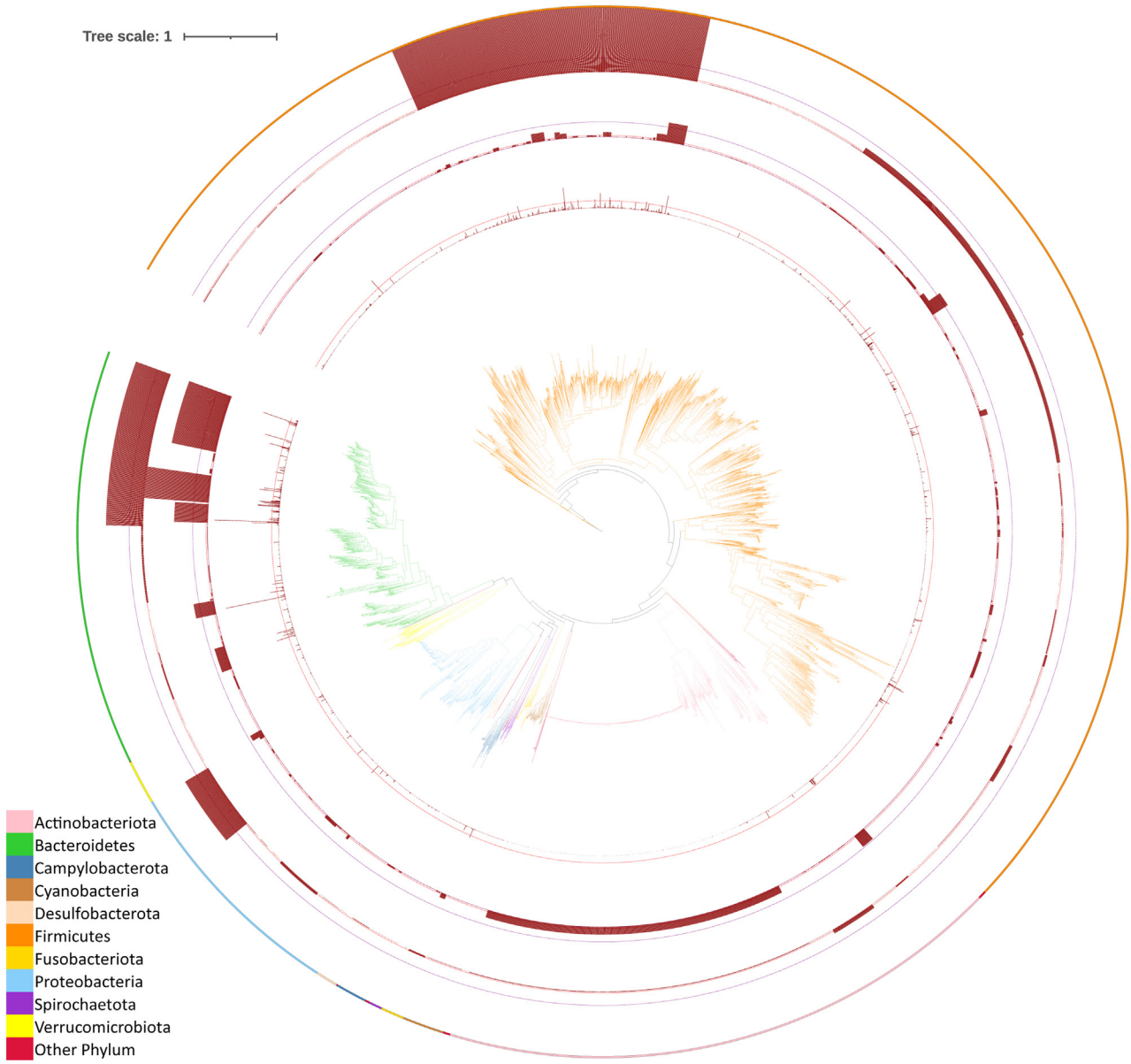


FIG. 1. **Phylogenetic distribution of identified peptides in metaproteomics studies.** The innermost is the phylogenetic tree of 4716 bacterial species extracted from the UHGG dataset. From the inner to the outer circle, 3 bar plots show the number of genome-distinct peptides, peptides assigned to genus-level lowest common ancestors (LCA), and peptides assigned to family-level LCA for each node/clade. Dashed lines serve as references for the number of peptides: *red dash line* (1000 peptides), *purple dash line* (10,000 peptides). Phylum-level taxonomy is indicated by *different colors* in both the phylogenetic tree and the color strip on the outermost ring.

abundance in real microbial samples. This lower abundance makes it more challenging to detect genome-distinct peptides, resulting in fewer of them being collected in MetaPep. Focusing on the analysis of MetaPep compiled peptides, the number of genome-distinct peptides of each bacterial species varied from 0 to 9093. Peptides assigned to higher taxonomic levels, such as genus and family, were also incorporated into the phylogenetic mapping (Fig. 1). For peptides assigned to genus level LCA, five genera (*Bacteroides*, *Prevotella*, *Phocaeicola*, *Parabacteroides*, *Blautia_A*, and *Faecalibacterium*) have >10,000 peptides. For peptides assigned to family level LCA, three families (Lachnospiraceae, Bacteroidaceae, and Enterobacteriaceae) have >10,000 peptides. Species from these taxonomic units also had a larger number of genome-specific peptides (Fig. 1), indicating these bacteria were frequently identified from human gut metaproteomics. While it is essential to consider these taxonomic units with numerous identified peptides, it is equally important to note that many bacterial species (2198) were represented by only a limited number of genome-distinct peptides (≤5), including 646 species that had no identifiable genome-distinct peptides. These 2198 species accounted for 46.6% of the 4716 representative bacterial species in the UHGG database. At the family level, the most represented sources of these species were Coriobacteriaceae (566), Acutalibacteraceae (104), Oscillospiraceae (99), UBA660 (91), and CAG-508 (40). At the genus level, the most represented were *Collinsella* (564), *Streptococcus* (41), *Veillonella* (28), *CAG-1427* (25), and *CAG-269* (23). There was no overlap between the taxa with the highest number of identified peptides and those with the most species lacking identifiable peptides, highlighting the uneven phylogenetic distribution of identified peptides across species and indicating that some bacterial taxa are rarely detected in the compiled metaproteomics datasets.

Second, we observed that the number of identified peptides for each bacterial species in metaproteomics constitutes only a small fraction of that species' entire proteome. *Phocaeicola dorei*, for example, has the highest number of genome-distinct peptides (9093) in MetaPep, with 60 additional species also having more than 1000 genome-distinct peptides each (Fig. 1). Focusing on *P. dorei*, 38,293 peptides from MetaPep could be assigned to this bacteria species, as more peptides assigned to higher taxonomic levels could also be found in this species. However, these 38,293 peptides only account for 9.71% of all 394,374 *in silico*-digested peptides of this species (in-silico digestion parameters: minimum peptide length = 7, maximum missed cleavage = 2). For the other 60 species with over 1000 genome-distinct peptides, an average of 12,748 peptides from MetaPep corresponded to 1.61 to 9.58% of each species' in-silico digested peptides. This indicates that even for species with the most identified peptides, only a small proportion of their theoretical proteome is detected. The number of identified peptides for all bacterial species is shown in Supplemental Table S1.

Protein-Level Proteome Coverage—We also investigated proteome coverage at the protein level by re-examining protein identification results from metaproteomics studies collected in MetaPep (29). In total, 306,413 unique head proteins were extracted from all identified protein groups across 15 studies. The head protein, which is the first protein listed in a protein group identified by MetaLab-MAG (30), shares identified peptides with all other proteins in the group. The overall landscape of metaproteomics studies, as revealed by analyzing these head proteins, was similar to the findings at the peptide level.

First, similar to the peptide-level analysis, an uneven phylogenetic distribution was observed among the identified proteins. The head proteins from identified protein groups originated from 4288 of the 4744 representative prokaryotic species in the UHGG, and the number of identified proteins per species varied widely, from 0 to 2282 (Fig. 2). Consistent with peptide-level results, in protein-level analysis, the same five genera (*Bacteroides*, *Prevotella*, *Phocaeicola*, *Parabacteroides*, *Blautia_A*) had the most identified head proteins, exceeding 10,000 each. At the family level, the same three families (Lachnospiraceae, Bacteroidaceae, Ruminococcaceae) had the most identified head proteins, with two other families Enterobacteriaceae and Oscillospiraceae, also having more than 10,000 identified head proteins. In parallel with taxa with numerous identified proteins, a significant number of bacterial species and taxa were represented by only a limited number of identified proteins (Fig. 2).

Second, the number of identified proteins covered a limited portion of the proteome of bacterial species. Similar to the peptide-level coverage results, at the protein level, *P. dorei* also has the largest number of identified head proteins (2282), covering 50.5% of the species proteome (all 4522 CDS). However, the majority of species (3655) exhibited limited proteome coverage (<5%), with fewer species showing higher coverage (199 genomes with 10 to 20% coverage, 83 genomes with 20 to 50% coverage, and only two genomes with ≥50% coverage). The number of identified head proteins of each bacterial species is also shown in Supplemental Table S1. While it is hard to estimate how many in-silico digested peptides are present in the sample (40), it is well-established that the majority of bacterial coding sequences are indeed expressed as proteins (41, 42). Compared to peptide-level analysis, analyzing at the protein level provides a more intuitive understanding of proteome coverage.

Comparison of Taxonomic Composition Between Bacterial Species Identified in Metaproteomics Studies and the Species in the Metagenomics Reference Database

In addition to only focusing on the metaproteomics dataset, we further compared the taxonomic composition of bacterial species identified in metaproteomics studies and the species in the metagenomics reference database. We extracted the taxonomic composition at various taxonomic ranks from the

FIG. 2. **Phylogenetic distribution of identified proteins in metaproteomics studies.** The innermost is the phylogenetic tree of 4716 bacterial species extracted from the UHGG dataset. The bar plot shows the number of identified proteins from each bacterial species. *Dashed lines* serve as references for the number of proteins: *red dash line* (100 proteins), *purple dash line* (1000 proteins). Phylum-level taxonomy is indicated by *different colors* in both the phylogenetic tree and the color strip on the outermost ring.

metagenomics reference database, UHGG (38), as presented in Figure 3*A*. On the other hand, the number of identified peptides collected in the MetaPep and their taxonomic sources at different taxonomic ranks are shown in Figure 3*B*.

At higher taxonomic ranks, the overarching taxonomic composition of the bacterial species identified in metaproteomic analysis aligns with the composition of representative genomes derived from the UHGG (Fig. 3, *A* and *B*). At the phylum level, the predominant groups—Firmicutes_A, Bacteroidota, Proteobacteria, Firmicutes, and Actinobacteriota—were not only the most genome-rich in the UHGG but also yielded the highest number of identified peptides in MetaPep. However, their relative abundances vary between the metagenomic reference database and metaproteomic datasets. For instance, Bacteroidota comprises approximately 15% of the UHGG's genomes but contributes around 34.6% of MetaPep's peptides. Conversely, Actinobacteriota accounts for about 18% of UHGG genomes, but only3.7% of peptides in MetaPep. The top taxa at the class and order levels were generally consistent between the two datasets, with minor exceptions (Fig. 3, *A* and *B*).

At the family and genus levels, however, this congruity wanes, with only two taxa shared across the five most represented taxa in the metagenomics reference database and

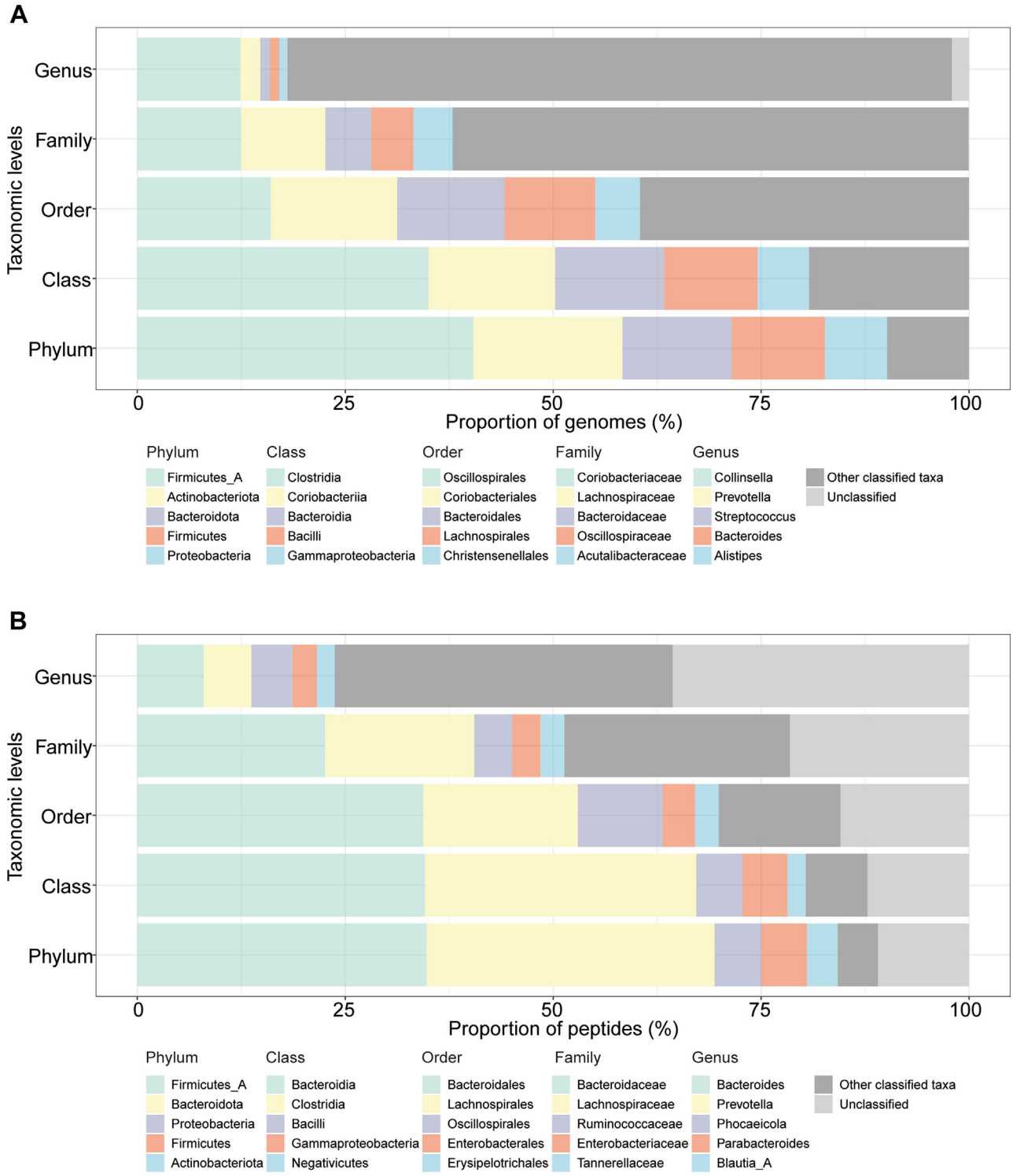


FIG. 3. **Comparison of taxonomic composition of bacterial species identified in metaproteomic studies and the species in the metagenomics reference database, UHGG.** *A*, taxonomic affiliation of the bacterial species collected in the UHGG at different taxonomic ranks. *B*, taxonomic affiliation of the metaproteomics-identified peptides (peptides compiled in MetaPep) at different taxonomic ranks. The legend only depicts the five most highly represented taxa per rank.

metaproteomics datasets (Family level: Bacteroidaceae and Lachnospiraceae; Genus level: *Bacteroides* and *Prevotella*). Notable discrepancies include *Collinsella*, which, despite possessing the highest number of genomes at the genus level in the UHGG database (584, representing 12.3%), accounts for a mere 0.6% of peptides (5745) in MetaPep. Similarly,

Streptococcus, while being the third most genome-abundant genus (54, or 1.1% of UHGG genomes), corresponds to only 0.3% of MetaPep's peptides. In contrast, *Parabacteroides*, with only 27 genomes in the UHGG, constitutes 2.9% of the peptides (27,279) in MetaPep. The differences in the taxonomic composition at the family and genus levels were also observed when considering more representative taxa units. Among the 30 families and 30 genera with the highest number of identified peptides in metaproteomics studies, 17 families and 13 genera overlapped with taxa units that had the highest number of genomes in the UHGG.

In summary, while there are consistent representations of higher taxonomic ranks (phylum, class, and order) across both the metagenomic reference database and metaproteomic datasets, substantiating their abundance in the human gut microbiome, discrepancies at lower taxonomic ranks highlight the unique insights afforded by metaproteomic analysis. These differences underscore the value of metaproteomics in providing a complementary angle for analyzing the structure of microbial communities by assessing species biomass (3).

Metaproteomics View of Unannotated Proteins

Importance of Protein Functional Annotation—Protein functional annotation is fundamental for the understanding of biological processes, unfortunately, many genes from the microbiome are not or are poorly annotated and we need new approaches to address this challenge. The advent of sequencing technologies has resulted in the discovery of a plethora of new gene sequences, while, the experimental characterization of their protein products and function remains unaddressed, leading to a widening sequence-to-function gap (43). Even for well-studied organisms like *Homo sapiens* and *Escherichia coli*, 10% (44) and 7.2% (45) of proteins, respectively, remain unannotated. The challenge is more pronounced in the complex human gut microbiota, where many species lack cultured representatives, and a larger fraction of microbial proteins lack functional annotation. The UHGG database highlights this issue, revealing that 27.3% of genes do not match any functional database, and an additional 14.2% of genes match COG (Cluster of Orthologous Groups) categories with unknown functions (38), resulting in a total of 41.5% of genes poorly annotated. It is likely that considering ~50,000 metabolites potentially produced by the gut microbiome (46), some proteins are moonlighting which might not be captured in their annotation (47, 48).

Impact of Unannotated Proteins on Metaproteomics—The unannotated proteins have two significant impacts on metaproteomics studies. The first impact is that most metaproteomics analyses rely on protein database searches for peptide and protein identification. However, existing prokaryotic gene prediction tools accurately detect only 60 to 70% of start codons for specific bacterial species (49), leading to potential misannotations or omissions. This directly impacts the results of sequence database searches. The second impact is on the interpretation of metaproteomics data. Similar to other high-throughput omics studies, most metaproteomics studies start with pathway or network analysis to identify biological processes that significantly changed, where functional annotations of proteins are crucial. Annotation biases can result in a skewed understanding, focusing on a limited set of annotated proteins while overlooking the functions of many others (50). Indeed, some unannotated proteins have been implicated in disease development and may have essential roles (51, 52).

Unannotated Proteins in Genomics and Metaproteomics Datasets—Annotation coverage of different bacterial species has been systematically investigated in genomics data, with an average of 52 to 79% of the coding sequences in bacterial genomes could be functionally annotated, and the annotation coverage ranging from 14% in some species to 98% in others (53). This disparity suggests that taxonomy is a major factor influencing annotation completeness. However, the annotation coverage of proteins identified in metaproteomics studies has not been systemically investigated. Here, we revisited the annotations of 306,413 identified head proteins, sourced from studies collected in MetaPep (29). We discovered that 78,474 (25.6%) of these proteins have neither KEGG ko (54) nor Gene Ontology (GO) annotation (55, 56). Among these proteins, 28,100 (9.17%) proteins either had no COG category annotation or were assigned to COG category S (Function unknown), indicating limited functional information available for these proteins.

Interestingly, a higher proportion of proteins identified in metaproteomics have functional annotations compared to the overall gene datasets in UHGG. While 41.5% of genes in UHGG are poorly annotated, only 9.17% of metaproteomics-identified proteins lack annotations. Notably, among the 28,100 unannotated metaproteomics proteins, 2342 were highly abundant (within the top 10% of total intensity in specific studies), suggesting that these functionally unknown proteins could play significant biological roles. This underlines the potential of metaproteomics in investigating the functions of these enigmatic proteins.

When we analyze the ratio of annotated proteins among those identified from each bacterial species, we observe that the majority of identified proteins from different species already have functional annotations (see Supplemental Table S1). For species with at least 100 identified proteins, the percentage of proteins with annotations ranges from 79.0 to 100% of all the identified proteins. Notably, *Bacteroides* sp900765785 and *Bacteroides fragilis* have the lowest proportions of annotated proteins among the identified proteins, at 79.0% and 79.9%, respectively.

Peptide-Centric Analysis

Challenges of Protein Inference and the Advantages of Peptide-Centric Analysis—Protein inference poses another challenge in metaproteomics. Within a complex microbial

community, a single peptide could be shared by hundreds of different proteins, complicating the precise attribution of the peptide to its parent protein. As a result, in metaproteomics, protein inference usually does not yield a list of proteins, but rather a list of protein groups. These protein groups can encompass proteins from various bacterial species with different functions, resulting in the loss of valuable information during protein inference (57). Moreover, different protein inference algorithms can produce protein group lists with substantial differences (58, 59). All these factors in protein inference further impact subsequent taxonomic and functional analysis.

Since mass spectrometry intrinsically identifies peptides, not proteins, a peptide-centric approach for functional and taxonomic analysis emerges as a logical alternative in metaproteomics. Tools such as Unipept (60, 61), PepFunk (62), and MetaGOmics (63) facilitate this approach, linking peptides directly to their functional and taxonomic attributes. This method bypasses the protein inference step, building the microbial community profile based on peptide identifications and quantifications. Research indicates that peptide-centric analysis can offer enhanced sensitivity and uncover details that protein-level analysis might miss (64, 65). Moreover, our recent work on MetaPep (29) also substantiated the feasibility of initiating peptide-centric analysis by searching a peptide sequence database. Notably, this database is significantly smaller than its protein counterpart, offering a considerable advantage in terms of reducing search times throughout the peptide-centric analysis workflow.

Feasibility of Peptide-Centric Analysis—The feasibility of peptide-centric taxonomic analysis was demonstrated with the introduction of the first tryptic peptide-based metaproteomics biodiversity analysis method, Unipept (60), in 2012. Since then, peptide-centric taxonomic analysis has gained widespread adoption in metaproteomics (15, 16, 66), with subsequent developments in Unipept enabling functional analysis *via* GO terms and EC numbers (61). However, concerns have been raised about the limited sequence length of peptides and their capacity to convey meaningful functional information (67, 68). A proposed solution involves tailoring the protein sequence collection for in-silico digestion, ensuring that each peptide correlates to a protein with a specific function. Customized or research-specific databases for peptide digestion have been shown to enhance functional resolution in peptide-centric metaproteomics analyses (69).

To verify the feasibility of peptide-centric function analysis for human gut metaproteomics studies, we performed a preliminary test on the UHGG database for annotating in-silico digested peptides from the database to specific functions. Out of the 10,234,935 proteins from 4744 representative prokaryotic species within the UHGG database, 8,277,932 (80.9%) of them had specific COG family functional annotations. We collected proteins from each COG family, along with proteins lacking COG annotations, for in-silico digestion. This process yielded a total of 392,459,520 peptides, with 385,535,220 being unique peptide sequences. After digestion, 325,260,860 (84.3%) of these unique peptides were exclusively found in proteins from specific COG families and 55,404,896 (14.4%) were exclusive to proteins without COG family annotations. And these peptides were referred as to functional-distinct peptides. The remaining 4,869,464 unique peptides were shared among proteins from multiple COG families, accounting for only 1.3% of the total unique peptides. These findings suggest that a substantial majority of identifiable peptides can be confidently associated with specific functional categories.

PERSPECTIVES OF THE HUMAN GUT METAPROTEOMICS

Improving the Coverage of Human Gut Metaproteomics

Scaling up Metaproteomic Studies—The limitation in metaproteomics coverage was not only confined to individual studies but was also reflected in the overall limited number of metaproteomics studies. Despite an uptick in the quantity of metaproteomic research, there remains a stark contrast when compared to metagenomics. According to the search results from the *Web of Science*, during the past 10 years (2013–2022), 12,298 papers with topics in metagenomics have been published, while only 770 papers with topics in metaproteomics were published.

The human gut microbiome research has seen substantial expansion in gene catalogs such as IGC (70), UHGG (38), UNITN (71), and so on. In contrast, there is a paucity of efforts directed toward curating and reanalyzing peptides and proteins from available metaproteomic datasets in repositories like PRIDE (72), and ProteomeXchange (73). While Stamoulian *et al.*, examined over a thousand metaproteomics raw files to identify generalist species expressed across all samples and specialist species that are highly expressed in a small subset of samples associated with a certain phenotype (28), their article does not focus on the proteome coverage of each microbial species. Conversely, our recently developed MetaPep (29) dataset, which compiled identification results from over 2000 raw files across 15 published metaproteomics studies, may serve as a robust resource to enhance our understanding of metaproteomic coverage. The expectation is that the assembly of more extensive peptide and protein datasets will further enrich our knowledge of the human gut metaproteome.

Advancing Coverage with DIA and Emerging Technologies—Instrumental limitations notably affect metaproteomic coverage. Traditional mass spectrometry's capability to acquire tandem mass spectra within a given time frame is inherently limited. Even given enough scanning time, not all digested peptides in the sample are detected by mass spectrometry. MS-based proteomics tends to identify proteotypic peptides (38). For instance, in conventional data-dependent acquisition (DDA) methods typically, less than

1% of incoming precursor ions are fragmented and identified by MS2 (74).

Data-independent acquisition (DIA) metaproteomics, has shown improvements in proteome coverage, reproducibility, and accuracy in quantification over DDA methods (75). Instead of acquiring MS/MS scans with narrow isolation windows centered on peptide precursors detected in an MS scan in the traditional DDA, DIA acquires MS/MS scans with wide isolation windows that do not target any particular precursor (76). Additionally, the cutting-edge DIA-PASEF (Parallel Accumulation-Serial Fragmentation) is in theory sampling up to 100% of the peptide precursor ion (74). The application of DIA-PASEF in mouse microbiome studies, which potentially doubled protein identifications, underscores its promise (77). However, the application of such advanced methodologies to human gut microbiome studies is in its infancy, hindered by the nascent state of requisite bioinformatics pipelines and the complexity inherent in the technique.

However, it should be noted that DIA-based metaproteomics is only going to make a small dent in the dark field of the metaproteome. Novel enrichment techniques specifically designed for the human gut microbiome will likely be necessary to obtain a larger coverage of the metaproteome. For instance, the application of activity-based probes (ABPs) has facilitated the enrichment of proteins possessing distinct functionalities, thereby enabling the identification and quantification of proteins that are present at levels below conventional detection thresholds (78). Furthermore, various enrichment strategies also have substantially elevated the number of small proteins that are typically challenging to detect through standard metaproteomic methodologies (79).

Enhancing Protein Annotation Through Metaproteomics

Biochemical and genetic experiments are traditional methods for elucidating protein functions. In the absence of experimental data, proteins that have not been characterized are often assumed to have the same functions as proteins that have been experimentally characterized and share high sequence similarity. Surpassing traditional alignment-based techniques, recent advances have introduced machine learning and deep learning approaches to bolster protein function annotation (80, 81). Additionally, protein structure families based on clustering the predicted structure of nearly every known protein have expanded the dimensions of the protein universe, revealing that most protein structures are not functionally dark, shedding light on the functional annotation of a broader array of proteins (82, 83).

Proteomics data has been suggested as a valuable resource for enhancing protein functional annotation (37, 84). A study has cataloged the proteome-wide protein abundance of *E. coli* in response to more than 100 genetic perturbations, casting light on the modulation of functionally linked proteins and providing mechanistic insights into this model organism (85). Similarly, metaproteomics data can also contribute to protein function annotation by identifying proteins with consistent changes in abundance under different treatments. These changes indicate similarities or connections in their roles. However, in the analysis of metaproteomics samples, it is important to consider the abundance changes of different taxa. Given the presence of numerous functionally unknown proteins with high abundances in metaproteomics samples, mining metaproteomics data holds the potential to uncover the functions of these mysterious entities.

Performing Peptide-Centric Analysis in Metaproteomics

Our findings in this article demonstrate that confining the search space to meticulously curated such as the UHGG representative prokaryotic species, enables the comprehensive annotation of most peptides with detailed functional and taxonomic information. This outcome, coupled with the demonstrated utility of peptide-centric analysis tools like UniPept (61), pepFunk (62), and MetaGOmics (63) reinforces the feasibility of peptide-centric analysis in metaproteomics.

Nevertheless, the current peptide-centric analysis mainly focuses on downstream analysis after the sequence database search, while most metaproteomics studies still search protein sequence databases. Given that the identified peptides only account for a small part of all in-silico digested peptides from proteins (86), searching a protein database increases computational complexity. To achieve a comprehensive peptide-centric analysis workflow and reduce computing resource consumption, a refined and well-annotated peptide sequence database is required. Fortunately, the advent of innovative methodologies for predicting peptide detectability (87, 88) holds promise for condensing the search space, transitioning from an exhaustive protein sequence database to a more concise peptide sequence database. Implementing a complete peptide-centric analysis workflow initiated with searching a peptide database search, is anticipated to expedite metaproteomics data processing and enable more expansive studies in this field.

Enhancing Taxonomic and Taxon-Function Interplay Analysis

Refining Taxon Scope and Calculating Confidence Scores to Improve Taxonomic Analysis—The choice of protein sequence databases is known to significantly impact the outcome of metaproteomics (89). Moreover, the delineation of the search space establishes distinct parameters for analyzing the taxonomic composition of the sample. Employing preliminary amplicon sequencing, metagenomics, or using metaproteomics data alone for predetermining and refining the taxonomic scope could enhance the precision of taxonomic analyses within metaproteomics. However, the current metaproteomics analysis pipelines lack the flexibility to customize the taxonomic scope. The incorporation of this capacity is likely to enhance not just the detail of taxonomic resolution but also the efficiency of metaproteomics searches.

In addition, when applying metaproteomics to study complex microbiome samples with multiple species, inferring the presence of taxa based on identified peptides can be a complex endeavor. While existing metaproteomics analysis platforms such as MetaLab MAG employ Occam's razor for species referencing, the reliability of identified bacterial species remains largely unknown. A confidence score calculation method has been applied for strain-level taxonomic assignment of viral proteome samples (90). Developing meaningful confidence scores represents a promising direction for reinforcing the metaproteomic toolkit's ability to dissect microbial community structures.

Advancing Tool Development for Taxon-Function Crosstalk Analysis—In the realm of omics analyses, simply compiling a list of microbial taxa and gene/protein abundances is increasingly recognized as insufficient. Going beyond this, the identification of microbes and their corresponding functional contributions to the microbial community provides novel insights (91, 92). While tools such as BURRITO (93) and MetaFunc (94), have bridged taxonomic and functional data, the broader scientific community still faces challenges when attempting to conduct taxon-function crosstalk analysis. Such analyses are invaluable as peptides/proteins inherently embody both taxonomic and functional data. Therefore, the development of analytical tools and methodologies aimed at linking taxonomic identity with functional activity is a vital direction for advancing metaproteomic research.

SUMMARY

This paper presents a comprehensive analysis of the proteome coverage of human gut bacterial species identified in metaproteomics studies. To address the challenge of limited proteome coverage, there is a growing need for expanded metaproteomic research and the advancement of innovative methodologies. We also highlight the high annotation coverage of proteins identified through metaproteomics, indicating the significant potential of metaproteomics in improving protein annotation. Additionally, we demonstrate the feasibility of peptide-centric analysis as a promising approach to reduce computational demands in metaproteomics data analysis. The paper discusses various perspectives on enhancing the reliability of taxonomic analysis and facilitating taxon-function interactions in metaproteomics. These combined efforts aim to leverage metaproteomics in enhancing our understanding of microbial ecosystems and their complex interactions with host systems.

DATA AVAILABILITY

Additional data are available upon request.

Supplemental data—This article contains supplemental data.

Acknowledgments—The authors are grateful to the Natural Sciences and Engineering Research Council of Canada.

Funding and additional information—Substantial financial support was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the discovery grant (to D. F.). Z. S. was funded by a stipend from the NSERC CREATE in Technologies for Microbiome Science and Engineering (TECHNOMISE) Program.

Author contribution—Z. S. formal analysis; Z. S., Z. N., and D. F. writing–original draft.

Conflict of interest—D. F. is a co-founder of Biotagenics and MedBiome, both of which are clinical microbiomics companies. The other authors declare no competing interests.

Abbreviations—The abbreviations used are: CDS, coding sequences; COG, Cluster of Orthologous Groups; DDA, data-dependent acquisition; DIA, data-independent acquisition; GO, Gene Ontology; LCA, lowest common ancestor; LC-MS/MS, liquid chromatography and tandem mass spectrometry; PASEF, Parallel Accumulation-Serial Fragmentation; PTMs, post-translation modifications; UHGG, Unified Human Gastrointestinal Genome.

Received November 14, 2023, and in revised form, February 26, 2024 Published, MCPRO Papers in Press, April 10, 2024, https://doi.org/10.1016/j.mcpro.2024.100763

REFERENCES

1. de Vos, W. M., Tilg, H., Van Hul, M., and Cani, P. D. (2022) Gut microbiome and health: mechanistic insights. *Gut* **71**, 1020–1032
2. Milo, R. (2013) What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays.* **35**, 1050–1055
3. Kleiner, M., Thorson, E., Sharp, C. E., Dong, X., Liu, D., Li, C., *et al*. (2017) Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **8**, 1558
4. Wilmes, P., and Bond, P. L. (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ. Microbiol.* **6**, 911–920
5. Wilmes, P., and Bond, P. L. (2006) Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* **14**, 92–97
6. Ram, R. J., VerBerkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., *et al*. (2005) Community proteomics of a natural microbial Biofilm. *Science* **308**, 1915–1920
7. Li, L., Wang, T., Ning, Z., Zhang, X., Butcher, J., Serrana, J. M., *et al*. (2023) Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics. *Nat. Commun.* **14**, 3428
8. Zhao, J., Yang, Y., Xu, H., Zheng, J., Shen, C., Chen, T., *et al*. (2023) Data-independent acquisition boosts quantitative metaproteomics for deep characterization of gut microbiota. *NPJ Biofilms Microbiomes* **9**, 4
9. Salvato, F., Hettich, R. L., and Kleiner, M. (2021) Five key aspects of metaproteomics as a tool to understand functional interactions in host-associated microbiomes. *PLoS Pathog.* **17**, e1009245
10. Kleiner, M. (2019) Metaproteomics: much more than measuring gene expression in microbial communities. *mSystems* **4**, e00115–e00219
11. Duan, H., Zhang, X., and Figeys, D. (2023) An emerging field: post-translational modification in microbiome. *Proteomics* **23**, 2100389
12. Armengaud, J. (2023) Metaproteomics to understand how microbiota function: the crystal ball predicts a promising future. *Environ. Microbiol.* **25**, 115–125

13. Kleiner, M., Dong, X., Hinzke, T., Wippler, J., Thorson, E., Mayer, B., *et al*. (2018) Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5576–E5584
14. Zhang, X., Deeke, S. A., Ning, Z., Starr, A. E., Butcher, J., Li, J., *et al*. (2018) Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.* **9**, 2873
15. Gavin, P. G., Mullaney, J. A., Loo, D., Cao, K. L., Gottlieb, P. A., Hill, M. M., *et al*. (2018) Intestinal metaproteomics reveals host-microbiota interactions in Subjects at Risk for type 1 diabetes. *Diabetes Care* **41**, 2178–2186
16. Long, S., Yang, Y., Shen, C., Wang, Y., Deng, A., Qin, Q., *et al*. (2020) Metaproteomics characterizes human gut microbiome function in colorectal cancer. *NPJ Biofilms Microbiomes* **6**, 14
17. Tanca, A., Abbondio, M., Fiorito, G., Pira, G., Sau, R., Manca, A., *et al*. (2022) Metaproteomic profile of the Colonic Luminal microbiota from patients with Colon cancer. *Front. Microbiol.* **13**, 869523
18. Kolmeder, C. A., Ritari, J., Verdam, F. J., Muth, T., Keskitalo, S., Varjosalo, M., *et al*. (2015) Colonic metaproteomic signatures of active bacteria and the host in obesity. *Proteomics* **15**, 3544–3552
19. Calabrese, F. M., Porrelli, A., Vacca, M., Comte, B., Nimptsch, K., Pinart, M., *et al*. (2021) Metaproteomics approach and pathway modulation in obesity and diabetes: a narrative review. *Nutrients* **14**, 47
20. Grenga, L., Pible, O., Miotello, G., Culotta, K., Ruat, S., Roncato, M., *et al*. (2022) Taxonomical and functional changes in COVID -19 faecal microbiome could be related to SARS-CoV -2 faecal load. *Environ. Microbiol.* **24**, 4299–4316
21. He, F., Zhang, T., Xue, K., Fang, Z., Jiang, G., Huang, S., *et al*. (2021) Fecal multi-omics analysis reveals diverse molecular alterations of gut ecosystem in COVID-19 patients. *Anal. Chim. Acta* **1180**, 338881
22. Wolf, M., Schallert, K., Knipper, L., Sickmann, A., Sczyrba, A., Benndorf, D., *et al*. (2023) Advances in the clinical use of metaproteomics. *Expert Rev. Proteomics* **20**, 71–86
23. Zhang, X., and Figeys, D. (2019) Perspective and guidelines for metaproteomics in microbiome studies. *J. Proteome Res.* **18**, 2370–2380
24. Wang, Y., Zhou, Y., Xiao, X., Zheng, J., and Zhou, H. (2020) Metaproteomics: a strategy to study the taxonomy and functionality of the gut microbiota. *J. Proteomics* **219**, 103737
25. Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017) Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **261**, 24–36
26. Issa, I. N., Philippe, D., Nicholas, A., Raoult, D., and Eric, C. (2019) Metaproteomics of the human gut microbiota: challenges and contributions to other OMICS. *Clin. Mass Spectrom.* **14**, 18–30
27. Miura, N., and Okuda, S. (2023) Current progress and critical challenges to overcome in the bioinformatics of mass spectrometry-based metaproteomics. *Comput. Struct. Biotechnol. J.* **21**, 1140–1150
28. Stamboulian, M., Canderan, J., and Ye, Y. (2022) Metaproteomics as a tool for studying the protein landscape of human-gut bacterial species. *PLoS Comput. Biol.* **18**, e1009397
29. Sun, Z., Ning, Z., Cheng, K., Duan, H., Wu, Q., Mayne, J., *et al*. (2023) MetaPep: a core peptide database for faster human gut metaproteomics database searches. *Comput. Struct. Biotechnol. J.* **21**, 4228–4237
30. Cheng, K., Ning, Z., Li, L., Zhang, X., Serrana, J. M., Mayne, J., *et al*. (2023) MetaLab-MAG: a metaproteomic data analysis platform for genome-level characterization of microbiomes from the Metagenome-Assembled genomes database. *J. Proteome Res.* **22**, 387–398
31. Lohmann, P., Schäpe, S. S., Haange, S. B., Oliphant, K., Allen-Vercoe, E., Jehmlich, N., *et al*. (2020) Function is what counts: how microbial community complexity affects species, proteome and pathway coverage in metaproteomics. *Expert Rev. Proteomics* **17**, 163–173
32. Callister, S. J., Fillmore, T. L., Nicora, C. D., Shaw, J. B., Purvine, S. O., Orton, D. J., *et al*. (2018) Addressing the challenge of soil metaproteome complexity by improving metaproteome depth of coverage through two-dimensional liquid chromatography. *Soil Biol. Biochem.* **125**, 290–299
33. Van Den Bossche, T., Kunath, B. J., Schallert, K., Schäpe, S. S., Abraham, P. E., Armengaud, J., *et al*. (2021) Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nat. Commun.* **12**, 7305
34. Hou, M., Huang, J., Jia, T., Guan, Y., Yang, F., Zhou, H., *et al*. (2023) Deep profiling of the proteome dynamics of *Pseudomonas aeruginosa* reference strain PAO1 under different Growth Conditions. *J. Proteome Res.* **22**, 1747–1761
35. Wang, H., Wan, L., Shi, J., Zhang, T., Zhu, H., Jiang, S., *et al*. (2021) Quantitative proteomics reveals that dormancy-related proteins mediate the attenuation in mycobacterium strains. *Virulence* **12**, 2228–2246
36. Duan, H., Cheng, K., Ning, Z., Li, L., Mayne, J., Sun, Z., *et al*. (2022) Assessing the dark field of metaproteome. *Anal. Chem.* **94**, 15648–15654
37. Messner, C. B., Demichev, V., Wang, Z., Hartl, J., Kustatscher, G., Mülleder, M., *et al*. (2023) Mass spectrometry-based high-throughput proteomics and its role in biomedical studies and systems biology. *Proteomics* **23**, 2200013
38. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., *et al*. (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114
39. Letunic, I., and Bork, P. (2021) Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296
40. Son, J., Na, S., and Paek, E. (2023) DbyDeep: exploration of MS-detectable peptides via deep learning. *Anal. Chem.* **95**, 11193–11200
41. Price, M. N., Wetmore, K. M., Deutschbauer, A. M., and Arkin, A. P. (2016) A comparison of the costs and benefits of bacterial gene expression. *PLoS ONE* **11**, e0164314
42. Feng, Y., Wang, Z., Chien, K.-Y., Chen, H.-L., Liang, Y.-H., Hua, X., *et al*. (2022) Pseudo-pseudogenes" in bacterial genomes: proteogenomics reveals a wide but low protein expression of pseudogenes in *Salmonella enterica*. *Nucleic Acids Res* **50**, 5158–5170
43. Chang, Y. C., Hu, Z., Rachlin, J., Anton, B. P., Kasif, S., Roberts, R. J., *et al*. (2016) COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. *Nucleic Acids Res.* **44**, D330–D335
44. Paik, Y. K., Lane, L., Kawamura, T., Chen, Y. J., Cho, J. Y., LaBaer, J., *et al*. (2018) Launching the C-HPP neXt-CP50 Pilot Project for functional characterization of identified proteins with No known function. *J. Proteome Res.* **17**, 4042–4050
45. Ghatak, S., King, Z. A., Sastry, A., and Palsson, B. O. (2019) The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. *Nucleic Acids Res.* **47**, 2446–2454
46. Ghosh, S., Whitley, C. S., Haribabu, B., and Jala, V. R. (2021) Regulation of intestinal Barrier function by microbial metabolites. *Cell. Mol. Gastroenterol. Hepatol.* **11**, 1463–1482
47. Jeffery, C. J. (2018) Protein moonlighting: what is it, and why is it important? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **373**, 20160523
48. Beynon, R. J., Hammond, D., Harman, V., and Woolerton, Y. (2014) The role of proteomics in studies of protein moonlighting. *Biochem. Soc. Trans.* **42**, 1698–1703
49. Dimonaco, N. J., Aubrey, W., Kenobi, K., Clare, A., and Creevey, C. J. (2022) No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics* **38**, 1198–1207
50. Kustatscher, G., Collins, T., Gingras, A. C., Guo, T., Hermjakob, H., Ideker, T., *et al*. (2022) Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Methods* **19**, 774–779
51. Bashir, N., Kounsar, F., Mukhopadhyay, S., and Hasnain, S. E. (2010) *Mycobacterium tuberculosis* conserved hypothetical protein rRv2626c modulates macrophage effector functions. *Immunology* **130**, 34–45
52. Enany, S. (2014) Structural and functional analysis of hypothetical and conserved proteins of Clostridium tetani. *J. Infect. Public Health* **7**, 296–307
53. Lobb, B., Tremblay, B. J. M., Moreno-Hagelsieb, G., and Doxey, A. C. (2020) An assessment of genome annotation coverage across the bacterial tree of life. *Microb. Genomics* **6**, e000341
54. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462
55. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29
56. Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., *et al*. (2023) The gene ontology knowledgebase in 2023. *Genetics* **224**, iyad031
57. He, Z., Huang, T., Zhao, C., and Teng, B. (2016) Protein inference. In: Mirzaei, H., Carrasco, M., eds. *Modern Proteomics – Sample Preparation,*

Analysis and Practical Applications, Springer International Publishing, Cham: 237–242. Advances in Experimental Medicine and Biology; vol. 919
58. Schiebenhoefer, H., Van Den Bossche, T., Fuchs, S., Renard, B. Y., Muth, T., and Martens, L. (2019) Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Rev. Proteomics* **16**, 375–390
59. Audain, E., Uszkoreit, J., Sachsenberg, T., Pfeuffer, J., Liang, X., Hermjakob, H., *et al*. (2017) In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *J. Proteomics* **150**, 170–182
60. Mesuere, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., and Dawyndt, P. (2012) Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.* **11**, 5773–5780
61. Gurdeep Singh, R., Tanca, A., Palomba, A., Van der Jeugt, F., Verschaffelt, P., Uzzau, S., *et al*. (2019) Unipept 4.0: functional analysis of metaproteome data. *J. Proteome Res.* **18**, 606–615
62. Simopoulos, C. M. A., Ning, Z., Zhang, X., Li, L., Walker, K., Lavallée-Adam, M., *et al*. (2020) pepFunk: a tool for peptide-centric functional analysis of metaproteomic human gut microbiome studies. *Bioinformatics* **36**, 4171–4179
63. Riffle, M., May, D. H., Timmins-Schiffman, E., Mikan, M. P., Jaschob, D., Noble, W. S., *et al*. (2017) MetaGOmics: a Web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. *Proteomes* **6**, 2
64. Ning, Z., Zhang, X., Mayne, J., and Figeys, D. (2016) Peptide-centric approaches provide an alternative perspective to Re-examine quantitative proteomic data. *Anal. Chem.* **88**, 1973–1978
65. Lima, T., Rodrigues, J. E., Manadas, B., Henrique, R., Fardilha, M., and Vitorino, R. (2023) A peptide-centric approach to analyse quantitative proteomics data- an application to prostate cancer biomarker discovery. *J. Proteomics* **272**, 104774
66. Rechenberger, J., Samaras, P., Jarzab, A., Behr, J., Frejno, M., Djukovic, A., *et al*. (2019) Challenges in clinical metaproteomics highlighted by the analysis of Acute Leukemia patients with gut Colonization by Multidrug-Resistant Enterobacteriaceae. *Proteomes* **7**, 2
67. Xu, J., Zhang, H., Zheng, J., Dovoedo, P., and Yin, Y. (2020) eCAMI: simultaneous classification and motif identification for enzyme annotation. *Bioinformatics* **36**, 2068–2075
68. Vicedomini, R., Bouly, J. P., Laine, E., Falciatore, A., and Carbone, A. (2022) Multiple profile models Extract Features from protein sequence data and Resolve functional diversity of Very different protein families. *Mol. Biol. Evol.* **39**, msac070
69. Verschaffelt, P., Tanca, A., Abbondio, M., Van Den Bossche, T., Moortele, T. V., Dawyndt, P., *et al*. (2023) Unipept Desktop 2.0: construction of targeted reference protein databases for Metaproteogenomics analyses. *J. Proteome Res.* **22**, 2620–2628
70. Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., *et al*. (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841
71. Leviatan, S., Shoer, S., Rothschild, D., Gorodetski, M., and Segal, E. (2022) An expanded reference map of the human gut microbiome reveals hundreds of previously unknown species. *Nat. Commun.* **13**, 3863
72. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., *et al*. (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552
73. Deutsch, E. W., Bandeira, N., Sharma, V., Perez-Riverol, Y., Carver, J. J., Kundu, D. J., *et al*. (2019) The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res.* **5**, gkz984
74. Meier, F., Brunner, A. D., Frank, M., Ha, A., Bludau, I., Voytik, E., *et al*. (2020) diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236
75. Aakko, J., Pietilä, S., Suomi, T., Mahmoudian, M., Toivonen, R., Kouvonen, P., *et al*. (2020) Data-independent acquisition mass spectrometry in metaproteomics of gut microbiota—Implementation and computational analysis. *J. Proteome Res.* **19**, 432–436
76. Ludwig, C., and Aebersold, R. (2018) Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutoria. *Mol. Syst. Biol.* **14**, e8126
77. Gómez-Varela, D., Xian, F., Grundtner, S., Sondermann, J. R., Carta, G., and Schmidt, M. (2023) Increasing taxonomic and functional characterization of host-microbiome interactions by DIA-PASEF metaproteomics. *Front. Microbiol.* **14**, 1258703
78. Mayers, M. D., Moon, C., Stupp, G. S., Su, A. I., and Wolan, D. W. (2017) Quantitative metaproteomics and activity-based probe enrichment reveals significant alterations in protein expression from a mouse model of inflammatory bowel disease. *J. Proteome Res.* **16**, 1014–1026
79. Petruschke, H., Anders, J., Stadler, P. F., Jehmlich, N., and Von Bergen, M. (2020) Enrichment and identification of small proteins in a simplified human gut microbiome. *J. Proteomics* **213**, 103604
80. Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., *et al*. (2022) Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–937
81. Maranga, M., Szczerbiak, P., Bezshapkin, V., Gligorijevic, V., Chandler, C., Bonneau, R., *et al*. (2023) Comprehensive functional annotation of Metagenomes and microbial genomes using a deep learning-based method. *mSystems* **8**, e01178–e01222
82. Durairaj, J., Waterhouse, A. M., Mets, T., Brodiazhenko, T., Abdullah, M., Studer, G., *et al*. (2023) Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653
83. Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C. L. M., Wein, T., *et al*. (2023) Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645
84. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355
85. Mateus, A., Hevler, J., Bobonis, J., Kurzawa, N., Shah, M., Mitosch, K., *et al*. (2020) The functional proteome landscape of Escherichia coli. *Nature* **588**, 473–478
86. Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., *et al*. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25**, 125–131
87. Serrano, G., Guruceaga, E., and Segura, V. (2020) DeepMSPeptide: peptide detectability prediction using deep learning. *Bioinformatics* **36**, 1279–1280
88. Yu, M., Duan, Y., Li, Z., and Zhang, Y. (2021) Prediction of peptide detectability based on capsnet and convolutional block attention module. *Int. J. Mol. Sci.* **22**, 12080
89. Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., *et al*. (2016) The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* **4**, 51
90. Holstein, T., Kistner, F., Martens, L., and Muth, T. (2023) PepGM: a probabilistic graphical model for taxonomic inference of viral proteome samples with associated confidence scores. *Bioinformatics* **39**, btad289
91. Langille, M. G. I. (2018) Exploring Linkages between taxonomic and functional profiles of the human microbiome. *mSystems* **3**, e00163–e00217
92. Greslehner, G. P. (2020) Microbiome structure and function: a new framework for interpreting data. *BioEssays* **42**, 1900255
93. McNally, C. P., Eng, A., Noecker, C., Gagne-Maynard, W. C., and Borenstein, E. (2018) BURRITO: an interactive multi-Omic tool for visualizing taxa–function Relationships in microbiome data. *Front. Microbiol.* **9**, 365
94. Sulit, A. K., Kolisnik, T., Frizelle, F. A., Purcell, R., and Schmeier, S. (2023) MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes. *Gut Microb.* **4**, e4