Shining a Light on the Dark-Field of the Metaproteome

Haonan Duan

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements
for the Doctorate in Philosophy degree in Biochemistry

Ottawa Institute of Systems Biology
Department of Biochemistry, Microbiology and Immunology
Faculty of Medicine
University of Ottawa

Abstract

Metaproteomics has emerged as a powerful tool for studying human gut microbiomes. However, the intricate nature of these microbial communities often surpasses the analytical depth of current metaproteomic methodologies, particularly those relying on traditional Data-Dependent Acquisition (DDA) techniques. In this thesis, we first quantitatively assess the limitations of DDA-based metaproteomics, revealing a significant underrepresentation of proteins from low-abundance species. This finding underscores the necessity for analytical approaches capable of capturing the full spectrum of microbiome samples.

Further investigation into the capabilities of mass spectrometry to discern the complexity of human gut microbiomes revealed that the complexity of MS1 spectra from microbiome samples significantly exceeds that of individual species. This suggests that many species are potentially detectable at the MS1 level but are systematically overlooked by conventional shotgun metaproteomic strategies, which prioritize peptides for MS2 fragmentation based on abundance.

In response to these challenges, we explored the application of Data-Independent Acquisition (DIA) in metaproteomics. We developed and optimized MetaDIA, a novel workflow for the analysis of microbiome DIA data. MetaDIA operates independently of DDA-derived spectral libraries, offering a robust alternative that enhances the detection and characterization of microbial proteins. Our comparative analyses demonstrate that

MetaDIA achieves strong consistency with the result from traditional DDA-derived spectral libraries, both in terms of protein identification and functional annotation.

This work not only highlights the limitations of current metaproteomic methodologies in capturing the full complexity of microbiomes but also introduces MetaDIA as a promising strategy for advancing the depth and accuracy of microbial community analysis.

Acknowledgment

Foremost I would like to express my gratitude to my supervisors Dr. Daniel Figeys who provided me with the opportunity to study in such a supportive lab. Every time I come up with wild ideas, I always get thoughtful responses after careful consideration. Your encouragement has helped me gradually build confidence in research.

I would like to express my gratitude to Dr. Zhibin Ning. You generously and patiently taught me about mass spectrometry and chromatography, knowledge that cannot be learned from books. Your guidance in my work is much appreciated. I am forever grateful for your support.

I owe an enormous debt of gratitude to my labmate Zhongzhi Sun. Sorry for always interrupting your work. But thank you deeply for always listening attentively to my descriptions of the problems encountered in my project. Your listening and feedback always help me clarify my thoughts and are of great assistance to me.

Ottawa was once an unfamiliar city to me. I would like to extend the deepest thanks to Alexander Dewar, Peter Dobranowski and Krystal Walker who were very friendly and helped me adapt to life here.

I would like to thank my thesis advisory committee members Dr. Mathieu Lavallee, Dr. Alain Stintzi and Dr. Jean-François Couture for guiding me through these years of my PhD and by providing me with helpful suggestions and advice. All your support and constructive criticism are much appreciated.

I would also like to express my gratitude to the past and present members of the Figeys lab for their friendship, feedback, and insights, and for providing me with an outstanding support network. Thank you, Dr. Galen Guo, for providing advice for the thesis writing. A special thanks to Dr. Cheng Kai, Dr. Janice Mayen, Qing Wu, Ailing Zhang, Dr. Xu Zhang, Dr. Leyuan Li, Adrian Beltran and Hongye Qin. It was a great pleasure working with you.

Lastly, I would like to sincerely thank my family and friends. Without your support, I would not have been able to reach this point.

Table of Contents

Contents

Abbreviations

C

CKD: Chronic kidney disease

CDI: *C. difficile* infection

D

DDA: data-dependent acquisition

DIA: data-independent acquisition

F

FDR: false discovery rate

H

HAPs: high-abundance proteins

HMP: Human Microbiome Project

I

IBD: inflammatory bowel disease

IGC: Integrated Gene Catalog

M

m/z: mass-to-charge ratio

MBR: match-between-run

MetaHit: Human Intestinal Tract

MAGs: metagenome-assembled genomes

P

PBS: phosphate-buffered saline

PSMs: peptide-spectrum match

S

SCFAs: Short-chain fatty acids

SNPs: single nucleotide polymorphisms

T

TMAO: trimethylamine N-oxide

U

UHGG: Unified Human Gastrointestinal Genome

List of Figures

Figures

Supplementary Figures

Chapter 1. General introduction

1.1 Human Gut Microbiome: importance and complexity

1.1.1 A complex “organ”

The human gut represents a semi-closed, nutrient-rich anaerobic system that provides a conducive environment for microorganisms. Approximately the same number of microbial cells as human body cells reside within the human gastrointestinal tract[1]. The total dry weight of bacteria in the body is about 50–100g[1]. These microbes interact with host cells, perform biological functions, and significantly impact the host's health[2]. Given their vast numbers and critical biological roles, the gut microbiota is considered an organ in its own right[3].

The composition of the gut microbiome is remarkably diverse, encompassing viruses, prokaryotes, and eukaryotes[4]. Genomic studies have demonstrated that the gut of a healthy individual typically harbors approximately 200 species[5], the majority of which are bacteria. Due to the diversity of gut microbiome, although the number of cells in the human gut microbiome is roughly equivalent to that of the host, the size of its genome is much larger than that of the host[6, 7]. Furthermore, there is significant inter-individual variability in the gut microbiota, with different individuals hosting distinct microbial communities. This variability contributes to the continuous expansion of the reference gene library for human gut microbiota. The latest human gut genome catalog includes 4,744 genomes[8], which significantly surpasses the number of species found in any single individual. However, it is important to note that despite differences in species

composition, the functional capacity of these microbiomes is predicted to be highly conserved[9-11].

1.1.2 Human gut microbiome and healthy

The human gut microbiome performs numerous roles in maintaining human health and is related to many diseases such as obesity[12], chronic kidney disease (CKD)[13] and inflammatory bowel disease (IBD)[14]. The gut microbiome produces an extensive array of metabolites, some of which have been shown to be important to human health. For example, short-chain fatty acids (SCFAs) are produced primarily through the fermentation of dietary fibers by the anaerobic bacteria in the colon and play a crucial role in host health. SCFAs provide an important source of energy for the body. Butyrate is the main energy source for colonocytes[15]. Increasing fiber intake or oral butyrate supplementation has been shown to reduce obesity and improve insulin sensitivity[16]. The gut microbiota is capable of metabolizing tyrosine and tryptophan into p-cresol and indole, respectively. These substances will be absorbed by the intestines and subsequently metabolized by the liver into p-cresol sulfate and indoxyl sulfate, which are uremic toxins and can cause renal damage and kidney disease[17, 18].

The gut microbiome is also associated with the intestinal barrier, serving as the first line of defense. A healthy microbiome can inhibit pathogen colonization through several mechanisms, including direct killing, competition for limited nutrients, and enhancement of immune responses[19]. Furthermore, the integrity of the gut barrier, particularly its mucus component, relies on the maintenance by gut microbiome. Mucus is primarily composed of mucins, which are host-produced glycoproteins[20]. The gut microbiome

possesses a rich array of glycan-degrading enzymes, enabling them to utilize mucin as an energy source and produce SCFAs[21]. This interaction between mucus and the intestinal microbiota establishes a dynamic equilibrium. A diet devoid of fiber can lead the microbiome to excessively exploit mucus, resulting in gut barrier dysfunction[22]. Conversely, in germ-free mice, the absence of microbiota consuming the mucus results in a thicker and more viscous mucus layer that is paradoxically more penetrable, allowing harmful substances to cross the gut barrier more easily[23].

Given the relationship between the microbiota and health, some pharmaceuticals and therapeutic approaches targeting the gut microbiome have been developed and implemented. *Clostridioides difficile*, a pathogen capable of producing toxins that lead to intestine infections (*C. difficile* infection, CDI), is traditionally treated with antibiotic medications such as vancomycin or fidaxomicin[24]. However, antibiotics are ineffective against *C. difficile* spores and concurrently disrupt the microbiome, resulting in a recurrence rate of 45–65% of patients after antibacterial therapy[25]. Faecal microbiota transplantation (FMT), which involves the transfer of feces from a healthy donor to a recipient, has been utilized to decrease the risk of CDI recurrence. Nonetheless, the efficacy of FMT is inconsistent. In response, RBX2660 was developed and has received FDA approval[26]. RBX2660 is a live fecal microbiota suspension derived from donated human fecal matter for rectal administration. Subsequently, SER-109, an orally administration product focusing on a narrower consortium of microbes, has also shown positive results and received FDA approval[26]. Another preclinical study in mice has demonstrated that the addition of competitive inhibitors can significantly inhibit the

production of the gut microbiota-associated uremic toxin precursor, p-cresol, thereby alleviating chronic kidney disease[27].

1.1.3 Methods to study human gut microbiome

Due to the fact that the majority of gut microbes are uncultivable[28], it is challenging to study individual species by isolation. The prevalent approach involves examining the gut microbiome as a whole, which facilitates understanding the interactions among different species. Human fecal samples are commonly used for studying the gut microbiome. The advantage of using fecal samples is their ease of collection; however, they have several drawbacks, including poor sample stability, susceptibility to various interfering factors, and significant variation in sample quality depending on the method and timing of collection[29]. Moreover, it is difficult to assess the impact of exposure factors on the gut microbiota using these samples. An alternative strategy involves simulating the gut environment and cultivating gut microbial community *in vitro*. These in vitro culture models, which allow for controlled experimental conditions, can produce highly reproducible results[30]. Some models support high-throughput experiments while maintaining the stability of taxonomic and functional composition, thereby facilitating research into the effects of exposure factors, like drugs, on the gut microbiome[31, 32].

The current methodologies for studying the gut microbiome primarily involve nucleic acid sequencing techniques such as 16S rRNA gene sequencing[33], metagenomics[34], and metatranscriptomics[35], as well as mass spectrometry-based approaches like metaproteomics and metabolomics[36]. The first three techniques enable the high-throughput and rapid acquisition of information on the microbiome taxonomic composition

and their relative abundance through DNA or RNA sequencing. Moreover, metagenomics and metatranscriptomics can provide insights into the potential functions of the microbiome. However, not all genes are expressed, and not all mRNA transcripts are translated into proteins. Thus, the functional information obtained from gene sequencing technologies represents only the potential capabilities. Research has found that there is a very low correlation between the results of proteomics and transcriptomics[37, 38].

Therefore, proteomics and metabolomics are essential tools in studying the function of the microbiome. Some metabolites derived from the gut microbiome, such as SCFAs[39], trimethylamine N-oxide (TMAO), secondary bile acids[40], and tryptophan derivatives[41], significantly impact the host's health. While identifying and quantifying these metabolites is important, it is crucial to understand the mechanisms of their production and their effects under conditions of health or illness. Metaproteomics, capable of quantitatively analyzing thousands of proteins simultaneously, can provide the necessary information to elucidate these biological processes.

1.2 Metaproteomics: A Continually Optimized Method

Metaproteomics is a specialized branch of proteomics. It extends the approach to the study of proteins from complex microbial communities rather than a single organism. The research objective of bottom-up metaproteomics are peptides. These peptides, originating from the intricate microbiome system, exhibit a substantial dynamic range in their abundance (the ratio between the maximum and minimum values). For instance, a peptide derived from a high-abundance species and protein will exhibit a significant disparity compared to one from a low-abundance species and protein. This variance in

abundance at the species level is absent in the proteomics of a single species. Similar to conventional mass spectrometry-based proteomics, metaproteomics encompasses several critical processes, including protein extraction, protein digestion, mass spectrometry analysis, protein identification and quantification, and subsequent taxonomic and functional analysis.

1.2.1 Protein extraction

Many microorganisms possess cell walls that render conventional protein disruption methods ineffective for the extraction of bacterial proteins[42]. According to previous studies, mechanical cell disruptions, such as bead beating or ultrasonication, are required for exhaustive protein extractions from human gut microbiome[43]. Additionally, it is found that different lysis buffers can selectively extract proteins from either Gram-positive or Gram-negative bacteria due to variations in cell wall thickness[44].

For the studies of the human gut microbiome, fecal samples are most commonly used. The composition of feces can be influenced by diet and drug intake, making the selection of an appropriate microbial cell extraction protocol critical. Food debris may contain contaminant proteins, inorganic salts, drug molecules and their metabolites, and natural compounds, all of which could potentially interfere with protein digestion efficiency. Therefore, employing phosphate-buffered saline (PBS) or normal saline to wash and centrifuge the sample multiple times is essential to mitigate these effects[45].

1.2.2 Mass spectrometry analysis

The proteins extracted from the microbiome are enzymatically digested into peptides, which are subsequently ionized and injected into a mass spectrometer. The peptides

mass-to-charge ratio (m/z) and relative abundance are measured and recorded in MS1. Additionally, tandem mass spectrometry (MS/MS) is employed to selectively fragment peptides in the gas phase and to measure the m/z of the generated fragments, further aiding in the determination of peptide sequences. The comprehensive data obtained from the mass spectrometer are analyzed with bioinformatics tools, comparing the observed spectra against theoretical or database spectra to infer the sequence information of the peptides, thereby elucidating the protein composition of the microbiome.

There are two major methods in bottom-up proteomics: data-dependent acquisition (DDA) and data-independent acquisition (DIA). DDA is extensively used in metaproteomics, whereas DIA, a more recent approach, has not yet to be extensively applied. In DDA, the most intense peaks in the MS1 are selected for fragmentation and identification. Thus, DDA a bias based upon abundance is introduced. The selection of low abundant peptides is usually stochastic in DDA. As well, a complex background can decrease the sensitivity in DDA mode[46]. Strategies like dynamic exclusion and match-between-run (MBR) has been used to increase the identification rate of DDA[47]. Nevertheless, when DDA is used to analyze complex samples having a larger dynamic range of proteins like microbiota communities, a great deal of information from low-intensity peptides may be lost. To the best of our knowledge, the largest number of peptides identified in a single metaproteomic study is approximately 45,000 using an ultra-deep analysis method with DDA mode[48]. This method requires a pre-fractionation before loading samples onto the MS and much longer MS run times. Despite these measures, it remains noncomprehensive. Considering the high diversity of the human gut microbiota, many peptides remain to be

characterized. Therefore, the depth and resolution of MS-based metaproteomics by DDA mode on the human microbiota need to be evaluated and improved.

In contrast, the DIA approach simultaneously isolates and fragments peptides per m/z selection windows, collecting all the fragments ions indiscriminately[49]. All the peptides in the specific window can be fragmented thus introducing no bias during sampling, including those peptides that lack detectable precursors in MS1[50]. In experiments involving large sample sizes, the randomness inherent in the DDA mode results in a significant number of missing values. Consequently, despite the total number of identified peptides increased by the number of samples, the subset of peptides available for analysis may remain constant or even decrease. Conversely DIA data act akin to a library, compiling all peptide fragments, thereby demonstrating notable robustness, sensitivity, and reproducibility, with fewer missing values[51].

The DDA mode typically necessitates extended chromatographic gradients to diminish the complexity in MS1, thereby facilitating deeper detection. Whereas the DIA mode does not require this. The DIA mode can be integrated with microLC which enables high-throughput analysis[52]. Consequently, this attribute renders it especially appropriate for performing large-scale analyses. The DIA-PASEF[53] method incorporates ion mobility separation into the DIA workflow, introducing a fourth dimension of ion mobility analysis to the conventional three-dimensional dataset (retention time, m/z, intensity). This enhancement not only augments the structural information of analytes but also improves ion utilization efficiency by exploiting the linear relationship between ion mobility and the m/z. An enhancement in the scanning speed of mass spectrometers has facilitated the use of narrower isolation windows in DIA, a technique referred to as narrow-window

DIA[54]. This method employs a scanning speed of approximately 200 Hz, utilizing 2-Th isolation windows, which approaches the isolation windows typically used in DDA mode. With this approach, a duty cycle that spans from 380 to 980 m/z requires only about 1.5 seconds to complete. This technique effectively blurs the distinction between DIA and DDA, achieving comprehensive coverage of peptide precursors as well as high levels of quantitative precision and accuracy.

Given the advantages of DIA introduced above, it has become increasingly popular in regular proteomics. However, the benefits of these techniques have not been fully realized in the field of metaproteomics. The primary challenge lies in the requirement to use the information from the mass spectrometer to find the best matching peptides within an extensive database of proteins potentially present in a complex and unknown samples.

1.2.3 Reference database

As previously mentioned, in proteomics, the identification of peptides relies on comparing the mass spectra with theoretical fragments, which are derived from protein sequences obtained through genomic sequencing.

One study individually sequenced sample to construct a matched database from metagenomic sequencing results[55]. Additionally, other research has involved manually selecting potential genomes contained within the samples and combined them into a metagenome database[56]. However, performing metagenomic sequencing on each sample significantly increases the cost of metaproteomics, thereby impeding its widespread adoption.

Several metagenomics projects have endeavored to construct a comprehensive database for general use by employing large sample sizes. The Metagenomics of the Human Intestinal Tract (MetaHit) project[57] includes samples of the gut microbiota from 124 Europeans, while the Human Microbiome Project (HMP)[58] encompasses samples from 136 Americans. A subsequent study, Integrated Gene Catalog (IGC) merged these two databases with additional newly sequenced data, resulting in a database that contains thousands of samples from three continents[6]. Ultimately, the IGC database comprises over 9 million genes. But the IGC aggregated all genes into a single file without providing taxonomic information for each gene which caused inconvenience for subsequent taxonomic and functional analysis.

In a more recent study, the Unified Human Gastrointestinal Genome (UHGG) project significantly expanded the sample size by integrating sequencing results from over 500 human gut-related bacteria and three metagenomics datasets[8]. The number of proteins contained within the UHGG database is twice that of the IGC. Furthermore, the UHGG database has binned proteins into 4,744 representing genomes, referred to as metagenome-assembled genomes (MAGs). Each protein is not only linked to taxonomic information through eggNOG annotations but also functional information, substantially facilitating subsequent taxonomic and functional analysis.

The databases mentioned above are constructed from metagenomic sequencing data to create theoretical protein libraries. As they encompass an increasingly large number of samples, the databases become more comprehensive. However, this characteristic serves as both an advantage and a disadvantage. The extensive databases consume significant computational resources during matching spectra to peptides. Furthermore,

since the identification of peptides relies on a target-decoy strategy, the vast search space leads a high number of random matches. This situation makes it challenging to find true peptide-spectrum matches (PSMs) when applying a false discovery rate (FDR) control. It is important to note that not all genes are expressed, and not all expressed proteins can be detected. Ideally, a database should contain all expressed proteins that are detectable.

The MetaPep database has a different perspective summarize the identification results from 2,134 samples across 415 individuals[59]. It constructs a core peptide database using all identified peptides. This database contains approximately 1 million high-detectability peptides, which extremely reduces the search space. However, the results from MetaPep are limited by the detection depth of the current DDA based data and the representativeness of the samples. Applying it to deep metaproteomics or a broader range of samples may result in a lower number of identified peptides than expected.

1.2.4 Database searching strategy and software for DDA data.

In mass spectrometry-based proteomics, the information of peptides and its fragments including retention time, abundance and m/z will be recorded in the mass spectra. The peptide identification relying on matching the experimental spectra to the peptide sequence based on the information above. However, not all identified spectra perfectly match known peptides from the database. Some matches can be random noise or incorrect assignments. To distinguish the true matches from the random matches, the target-decoy strategy is commonly used[60]. In this method, a decoy databased, which is a reversed or scrambled version of the original (target) database, is created. The experimental spectra will be searched against both the target and decoy databases. Hits

to the decoy database serve as a control to estimate the FDR of identifications, allowing researchers to filter results based on a desired level of confidence and thus enhancing the reliability of peptide identification.

As previously mentioned, the reference gene database of the human gut microbiome is excessively large, which necessitates substantial computational resources and elevates the rate of Type II errors (false negatives)[45]. In response to this, the iterative searching strategy has been gradually established in metaproteomics. Jagtap *et al*.[61, 62] firstly introduced a two-step searching method and applied it on oral microbiome. In this method, the first search was performed in a target-only version where all the peptide sequence matched to the target database will be collected to construct a much smaller database for the second search. In the second search, a strict FDR control using decoy database will be applied to ensure the matching accuracy. Zhang *et al*.[63] applied this strategy on gut microbiome using IGC as the database in the first search. They built a workflow named MetaPro-IQ and enabled comparable identifications with the matched metagenome database search strategy. The strategy was further implemented in MetaLab, a user-friendly software enabling an integrated pipeline including taxonomic and functional analysis[64]. In the subsequent updated versions, MetaLab 2.0 has added the function of open searching, enabling identification of post-translational modified peptides[65].

However, in those methods, the first search inevitably involves using a large database like IGC which cost a lot computing resource. Stamboulian *et al*.[66] introduced a novel iterative searching method. They do the first search only using the high-abundance proteins (HAPs) (ribosomal proteins and elongation factors) which is a much smaller searching space compared to using the whole genomes. The result of the first search will

be used to infer the taxonomic composition of the underlying microbiome samples. The second step is to expand the search database by including all proteins from identified species. This method is faster and identifies more peptides than MetaPro-IQ probably due to the concise searching space. The latest version of MetaLab, MetaLab MAG adopted this strategy and integrated it with the more comprehensive UHGG database[67]. In addition to theses methods, there are some attempts using de novo peptide sequencing independently of protein database[68, 69]. This method can identify sequence variants, such as single nucleotide polymorphisms (SNPs) or mutations, which may be missed by database searching methods.

1.2.5 Database searching strategy and software for DIA data.

Currently, DDA continues to dominate the field of metaproteomics. The adoption of DIA in metaproteomics is still limited. The search strategies for DIA data can be categorized into library-based and library-free approaches. The library-free strategies are further subdivided is into spectrum-centric and peptide-centric methods.

The library-based approach involves analyzing samples using DDA to collect spectra of identified peptides, which are then used to construct a spectral library for DIA searching. This strategy has been used in some metaproteomics research[70-72]. Typically, this method generates a very small library, costing fewer computing resources and its results also inherit the advantage of fewer missing values in DIA data. However, the drawbacks of this approach are significant. It necessitates processing samples twice, thereby consuming both samples and mass spectrometry time. Moreover, this method is limited

to identifying peptides that have been previously identified in DDA mode, thus not fully exploiting the potential for deeper analysis offered by DIA.

The spectrum-centric algorithms identify peptides by initially associating each individual spectrum with a corresponding peptide sequence. The conventional searching strategy for DDA data can be classified as spectrum-centric[73]. When applying this strategy for DIA data, which is spectrum-centric library-free approach, it often involves deconvoluting DIA data into pseudospectra. Each pseudospectrum corresponds to a single peptide, similar to DDA data. Consequently, this enables the utilization of conventional DDA search strategies for the analysis. Tools using this strategy, such as DIA-Umpire[74] and Group-DIA[75] have been developed. However, due to the complexity of DIA data, there is currently no algorithm capable of perfectly reconstructing DIA data into DDA data. Pietilä *et al*.[76] applied DIA-Umpire in microbiome data, resulting in a fewer number of identified peptides than a typical DDA experiment[63].

An increasing number of algorithms adopt a peptide-centric strategy to handle DIA data. Unlike spectrum-centric approaches, peptide-centric approaches query the data for the best supporting evidence of detection for each query peptide. Many tools have been developed using this strategy such as FT-ARM[77], PECAN[78] and DIA-NN[79]. Currently, DIA-NN using deep neural networks to distinguish between the target and decoy precursors is the state-of-art tool for peptide identification from DIA data. However, this strategy has not yet been applied in metaproteomics. Because the query unit in peptide-centric strategy is peptide instead of spectrum, the computational work will increase as the size of the database grows. Applying the strategy directly on database of human gut microbiome is almost impossible. Based on genomic data, the average human

gut microbiome contains approximately 200 species[5]. Even employing the HAPs-guided method to extract potentially present MAGs, the size of the resulting database still exceeds the processing capabilities of current DIA search algorithms. Consequently, there is an urgent need for a more powerful strategy to reduce the search space for DIA metaproteomics.

1.3 Rationale and Hypothesis

The human gut microbiome constitutes a complex system intricately linked to health. Metaproteomics serves as a crucial approach for investigating the gut microbiota. Despite its widespread application, the depth of detection achieved by metaproteomics has not yet to be evaluated. I hypothesize that current metaproteomic DDA-based approaches may predominantly profile high-abundance species, leaving many others poorly characterized. Furthermore, I hypothesize that an appropriate database reduction method could facilitate the use of DIA in metaproteomics, thereby enhancing the analytical depth of metaproteomic studies.

Chapter 2. Assessing the dark field of metaproteome

Preface

To validate our first hypothesis, in this investigation, we endeavored to assess the depth of existing DDA-based metaproteomic methodologies. We prepared various samples by mixing *E. coli* with a microbiome in differing ratios, thus creating a gradient of *E. coli* abundance within these samples. The same analytical procedure was applied across all samples to maintain consistency in our examination. Our objective was to ascertain the current DDA-based metaproteomics' capability in detecting species of varying abundances and to determine the detection threshold for individual species. The findings of this study may unveil deficiencies in current methodology, laying the groundwork for our subsequent research endeavors.

The following chapter consists of a research article for which I am the primary author. The article was published in *Analytical Chemistry*.

H. Duan, K. Cheng, Z. Ning, L. Li, J. Mayne, Z. Sun, D. Figeys, Assessing the Dark Field of Metaproteome, Anal Chem 94(45) (2022) 15648-15654.

HD, DF and LL developed the concept and designed the experimental outline. HD performed the experiment. HD and KC performed bioinformatics analysis. ZN and ZS provided conceptual advice on data analysis. HD, JM and DF wrote the manuscript.

2.1 Abstract

The human gut microbiome is a complex system composed of hundreds of species and metaproteomics can be used to explore their expressed functions. However, we do not know the minimal abundance of a bacterium in a microbiome(depth) that can be detected by shotgun metaproteomics. In this study, we spiked ^{15}N-labelled *E. coli* peptides at different percentages into peptides mixture derived from the human gut microbiome to evaluate the depth that can be achieved by shotgun metaproteomics. We observed that the number of identified peptides and peptide intensity from ^{15}N-labelled *E. coli* were linearly correlated with the spike-in levels even when ^{15}N-labelled *E. coli* was down to 5% of the biomass. Below that level, it was not detected. Interestingly, the match-between-run strategy significantly increased the number of quantified peptides even when ^{15}N-labelled *E. coli* peptides were at low abundance. This is indicative that in metaproteomics of complex gut microbiomes many peptides from low abundant species are likely observable in MS1 but are not selected for MS2 by standard shotgun strategies.

2.2 Introduction

The human gut microbiome is an important factor for human health and its dysbiosis is related to many chronic disease[80]. The gut microbiome is a complex system having a relatively equal number of cells to that of its host[1, 81]. Metagenomics has been commonly used in human microbiome studies[6, 34]. Whereas it only reveals the functional potential of the microbiome, metaproteomics can reveal direct functional information and is more suitable to study functional alterations. However, the analysis of lower abundant proteins by metaproteomics remains an issue.

Genome sequencing has revealed that a healthy individual harbours around 200 bacterial species in their guts[5, 57]. However, bacterial abundances in the microbiome are not uniform with a few species forming the bulk of the microbiome[57]. Therefore, most bacteria species in a microbiome are of low abundance and their proteins would also be in low abundance. Although gene copy number does not exactly correspond to protein biomass, it highlights the high diversity of the human gut microbiome.

The protein analysis depth that can be achieved by metaproteomics for the human gut microbiome is not well understood. Data-dependent acquisition (DDA) is extensively used in metaproteomics. In DDA, the most intense peaks in the mass spectrometer (MS)[1] are selected for fragmentation and thus a bias based upon abundance is introduced. The selection of low abundant peptides is usually stochastic in DDA. As well, background proteins can decrease the sensitivity in DDA mode[46]. Strategies like dynamic exclusion and match-between-run (MBR) can be used to increase the identification rate of DDA. Nevertheless, when DDA is used to analyze complex samples like microbiome communities, a great deal of information from low abundance peptides may be lost. To the best of our knowledge, the largest number of peptides identified in a single metaproteomic study is approximately 45,000 using an ultra-deep analysis method with DDA mode[48]. This method requires a pre-fractionation before loading samples onto the MS and much longer MS run times. However, in this study, a single bacterial species generated ~45% of those peptides. Considering the high diversity of the human gut microbiome, many peptides remain to be characterized. Therefore, the depth and resolution of MS-based metaproteomics on the human microbiome need to be evaluated.

In this study, we labelled an *E. coli* strain with ^{15}N, then mixed the heavy-labelled *E. coli* peptides with unlabeled human gut microbiome peptides at different percentages. We then evaluated the performance of the DDA mode for the detection of ^{15}N-labelled *E. coli* peptides at the different spike-in percentages. We also evaluated the MBR strategy used in DDA.

2.3 Experimental section

2.3.1 Bacteria culture and ^{15}N heavy labelled *E. coli*.

The *Escherichia coli* (DSM 101114; Leibniz Institute DSMZ- German collection of microorganisms and cell cultures) powder was rehydrated in LB broth (Millipore Sigma, ON, CAN) then streaked onto sheep blood agar (tryptic soy agar (Fisher Scientific, ON, CAN) with 5% (v/v) sheep blood (Cedarlane, ON, CAN)) and was anaerobically cultured at 37°C overnight. One isolated colony was transferred into ^{15}N media (Cambridge Isotope Laboratories, Inc. CGM-1000-NS, QC, CAN) and cultured at 37°C for 30 hours. The samples were centrifuged at 14,000g at 4°C for 5 mins to pellet the cells. After the supernatant was removed, the pelleted cells were washed with ice-cold 1X phosphate-buffered saline (PBS; pH 7.4) three times and pellets were stored at -80°C prior to protein extraction.

Blautia hydrogenotrophica (DSM10507; Leibniz Institute DSMZ) and *Bacteroides uniformis* (ATCC8492; Cedarlane, ON, CAN) were processed in the same method but cultured in ^{14}N medium.

2.3.2 Human stool microbiome.

The Human stool was collected from a healthy adult volunteer at the University of Ottawa, Ottawa, ON, CAN. The protocol (# 20160585-01H) was approved by Ottawa Health Science Network Research Ethics. Briefly, the fresh stool was resuspended to 20% (w/v) in cold PBS containing protease inhibitors, then mixed with 2.5-mm glass beads. The fecal slurry was centrifuged at 700 g, 4°C for 5 mins to remove debris. The supernatant was collected and centrifuged at 14,000 g for 30 mins. The pellet was then resuspended with cold PBS and centrifuged for another 30 mins at 14,000 g, 4°C. The pellets were stored at 80°C for protein extraction.

2.3.3 Cultured microbiome.

The culture was based on our previous published method[32]. Briefly, 100 µL fecal inoculum was mixed with 900 µL optimized medium in a 96-well plate. All the operations were carried out in an anaerobic chamber. Bacteria were harvested after 24-hours culture at 37°C. Following culture, the plate was centrifuged at 2272g at 4°C for 45 mins. The supernatant was removed. With the plate sitting on ice, 1 ml cold PBS was added to each well and mixed thoroughly to wash the cells. Two additional washes were carried out. The plate was centrifuged at 300g at 4°C for 5 mins to spin down the debris following each wash. Following the 300g spins, the cell suspension was centrifuged at 2272g at 4°C for 45 mins. The supernatant was removed. Cell pellets were stored at -80°C for protein extraction.

2.3.4 Protein extraction and tryptic digestion.

Bacteria cells were lysed in lysis buffer (6 M urea (Millipore Sigma) in 50 mM Tris-HCl buffer (pH = 8.0; Millipore Sigma), 4% (w/v) sodium dodecyl sulfate (SDS; Millipore Sigma), and Roche PhosSTOP™ and Roche cOmplete™ Mini tablets). The bacterial lysate was ultrasonicated for 10 mins at 8°C (Q125 Qsonica, USA) using a round of 10 s ultrasonication and 10 s cooling down at 50% amplitude. The lysate was centrifuged at 16,000 g to remove the debris. Total protein was precipitated by adding ice-cold acetone/ethanol/acetic acid (50:50:0.1; Fisher Scientific) at a 1:5 (v/v) overnight at -20°C. Proteins were pelleted at 16,000g and 4°C. Protein pellets were washed with 100% acetone three times and pellets were dissolved in 6 M urea, 50 mM ammonium bicarbonate (pH 8.0; Millipore Sigma). Protein concentration was measured by the detergent compatible (DC) assay (Bio-Rad, USA). 50 µg of protein was reduced and alkylated with 10 mM dithiothreitol (37°C for 1 h) and 20 mM iodoacetamide (room temperature in the dark for 45 min). After 10× dilution with 50 mM ammonium bicarbonate (pH 8.0), the protein was digested by 1 µg trypsin (Worthington biochemicals, USA) per 50 µg proteins at 37°C for 24 h. Peptides were desalted by 10 µm C18 column (Dr.Maisch HPLC GmbH, Ammerbuch, Germany). After freeze-drying, each sample was re-dissolved in 0.1% (v/v) formic acid (Millipore Sigma).

2.3.5 Generation of serial dilution mixtures.

Peptide concentrations were measured using Thermo Scientific™ Pierce™ Quantitative Colorimetric Peptide Assays according to the manufacturer's directions. The serial dilution mixtures were prepared from 0.0005% to 100% of ^{15}N labelled peptides with triplicates.

Dried peptide mixtures were dissolved into an equal volume of 0.1% (v/v) formic acid before MS analysis.

2.3.6 LC−MS/MS analysis.

Samples were analyzed by an UltiMate 3000 RSLCnano system (Thermo Fisher Scientific, USA) coupled to Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific, USA). Peptides were loaded onto a tip column (75 μm inner diameter × 15 cm) packed with reverse phase beads (3 μm/120 Å ReproSil-Pur C18 resin, Dr. Maisch HPLC GmbH). A 60-min gradient of 5 to 35% (v/v) from buffer A (0.1% (v/v) formic acid) to B (0.1% (v/v) formic acid with 80% (v/v) acetonitrile) at a flowrate of 300 μL/min was used. The mass spectrometer was in data-dependant mode with top15 method. Dynamic exclusion repeat count was set to one and repeat exclusion duration of 20 s. Peptides are fragmented using higher-energy collisional dissociation (HCD) with a normalized collision energy of 30%. MS/MS scans were acquired at a resolution of 15,000 at m/z 200. The full mass scan was from 350-1200 (m/z). The samples were analyzed in a randomized order.

2.3.7 Database searching and peptides quantification.

The raw files were firstly searched against the human gut microbial IGC (integrated gene catalogue) database[6] using Metalab 2.0[65] to generate a refined fasta database. The *E. coli* database was downloaded from the Uniprot by downloading all protein fasta sequences of *Escherichia coli* (strain K12). A combined database of the two databases above was generated which was used for the open search by pFind 3.0[82] for both ^{15}N and background peptide identification. Enzyme specificity was assigned to trypsin with a maximum of two missed cleavages allowed. The precursor and fragment ions mass

tolerance were set to 20 ppm. Carbamidomethylation of cysteine was set as a fixed modification. The false discovery rate (FDR) for peptide identifications was controlled below 1% using the target-decoy approach.

The quantification was performed by FlashLFQ[83] with and without MBR. Isotope PPM tolerance was set at 5. At least 2 isotopes were required. The maximum window for MBR was 2.5 mins. The peptides intensities used for analysis were raw intensity without any normalization or scaling.

2.3.8 Functional analysis.

All identified proteins were blasted against the latest Clusters of Orthologous Genes (COG) database[84]. We carried out compositional analysis using an R Shiny app (https://shiny.imetalab.ca/playtable/) and calculated the Bray-Curtis distance using the R package “vegan”. Briefly, the identified proteins were annotated to their COG category. The intensity of the COG category in each sample was scaled by percentage by the total intensity. The averages of three replicates were used to represent the COG category percentage. The distance was calculated using the percentage between the 100% *E. coli* sample and the other samples.

2.4 Result and discussion

2.4.1 Experiment overview.

Human gut microbiomes are complex systems composed of hundreds of microbial species[5]. Metaproteomics has been used to better understand the biology of the gut microbiome. Nevertheless, metaproteomics principally provides information on the more

abundant species representing the bulk of the biomass. Here we systematically evaluated the depth and coverage of the gut microbiome that can be achieved by DDA-based metaproteomics with 1-hour nano-LC analysis. Briefly, a ^{15}N-labelled *E. coli* proteome was spiked at different percentages (w/w) into a background peptide mixture either directly from human stool or a microbiome isolate following *in vitro* culture for 24 h[32] (Figure 2.1). The mixtures were analyzed by HPLC-ESI-MS/MS. The peptides were then identified for ^{15}N *E. coli* or the background human microbiome using pFind 3.0 in ^{15}N or ^{14}N mode. Unlike studies based on sequencing technologies which provide an indirect measure of cell numbers[85], MS-based proteomics directly provides protein biomass information[32]. So, the intensity of ^{15}N peptides is expected to be correlated with its quantity in the peptide mixture.

From the pure ^{15}N-labelled *E. coli*, on average, 22,148 ^{15}N labeled peptides (identification rate: 67%, Figure 2.2a) were found across 3 replicates, while only around 1,000 ^{14}N peptides were identified (identification rate: 2.80%). This suggested a high labelling efficiency of the *E. coli* culture. In contrast, 16,000 peptides were identified from the background microbiomes in the absence of ^{15}N-labelled *E. coli* which is comparable to previous studies (Figure 2.2a, and Supplementary Figure 2.2a)[32].

2.4.2 The number of identified peptides was correlated with concentrations of spiked-in labelled peptides.

As expected, as the concentration of spike-in ^{15}N *E. coli peptide* decreased, fewer ^{15}N-labelled peptides were detected. On average 117 and 173 peptides were identified for ^{15}N-labelled *E. coli* spiked at 0.5% in the two background microbiomes (Figure 2.2a and

Supplementary Figure 2.2a). However, when ^{15}N-labelled *E. coli* was spiked at 0.1% or lower, the identification of ^{15}N-labelled peptides was nearly impossible. The number (logged) of quantified ^{15}N-labelled *E. coli* peptides was linear-correlated with the spike-in percentage (logged) from 0.5% to 100% (Figure 2.2b and Supplementary Figure 2.2b).

Due to the diversity of bacteria size and mass, it is difficult to identify the biomass composition of the human microbiome. However, the gene expression composition can be used as a reference. One sequencing study including 120 healthy adults found that each individual harboured, on average, 186 species-level phylotypes (SLPs)[5] (Supplementary Figure 2.3a). Only around 30 SLPs had an abundance over 0.5% (Supplementary Figure 2.3b). Moreover, those high abundant species (> 0.5%) account for more than 80% of the total abundance (Supplementary Figure 2.3c). That means the gut microbiome is dominated by several very high abundant species (Supplementary Figure 2.4) and most species may represent less than 0.5% of the biomass in the microbiome and are undetectable in DDA mode using the parameters of this experiment. However, in regular experiments, that doesn't mean we will lose all the information of the species whose abundance are lower than 0.5%. Because some of the peptides are shared across many bacteria. The intensity of a shared peptide is contributed to all the species containing the peptide including the low abundant species.

In DDA mode, the mass spectrometer is unlikely to consistently select the same ions for fragmentation across multiple experiments. Fortunately, it is possible to infer the identification of peptides across multiple runs using retention time and parent m/z information using the function match between runs (MBR)[83, 86]. This function is clearly beneficial for lower concentrations of spike-in ^{15}N *E. coli* and, in particular, from 5% to 0.5%

(Figure 2.2c and Supplementary Figure 2.2c). At an *E. coli* spike-in of 0.5%, over 98% of peptides were from transferred peptide-spectrum matches (PSMs). We observed a sudden decrease in the number of transferred peptides when the ^{15}N peptides percentage was below 0.5%. That is likely due to a threshold of the MBR algorithm and the 0.5% spike-in approaches that threshold. Even so, the MBR strategy showed a remarkable effect on microbiome analysis.

2.4.3 Spike-in composition of ^{15}N *E. coli* peptides can be quantified accurately.

In this experiment, the decrease of total peptide intensity is due to the decrease of the number of detected peptides and the decrease of their concentrations. Interestingly, a positive linear relationship was observed between the total ^{15}N peptide intensity versus the spike-in percentage of *E. coli* in the background microbiome (Figure 2.2b). This means that the MS can be used to accurately characterize the biomass contributed by individual bacterial species in a microbiome. The biomass of the same source measured by MS is totally comparable across samples/conditions. Nevertheless, in the regular experiment without the ^{15}N labelling, it would be very hard to assign all the peptides to *E. coli* as many of the peptides would be also present in other species.

We then evaluate the quantification of each peptide. We defined the limit of repeatability (LOR) for each ^{15}N-labelled peptides as the lowest mixture percentage at which it is quantified in at least 2 out of 3 replicates. Peptides observed in only one of the three 100% *E. coli* replicates were considered as false identification and excluded in this analysis (prior MBR, 20,224 peptides remained). Interestingly, there was a strong correlation between the peptide intensity observed in the 100% ^{15}N *E. coli* sample and their LOR

(Figure 2.2d and S2d). Those results showed that high abundant peptides have lower LORs which reinforced the character of DDA. Most peptides had their LORs at 50%. Only a fraction of peptides (121 and 81 for each group) can reach LORs of 0.5%. The distribution of peptides among LOR groups can be partially explained by the distribution of the peptide intensity. The average $\log_2$ intensity of *E. coli* peptides is around 25 which is close to the 50% group (Figure 2.2d). We also found that low-intensity peptides always had high LORs, but not all the high-intensity peptides had low LORs (Figure 2.2d). Some high-intensity peptides have their LORs at 100%. That means one peptide selected at 100% is not necessarily found at 50% even if its intensity is higher than most of the peptides in the 50% group. Even so, the low LOR peptides showed good reproducibility. Most low LOR sequences were shared in both protein backgrounds (Supplementary Figure 2.5).

We selected all the peptides whose LORs could reach 0.5% and analyzed their correlation with *E. coli* percentage. Because the LOR could accept one missing value, we employed a robust regression model which can assign the missing value a lower weight[87]. The result showed that most of the correlations for these peptides had R^2 values > 0.99 (Supplementary Figure 2.6) with the lowest R^2 value = 0.93. Although only a small number of peptides can be detected in the 0.5% spike-in sample, all the identified peptides can be quantified accurately.

2.4.4 The functional composition based on detected peptides is not consistent across all the percentages.

As fewer ^{15}N peptides were found in the low spike-in percentage samples, we questioned whether the microbiome functional profile could be maintained. We annotated the detected protein group to each COG category. We found that as the percentage of spike-in decreased, some COG categories disappeared. And there was significant alteration across samples (Supplementary Figure 2.7). We used Bray-Curtis distance to describe the dissimilarity of COG composition between each sample and 100% *E. coli* sample. The distance was significantly increased when the spike-in percentage was below 50% (Supplementary Figure 2.8).

This result showed that the current DDA-based metaproteomics approach may not achieve sufficient depth to study the functional composition of low abundance species. Of note, when studying taxon-specific functional profiles, approach that we applied enabled accurate between-sample comparisons of functions of higher abundance taxa.

2.4.5 Evaluation of match-between-run.

MBR is a strategy using retention time and mass to charge ratios to transfer PSMs across samples. It is based on the assumption that the samples in the same batch are all similar and their proteins belong to the same proteome. The power of MBR depends on data quality and sample number which vary in different experiments. In this study, the three 100% ^{15}N *E. coli* replicates were an ideal reference for MBR. Using this approach, we noticed that the number of quantified peptides was increased significantly and the bias favouring high-intensity peptides reduced (Figure 2.3a). The number of peptides with LOR

of 0.5% increased 50 folds (stool: from 121 to 5511 and cultured: from 81 to 4146). These results suggested that MBR can greatly improve the performance of DDA by mining the information in MS1. It also revealed that peptide features are present in MS1 even for a bacterium that only represents 0.5% of the biomass. So, in metaproteomics experiment from gut microbiomes, there are large number of peptides observed in MS1 only within an MS experiment. Therefore, alternative strategies for the selection of peptides for MS2 analysis might be invaluable for deeper metaproteomics.

To validate this point, we further evaluated how many transferred peptides were reliable. When not using MBR, all the identified peptides had good linear relationships with *E. coli spike-in* percentage (Supplementary Figure 2.5). If the transferred peptides were selected correctly, they should also have a good linear relationship with *E. coli* spike-in percentage. We did linear regression using transferred peptides whose LOR was at 0.5% with *E. coli* spike-in percentage. The R^2 values distribution were not as good as those without MBR (Figure 2.3b). But to our delight, if we use 0.93 (the lowest R^2 value when not using MBR) as the threshold, 2294 (stool) and 1512 (cultured) transferred peptides have good linear relationships (Supplementary Figure 2.9 and 2.10) between their intensity and the spike-in percentage indicating that they were likely proper matches. For the 0.5% spike-in sample, the features added by MBR represents a 20-fold increase in quantified peptides from 121 to 2294(stool) and 81 to 1512(culture), respectively. This means that the MBR with proper quality control could be an invaluable approach for deeper metaproteomics.

We further tested whether a MBR based strategy could be developed to go deeper into the human gut microbiome by focusing on specific reference bacterial strains. To test this strategy, we performed a proof-of-concept study by performing three consecutive DDA

analysis of two bacterial strains (*Blautia hydrogenotrophica*, DSM10507 and *Bacteroides uniformis*, ATCC84892) and a human gut microbiome. First, the DDA MS/MS analysis of the human gut microbiome identified 16488 peptides of which 10 and 15 were from *Blautia hydrogenotrophica*, DSM10507; *Bacteroides uniformis*, ATCC84892. Interestingly, the MBR transferred another 9597 peptides from the individual analysis of the two bacterial strains to the human gut microbiome analysis (Supplementary Figure 2.11). Although not all the transferred peptides were reliable, we showed the potential of this strategy using the reference samples that are of interest to researchers to increase the quantified peptides.

2.5 Conclusion

Species representing less than 0.5% of the biomass of a microbiome are not likely to be detected using top N based DDA approaches in a 1-hour nano-LC metaproteomic experiment. Although by tweaking parameters, such as gradient time, target value, resolution and Boxcar, the DDA method could increase the number of identified peptides, it would be marginal increases in the number of identified species as the top N DDA experiments are biased towards higher abundance species. We found that although fewer peptides were found from low abundant bacteria, all the identified peptides could be quantified accurately.

Gut microbiome samples are very complex with huge dynamic range on each level, and it is likely that peptides from low abundant species are present in MS1 but are not prioritized for MS2 analysis in DDA experiments. Therefore, alternative approaches to top N DDA are needed for deeper metaproteomic analysis. We demonstrated such strategy

using MBR to match MS1 features between metaproteomics of human gut microbiomes as well as between specific bacterial strains and human gut microbiomes. We found that MBR can significantly increase the number of quantified peptides with half of them appearing to be correct matches and therefore recommend turning on MBR whenever possible for metaproteomics. This proof-of-concept experiment establishes that MBR is a promising strategy to increase quantification for low abundant bacteria. It also establishes that low abundant bacteria are predominantly represented in MS1 spectra and not MS2 spectra during DDA analysis of human gut microbiome. Therefore, further strategies to characterize these low abundant species are needed.

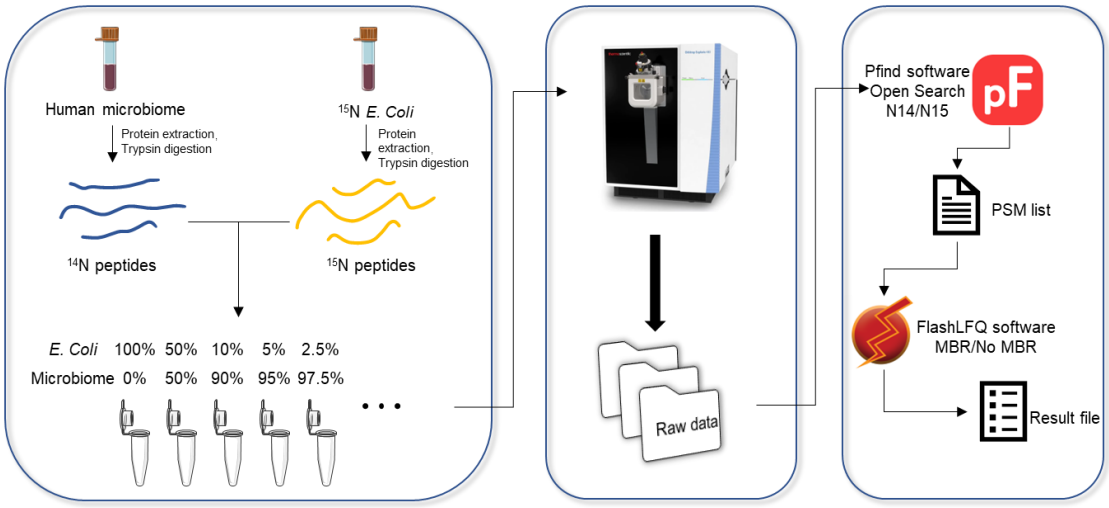


Figure 2.1 Experiment setup and identification profile.

a. Experiment workflow. *E. coli* was cultured in ^{15}N medium to label its proteins. Then the ^{15}N-labeled proteome was extracted and digested into peptides. ^{15}N peptides were mixed at different percentage with peptides from a human gut microbiome. The mixture was subjected to mass spectrometer analysis. Peptide identification was performed using

Pfind3.0 with open search in ^{14}N mode or ^{15}N mode. FlashLFQ was used to do quantification with or without MBR.

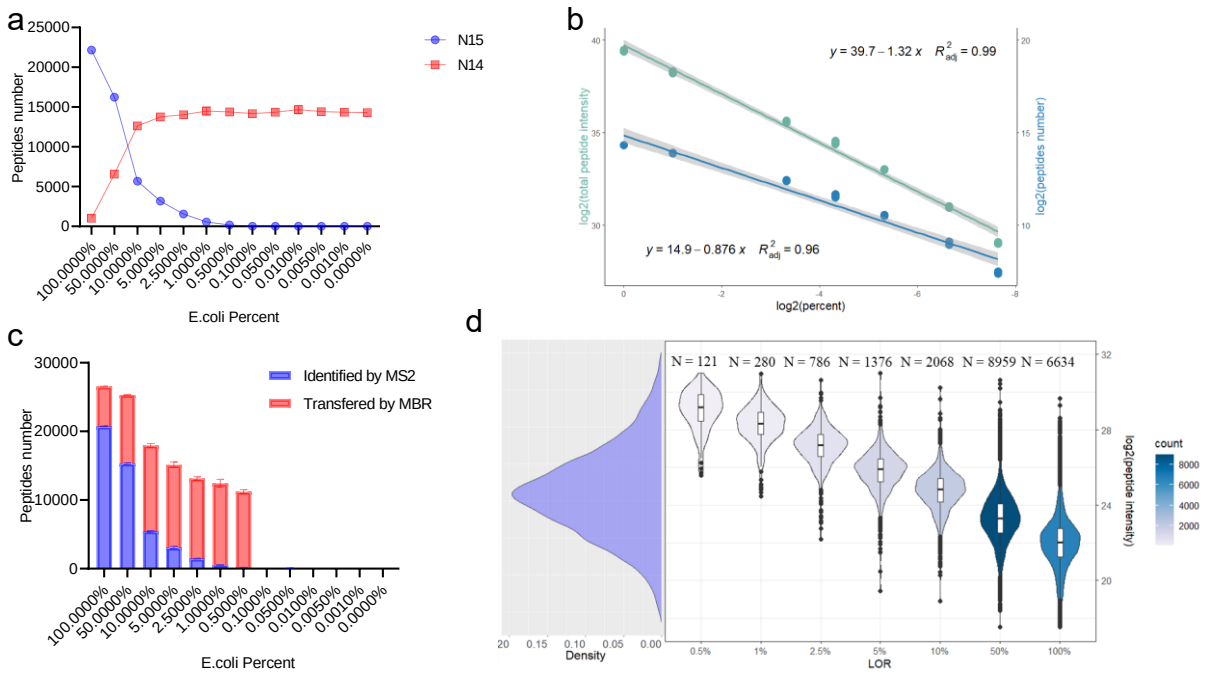


Figure 2.2 Performance of DDA method for peptide identification and quantification with the stool protein as background.

a. The number of identified peptides in each sample. b. Linear regression between log2 quantified peptide number and log2 of ^{15}N peptide spike-in percentage (blue). Linear regres-sion between log2 total peptide intensity and log2 of ^{15}N peptide spike-in percentage(green). c. The number of quantified peptides in each sample. Peptides identified directly by database searching (Pfind 3.0) are shown in blue. Peptides transferred by MBR (FlashLFQ) are shown in red. d. Density plot of the intensity of all quantified ^{15}N peptides (left.). Distribution of LODs for peptide intensity (right). The figures

above each LOD group represent the number of peptides in each group, which is also displayed by the color scale.

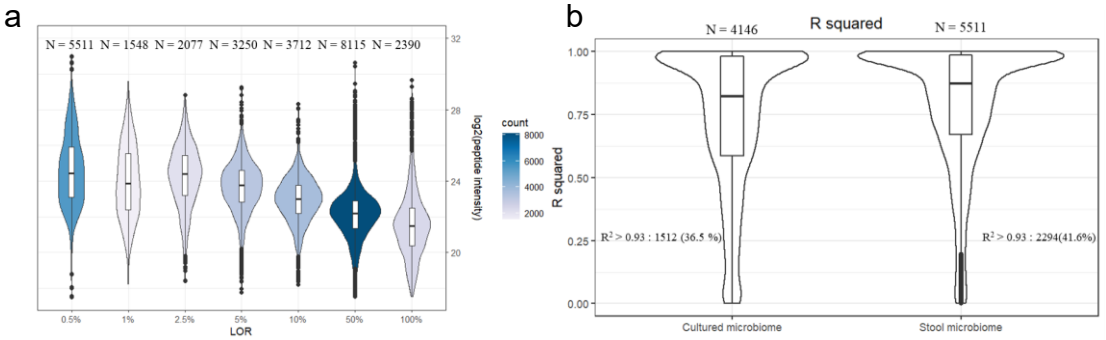
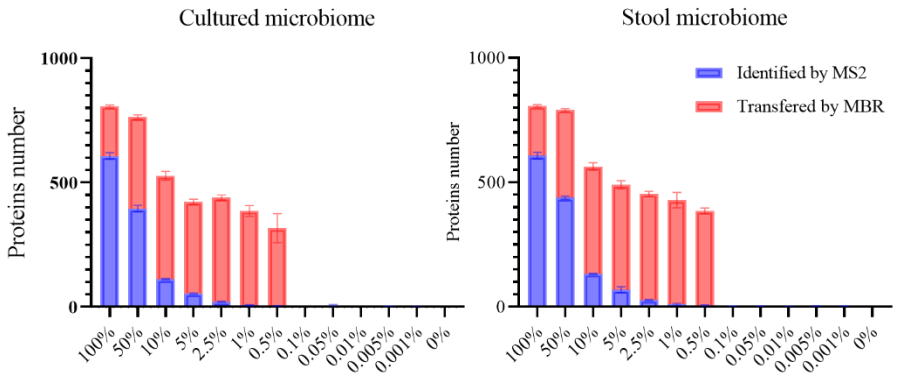


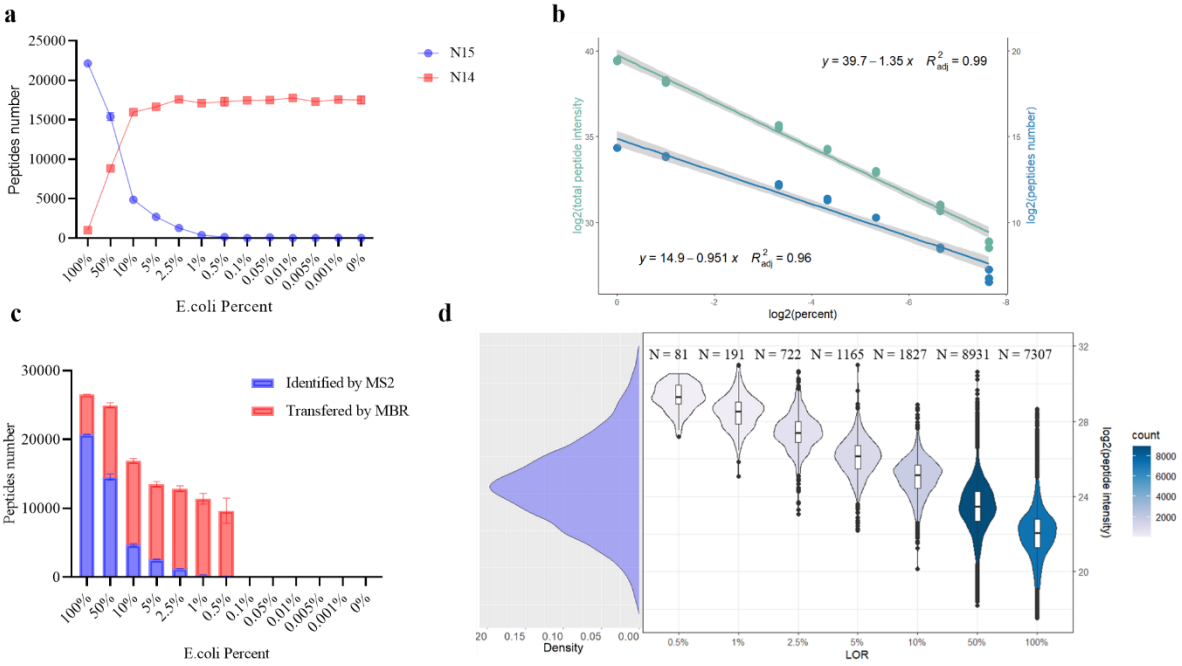
Figure 2.3 Performance of DDA method on peptide quantification with match-between-run.

a. Distribution of LOD on peptide intensity stool microbiome group. The figures above each LOD group represent the number of peptides in that group, which is also displayed by the color scale. b. R value from robust regression between log2 peptide intensity and log2 sample percentage of peptides whose LOD were at 0.5% when performing MBR. The figures above represent the number of peptides in each group. The figures on both sides represent the number (percent) of peptides whose R value is larger than 0.93.



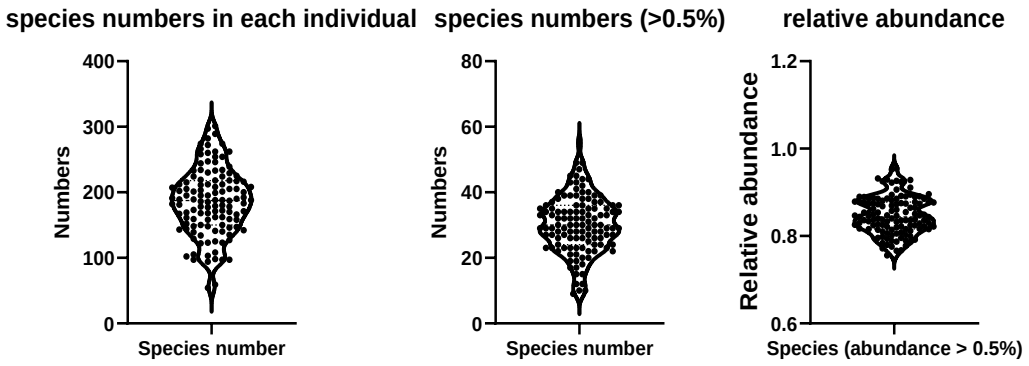
Supplementary Figure 2.1 Number of identified N15 protein groups.

Number of identified N15 protein groups along *E. coli* concentration with or without match-between-run in cultured microbiome group (a) and stool microbiome group (b).



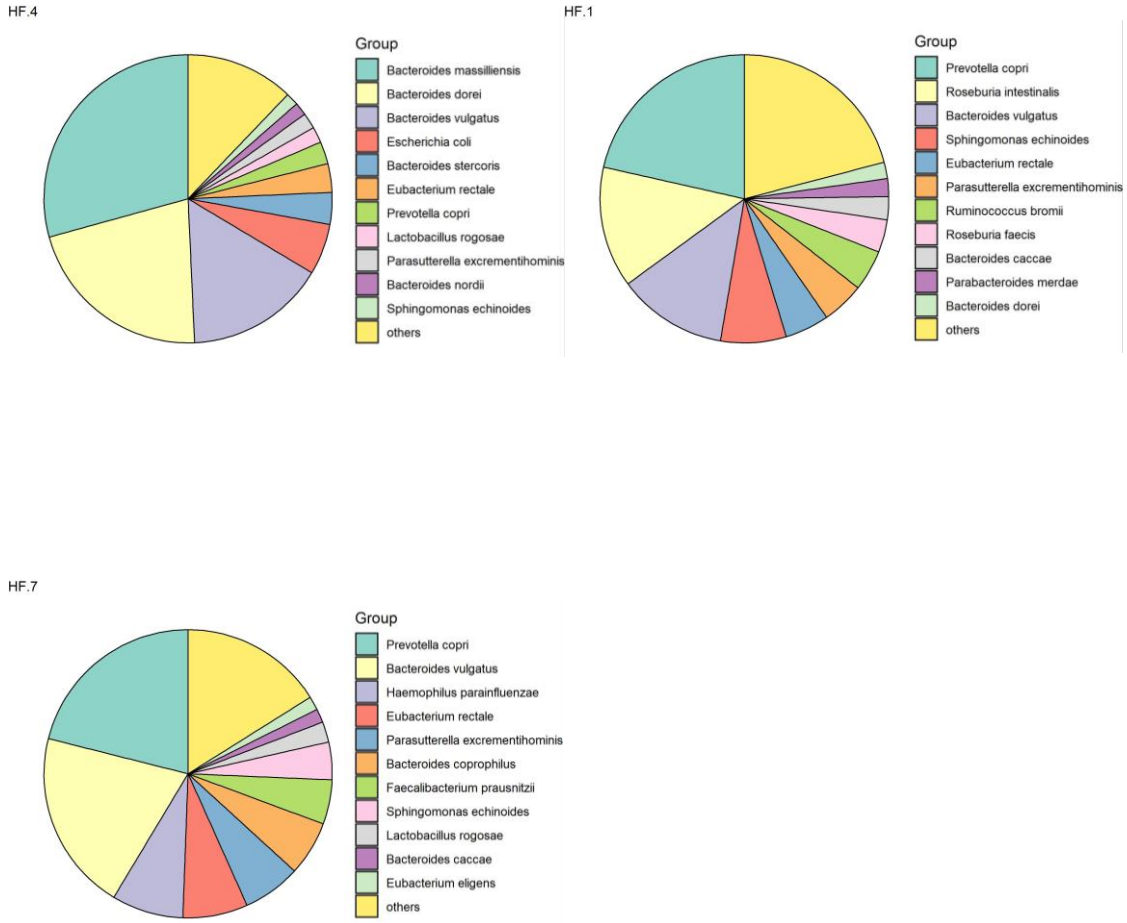
Supplementary Figure 2.2 Performance of DDA method on peptide identification and quantification with the cultured protein as background.

a. The number of identified peptides in each sample (Stool). b. Linear regression between log2 quantified peptide number and log2 percentage (blue). Linear regression between log2 total peptide intensity and log2 percentage (green). c. The number of quantified peptides in each sample. Peptides identified directly by database searching (Pfind 3.0) are shown in blue. Peptides transferred by MBR (FlashLFQ) are shown in red. d. Density plot of all quantified ^{15}N peptides on peptide intensity (left.). Distribution of LODs on peptide intensity (right). The figures above each LOD group represent the number of peptides in each group, which is also displayed by the color scale.



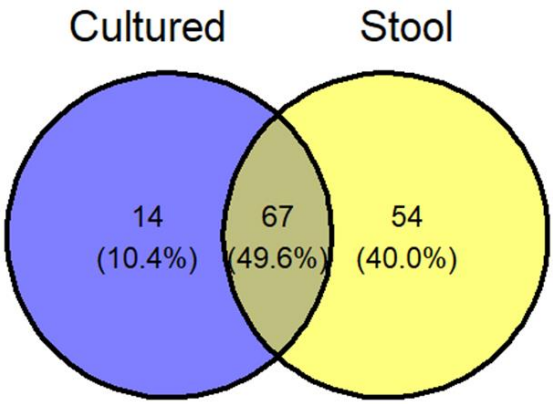
Supplementary Figure 2.3 Abundance distribution of gut microbiome.

a. Species number in each individual. b. The number of species whose abundance over 0.5%. c. Relative abundance of species whose abundance is over 0.5%. The analysis was based on previously published metagenomic data.

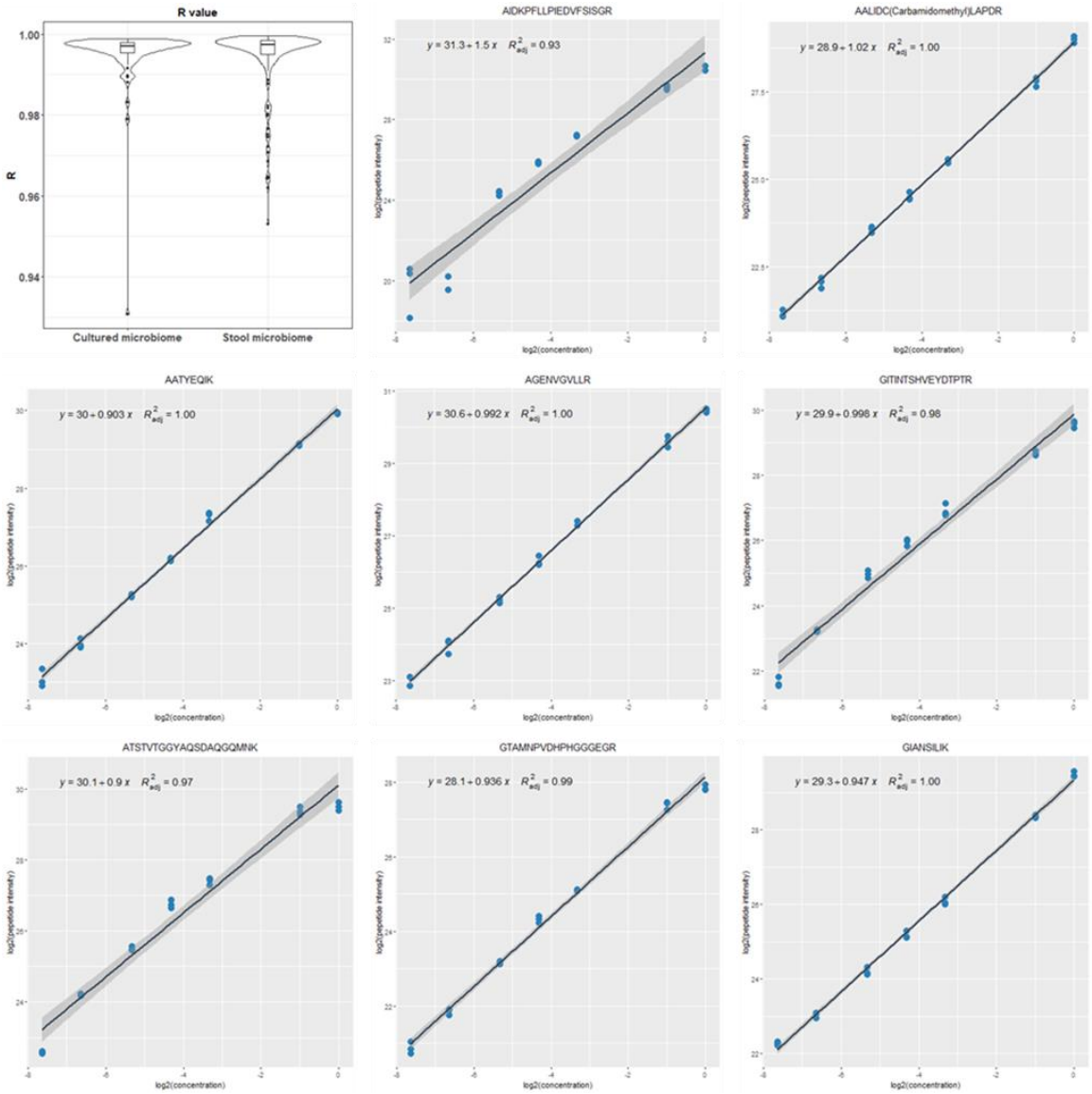


Supplementary Figure 2.4 Representative microbiome composition.

The three representative microbiome composition plots are based on the previously published metagenomic data. The pie charts were built by the species-level taxonomies. Sample indexes from original data are shown in the top left corner.

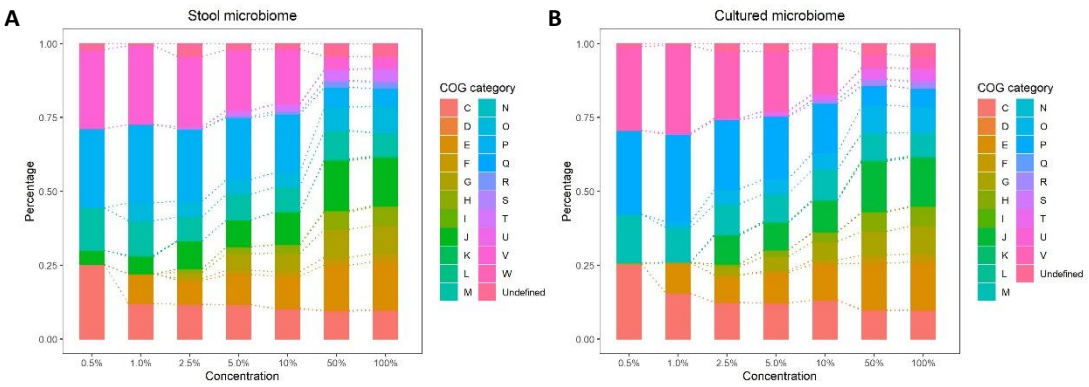


Supplementary Figure 2.5 Intersected peptides with LOR at 0.5% between two groups.



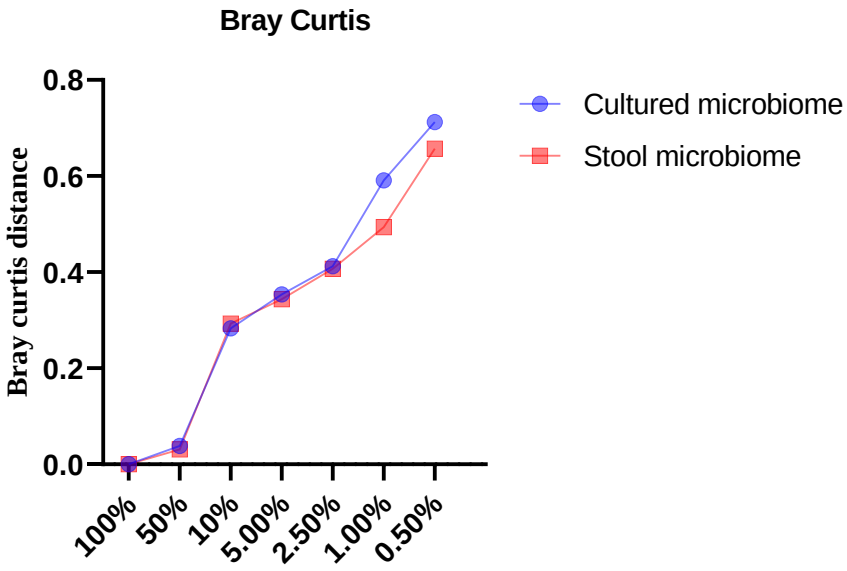
Supplementary Figure 2.6 R^2 value from robust regression.

R^2 value from robust regression between log2 intensity of peptides whose LORs were at 0.5% and *E. coli* percentage in mixtures (The first figure) and representative linear regression plot when not performing MBR. (The other eight figures, the first four are from the cultured microbiome group; the last four are from the stool microbiome group).

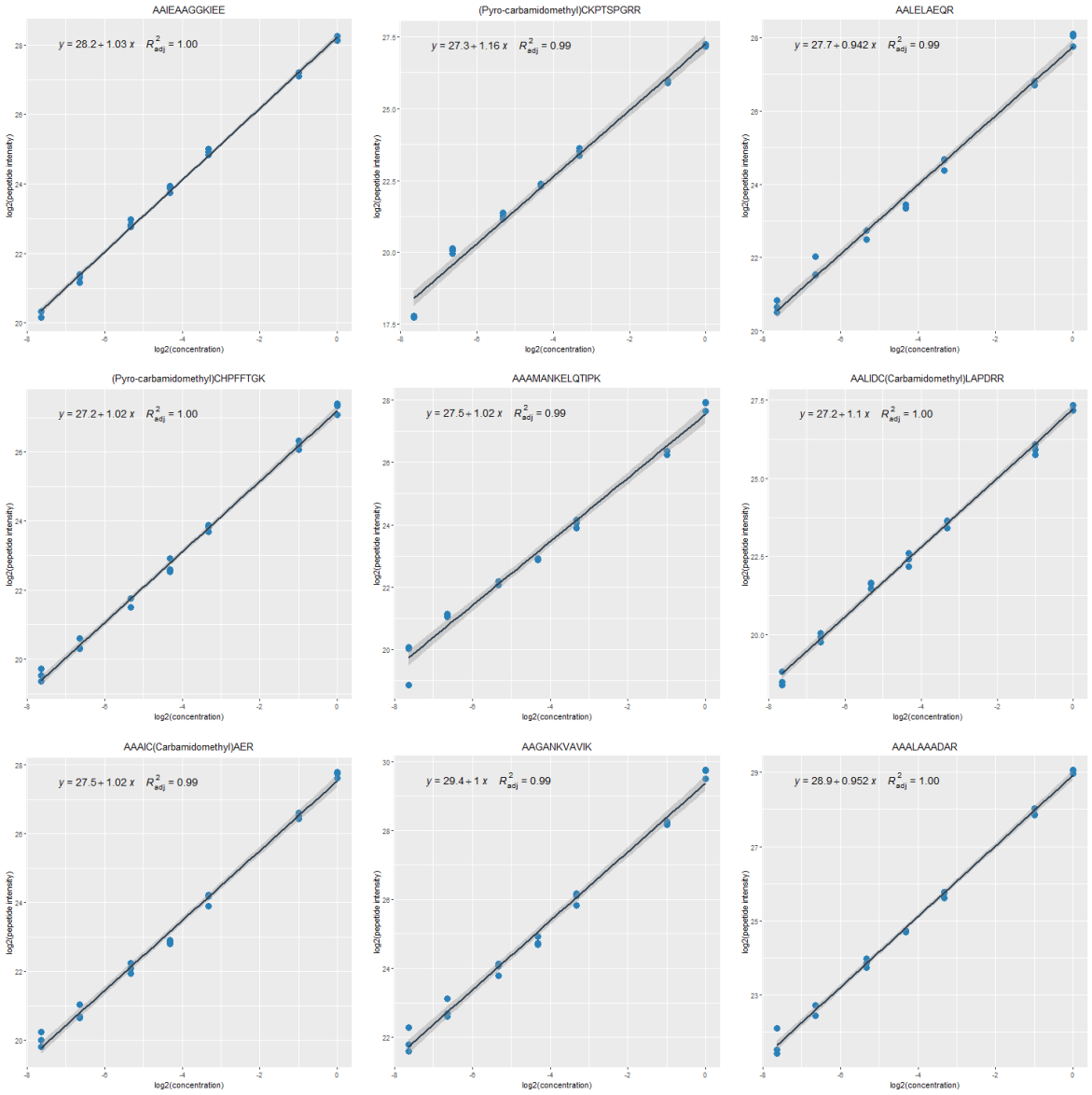


Supplementary Figure 2.7 Evaluation of taxonomy and function analysis.

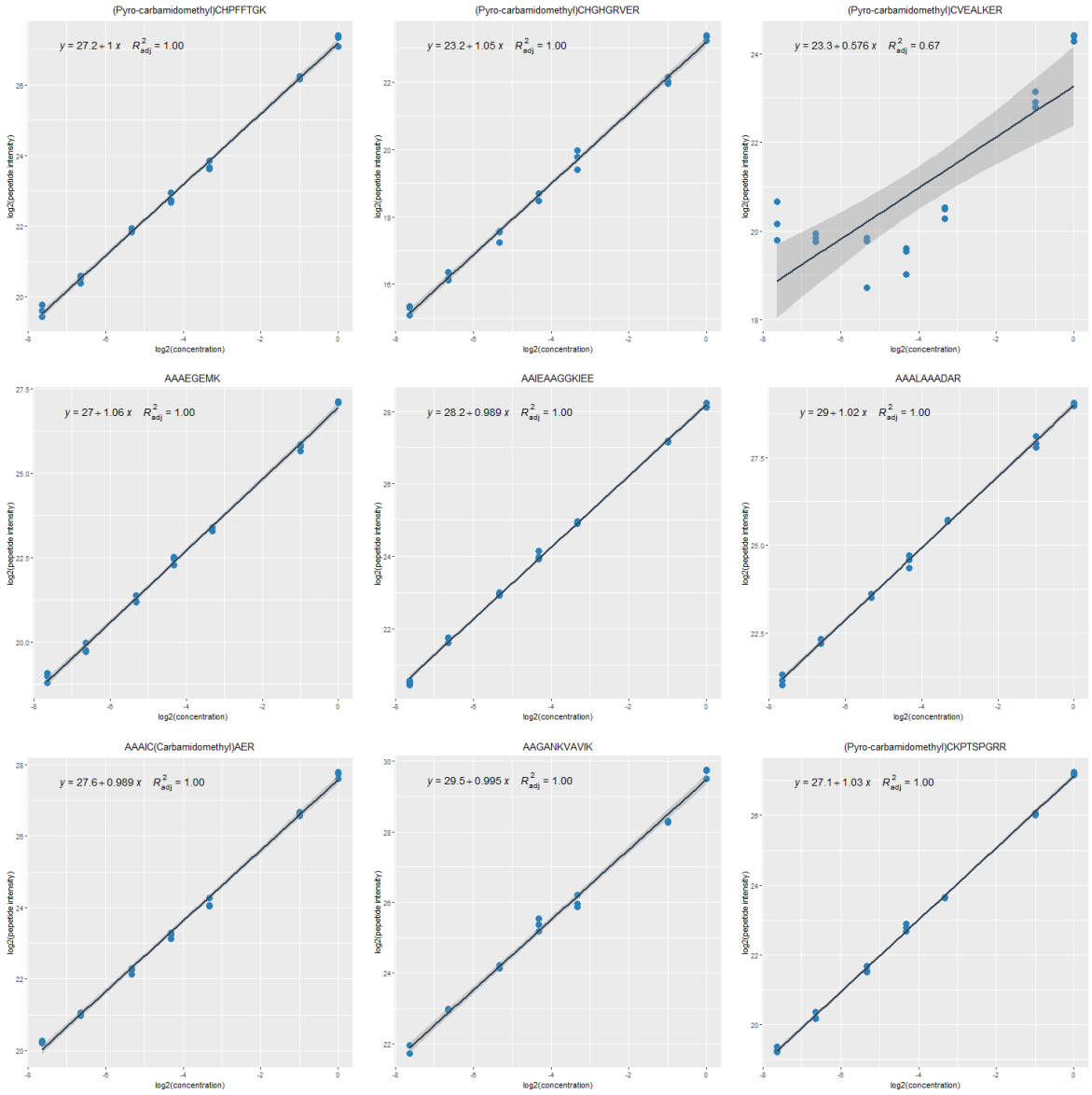
The COG category composition along with *E. coli* concentration (Stool microbiome: A; Cultured microbiome: B).



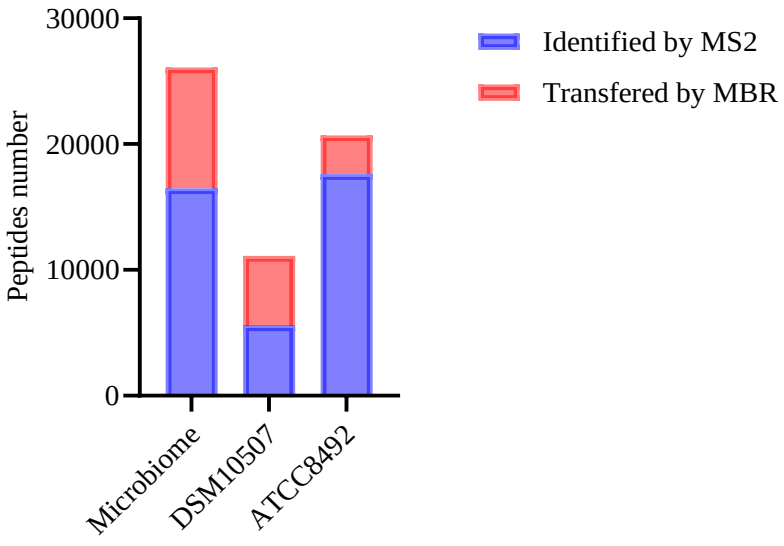
Supplementary Figure 2.8 Bray-Curtis distance between each and 100% *E. coli* on COG composition.



Supplementary Figure 2.9 Representative well-fitted linear regression plot from cultured microbiome group when using MBR.



Supplementary Figure 2.10 Representative well-fitted linear regression plot from stool microbiome group when using MBR.



Supplementary Figure 2.11 The number of quantified peptides.

Peptides identified directly by database searching are shown in blue. Peptides transferred by MBR are shown in red. DSM10507: *Blautia hydrogenotrophica*; ATCC84892: *Bacteroides uniformis*.

Chapter 3. Spectral entropy as a measure of the metaproteome complexity

Preface

In chapter 2, our research substantiated that current DDA-based metaproteomics exhibits a bias that favors species of high abundance. Moreover, when the abundance of a single species in the microbiome falls below 0.5%, it becomes undetectable. This indicates that the detection depth of DDA-based metaproteomics is inherently limited. The question then arises: what causes this limitation? Is it attributable to insufficient instrument sensitivity? If the limitation does stem from instrument sensitivity, then employing DIA would not enhance detection depth, given that the original data lacks information on species of low abundance. In this study, we employed spectral entropy as a tool to directly measure the volume of information in the original spectra, providing a feasible basis for deepening detection depth using DIA mode in subsequent steps.

The following chapter consists of a research article for which I am the primary author. The article was published on *Proteomics*.

H. Duan, Z. Ning, A. Zhang, D. Figeys, Spectral entropy as a measure of the metaproteome complexity, *Proteomics* 2024:e2300570.

HD and ZN developed the concept and designed the experimental outline. HD and AZ performed the experiment. HD performed bioinformatics analysis. ZN provided conceptual advice on data analysis. HD, ZN and DF wrote the manuscript.

3.1 Abstract

The diversity and complexity of the microbiome's genomic landscape are not always mirrored in its proteomic profile. Despite the anticipated proteomic diversity, observed complexities of microbiome sample are often lower than expected. Two main factors contribute to this discrepancy: limitations in mass spectrometry's detection sensitivity and bioinformatics challenges in metaproteomics identification. This study introduces a novel approach to evaluating sample complexity directly at the full mass spectrum (MS1) level rather than relying on peptide identifications. When analyzing under identical mass spectrometry conditions, microbiome samples displayed significantly higher complexity, as evidenced by the spectral entropy and peptide candidate entropy, compared to single-species samples. The research provides solid evidence for the complexity of microbiome in proteomics indicating the optimization potential of the bioinformatics workflow.

3.2 Introduction

The microbiome is a complex system, its complex genomic landscape highlights its inherent diversity[8, 88]. Intuitively, it might be expected that the proteomic profile of the microbiome would mirror this complexity and diversity. However, the observed proteomic complexity often does not meet this expectation. In some instances, the number of peptides identified from the microbiome is even fewer than that from a single species (Supplementary Figure 3.1a). This discrepancy can be attributed to two primary factors:

1) Limitations in Mass Spectrometry's Detection Sensitivity

The observed complexity by mass spectrometry of the metaproteome is often not greater than that of single species, due to functional redundancy[10] and the prevalence of shared

peptides. These shared peptides, collectively expressed across species, dominate the mass spectrometry detection in data-dependent experiments, analogous to the high abundance of albumin in plasma samples. Consequently, without enrichment procedures, detection of less abundant, species-specific peptides becomes challenging, potentially reducing the apparent complexity of microbiome samples in mass spectrometry analysis due to the limit on sensitivity of the mass spectrometer.

2) Bioinformatics Challenges in Metaproteomics Identification

Bioinformatics is one of the most challenging aspects of metaproteomics[45]. The huge size of the database used in metaproteomics increases the probability of random matches. Given that proteomics employs strategies like false discovery rate and decoy databases to filter false positives[89], the vast number of false positives can overshadow true matches, resulting in a low identification rate. Although methods that reduce the database size have significantly alleviated this issue[62, 63, 65-67], the size of the trimmed microbiome database remains hundreds of times larger than that of a single species database, indicating room for further optimization in bioinformatics. Besides, the accuracy of metagenomic sequencing will directly influence peptide identification.

Comparing the complexity of samples directly based on the number of peptide identifications is biased because of the differences in the bioinformatics workflows. The choice of database for searching profoundly influences the results of peptide identification[90]. Consequently, it is inappropriate to assert that a higher number of identified peptides corresponds to a greater diversity of peptide species within the sample. Therefore, we propose a method for comparing the complexity of single organism

samples and microbiome samples that does not rely on peptide identification and utilises instead the information available in MS1. In this approach, we evaluate the sample complexity directly at the mass spectrum level. The MS1 mass spectrum provides a deeper profile of the peptides present in the sample, giving valuable information about peptide ions, which is an ideal material for our comparative analysis. Spectral entropy, borrowing the concept of entropy from information theory, can be utilized to assess the complexity and informational content of a mass spectrum. A high spectral entropy value indicates a more uniform distribution of signals within the mass spectrum, signifying the presence of a diverse array of molecules in the sample; conversely, a low spectral entropy value may suggest that the signals are concentrated on a few ions, potentially indicating high sample purity or the analysis of a singular molecule.

In the present study, we conducted a comparative analysis of the human gut microbiomes and single strains at the MS1 level. To evaluate the complexity of the samples, we quantified various parameters, including the number of peaks in the spectra, spectral entropy, and the count of peptide features. Furthermore, we developed a computational tool to facilitate the calculation of spectral entropy. Our findings illustrate that spectral entropy is an effective and straightforward metric for assessing the complexity of microbiome samples.

3.3 Materials and methods

3.3.1 Sample preparation

The human stool was collected from 5 healthy adult volunteers at the University of Ottawa, Ottawa, ON, CAN. The protocol (# 20160585-01H) was approved by Ottawa Health

Science Network Research Ethics. The human microbiome was cultured in an optimized medium[32]. The strain samples were cultured in broth medium. Both cultures were under 37 °C for 24 h. Protein extraction and digestion were based on previously published method[91]. Briefly, cell pellets were harvested by centrifugation. Then bacterial cells were subjected to lysis in a specified buffer (6M urea, 50mM Tris-HCl, 4% SDS, Protease Inhibitor 1x), followed by ultrasonication at 8 °C. The lysate was then centrifuged at 16,000g to discard debris, and proteins were precipitated overnight using a cold acetone/ethanol/acetic acid mixture. The proteins were pelleted and washed with acetone before dissolving in urea and ammonium bicarbonate, with concentration determined by the DC assay. A sample of 50 µg protein underwent reduction, alkylation, and digestion by trypsin at 37 °C for 24 h. The resulting peptides were then desalted using a C18 column and, after freeze-drying, redissolved in 0.1% formic acid. For the experiment investigating the influence of loading amount on spectral complexity, the peptide concentration was determined using a quantitative peptide assay (Thermo Fisher Scientific, USA, catalog number 23275).

3.3.2 LC-MS/MS analysis

An UltiMate 3000 RSLCnano system was used with an Orbitrap Exploris 480 mass spectrometer for sample analysis (Thermo Fisher Scientific, USA). Peptides were loaded onto a column with reverse phase beads, followed by a 60-minute gradient from 5 to 35% buffer at a rate of 300 µL/min. For data-dependent acquisition experiment, the mass spectrometer was set to a top15 method, dynamic exclusion, and a full mass scan range of 350 to 1200 (m/z). Dynamic exclusion repeat count was set to one and repeat exclusion duration of 20 s. Peptides are fragmented using higher-energy collisional dissociation

(HCD) with a normalized collision energy of 30%. MS/MS scans were acquired at a resolution of 15,000 at m/z 200. For data-independent acquisition (DIA) experiment, m/z range is 380-980 with window widths set at 10 m/z.

3.3.3 Data processing

Raw files from single-species samples were searched against corresponding assembled genomes using Maxquant version 1.5.2.8[92]. The theoretical protein databases for five bacterial strains were retrieved from the National Center for Biotechnology Information (NCBI). The respective genome assembly identifiers for these bacteria are ASM15797v1, ASM16903v1, ASM1282v1, ASM2514748v1, and ASM15637v1. Raw files from microbiome samples were searched against the human gut microbial integrated gene catalogue database using Maxquant workflow in Metalab version 2.2[65]. We achieved satisfactory identification rates from all the samples, ranging between 23% and 34% (Supplementary Figure 3.1a). For the consecutive injections experiment, *Blautia hydrogenotrophica* was searched against corresponding strain database. The microbiome samples were searched against MetaPep, a peptide database specific to human gut microbiome samples[59]. The peptide identification for DDA and DIA samples was performed on Maxquant (Version 1.5.2.8) [92] and DIA-NN 1.8[79] under library free mode respectively.

The raw data generated by the mass spectrometer was converted from profile data into centroid data in the ms1 and mzML formats using MSconvert Version 3.0. An absolute intensity of 5000, which represents the lowest intensity to trigger an MS/MS scan, was employed as the threshold to filter out noise. The “feature finder centroided” algorithm in

TOPPAS version 3.0 was used to search for peptide candidates from mzML files using default parameters[93, 94]. The ms1 files were used to calculate spectral entropy using "Spectral Entropy Calculator". The calculation method for spectral entropy is illustrated by the following equation (*p*: peaks, the intensity was extract from raw data using MSconvert Version 3.0):

$$S = -\sum_p I_p \, ln \, I_p \quad (1)$$

$$Normalized\ S = -\frac{\sum_p I_p \, ln \, I_p}{\ln(number\ of\ p)} \quad (2)$$

$$I_P = \frac{Intensity\ of\ p}{Total\ intensity} \quad (3)$$

3.4 Results

Briefly, we performed LC-MS/MS analysis of five individual gut microbiome samples and five single strains. Around 4209-4882 MS1 spectra were collected per sample (Supplementary Figure 3.1b). The number of spectra collected per sample did not vary between single species and microbiome samples. We then counted the number of ion peaks with intensity exceeding 5000 (the lowest intensity to trigger MS2 scan) in each MS1 spectrum. The average peak number in all microbiome sample was significantly higher than that in single-specie samples (Figure 3.1a), which indicated higher complexity of microbiome sample. To further evaluate the amount of information contained in each sample, we employed spectral entropy. In information theory, entropy is used to quantify the amount of uncertainty or randomness associated with a set of data. The spectral entropy has been successful used for library matching in metabolomics[95, 96]. We developed a tool with GUI called "Spectral Entropy Calculator" to compute the spectral

entropy of mass spectra (https://github.com/northomics/spectra_entropy). As expected, the MS1 spectra from microbiome samples had significantly higher spectral entropy (Figure 3.1b). The spectral entropy achieves its maximum at ln(number of peak) when all ions exhibit identical intensity. Given that spectral entropy could increase with the number of peaks, we also compared the normalized spectral entropy by dividing the spectral entropy by ln(number of peak) which ranges from 0 to 1 (Equation 2). Once again, the microbiome samples displayed a higher normalized spectral entropy (Figure 3.1c). This indicates a more uniform distribution of ion intensities and greater diversity of ion species in microbiome samples. Consequently, this suggests that more information could be translated from the spectra of microbiome.

Increasing the loading amount might enable the detection of low-abundance peptides that were previously undetectable, potentially increasing the spectral complexity. We investigated the impact of the loading amount on peak numbers, spectral entropy, and normalized spectral entropy. One microbiome sample and a *Blautia hydrogenotrophica* sample were loaded with quantities ranging from 500 ng to 2000 ng in three replicates. Interestingly, the result shows that the spectral entropy and peak number remains relatively stable as the sample loading amount increases (Supplementary Figure 3.2). This indicates that within a certain range, spectral entropy is not affected by the sample loading amount, thereby accurately reflecting the complexity of the sample. In proteomics, feature detection algorithms play a crucial role by identifying potential peptide candidates[97]. These algorithms do not rely on theoretical protein databases. Instead, they directly search for peptide candidates in the MS1 spectra based on retention time, chromatographic peak shape, and isotopic patterns[94]. Using the same criteria, we

employed feature detection algorithms to search for peptide candidates in both microbiome and single-species samples. Interestingly, microbiome samples contained a larger number of peptide candidates, ranging from 45,000 to 61,000 (Figure 3.2c). However, in single-species samples, the number of peptide candidates never exceed 41,000 in the five species. The feature detection algorithms also quantified the intensity of candidates, enabling the application of a similar strategy to calculate candidate entropy in which the peptide candidate is similar to the peak in spectral entropy. Our findings reveal that both candidate entropy and normalized candidate entropy are markedly higher in microbiome samples when compared to single-species samples, suggesting a greater complexity and diversity in microbial communities (Figure 3.2a and 3.2b). We normalized the candidate intensity to the total intensity in each sample and plotted candidates in descending order of intensity. Both sets of data follow a normal distribution, displaying a higher number of moderate values and fewer extreme values. Microbiome samples displayed a higher count of candidates with moderate intensities (Figure 3.2c, Area II). Interestingly, we found that the intensity drops more quickly in microbiome for the high abundant area (Figure 3.2c, Area I). This may suggest that the accumulation of shared peptides results in a much higher abundance of these peptide compared to others. Additionally, despite the stability of spectral complexity across a range of loading amounts from 500 ng to 2000 ng (Supplementary Figure 3.2), our analysis indicated a positive correlation between the number of peptide candidates and the loading amount (Supplementary Figure 3.3). This suggests that although the signals generated by peptides are recorded, these peptides must still achieve a certain criterion (chromatographic peak shape, intensity) to be recognized by feature detection algorithm.

The microbiome samples yielded a greater number of peptide candidates compared to *Blautia hydrogenotrophica* samples when loading same amount. Moreover, the difference in the number of candidates between these samples expanded with an increase in the loading amount. In our experiment, we observed that the loading amount of microbiome samples was either comparable to or less than that of the single-species samples (Figure 3.1d), which implies that if an equal amount of peptides were loaded, the differences between the two sample types would be even more pronounced.

We conclude that the metaproteome complexity is reflected in the information in the MS1, entropy calculations provide good assessment of the sample complexity, and that it can detect differences in sample complexity.

3.5 Discussion

In this study, we found that both the spectral entropy and peptide feature that were directed derived from MS1 can reflect the complexity of samples. However, the spectral entropy is relatively unaffected by loading amounts. We observed excellent reproducibility of spectral entropy in each of the three replicate samples, indicating the potential use of spectral entropy as an assessment tool for sample quality. Without peptide identification, our research demonstrates that microbiome samples exhibit significantly higher complexity than single-species samples on MS1 under the current mass spectrometry sensitivity. In mass spectrometry, the DDA mode inherently possesses a degree of selection bias towards higher abundance. This characteristic means that although the number of identified peptides per injection remains relatively consistent, successive injections of a sample tend to yield an increased number of unique peptides[98, 99]

(Supplementary Figure 3.4). On the other hand, DIA operates without such bias. In an experiment involving 21 consecutive injections of the *Blautia hydrogenotrophica* samples (Supplementary Figure 3.4b), the DDA mode continuously revealed new peptides. In contrast, the peptide counts in DIA mode quickly reached a plateau. Interestingly, when this process was replicated with microbiome samples, both DDA and DIA modes consistently showed a rise in peptide identification as the number of injections increased (Supplementary Figure 3.4b). High reproducibility is one of the major advantages of the DIA mode[100]. However, it is not reflected in microbiome sample. This phenomenon may be attributed to the enhanced complexity of the microbiome samples, which potentially magnifies the bias in signal detection and peptide identification. In highly complex samples, numerous peptides may produce signals in the same or close m/z ranges. In DIA mode, although signals are not selectively collected based on ion intensity, these overlapping signals are still acquired simultaneously within the same mass window, forming a composite spectrum. The higher degree of signal overlap in complex samples makes the deconvolution process more complex and challenging, potentially leading to decreased identification accuracy and reproducibility. This suggests that when using DIA for highly complex samples like microbiome samples, it may be necessary to specifically optimize mass spectrometry parameters or employ longer chromatographic gradients to reduce sample complexity.

3.6 **Data availability**

The code and Spectral Entropy Calculator are available at:

https://github.com/northomics/spectra_entropy.

The raw data and searching result can be found at ProteomeXchange Consortium (dataset identifier: PXD050781).

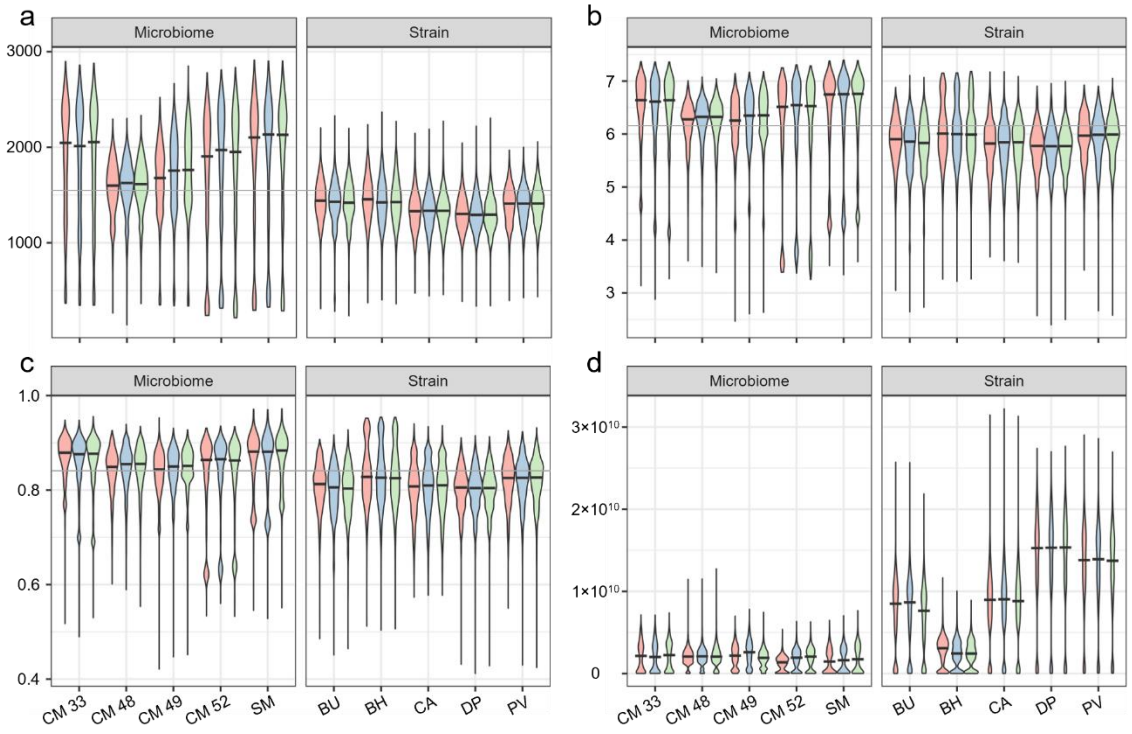


Figure 3.1 Spectra complexity comparison between microbiome and single species.

(a) Peaks number on MS1. (b). Spectral entropy of each MS1. (c). Normalized spectral entropy of each MS1. (d). The total ion current of each MS1. Each sample was injected three times under identical conditions. Three replicates were shown in different color. The black line represents the median values of the parameters within each sample. The gray line represents the boundary between microbiome and strain samples. The equations to calculate spectral entropy and normalized spectral entropy was shown in method. CM: cultured microbiome; SM: stool microbiome; BU: *Bacteroides uniformis*; BH: *Blautia*

hydrogenotrophica; CA: *Collinsella aerofaciens*; DP: *Desulfovibrio piger*; PV: *Phocaeicola vulgatus*

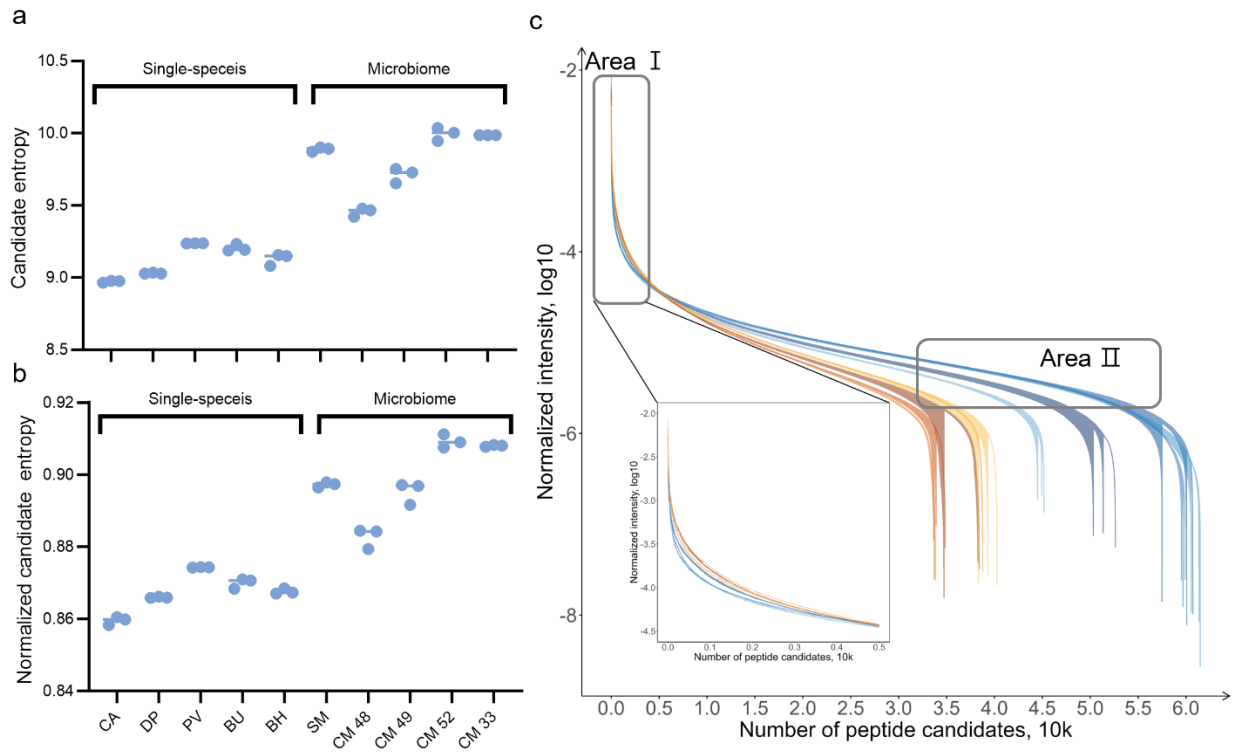
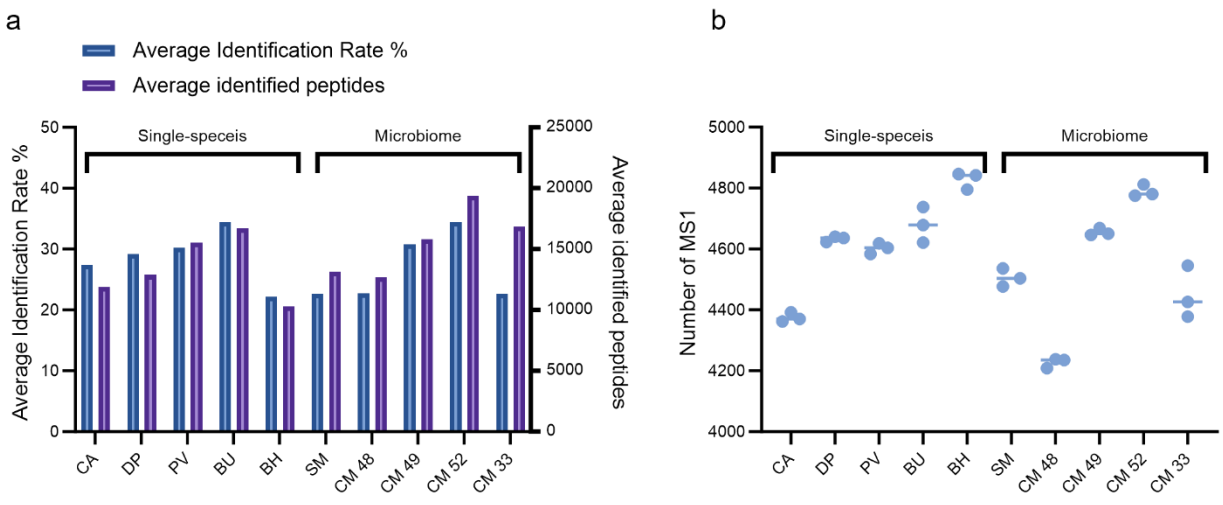


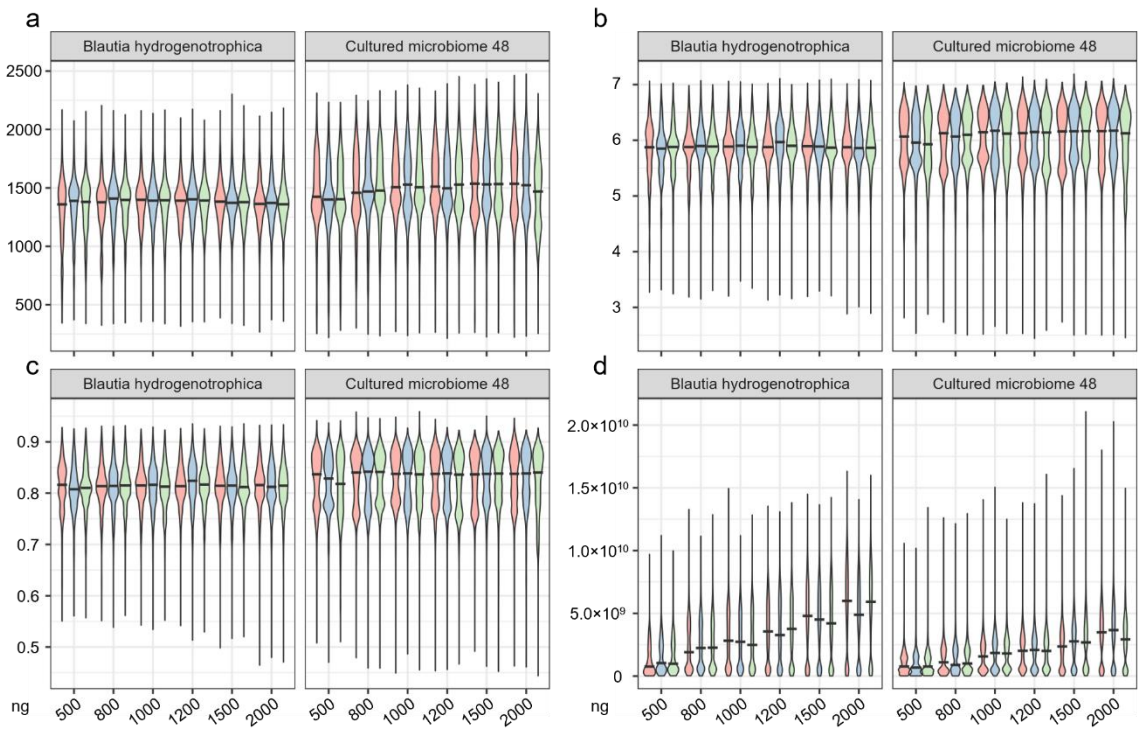
Figure 3.2 Candidates' complexity comparison between microbiome and single species.

(a) The candidate entropy of each sample. (b) The normalized candidate entropy of each sample. (c) The intensity was normalized to the total intensity in each sample and log-transformed (base 10). The intensity was arranged in descending order. Each sample had three replicates which were in same color. CM: cultured microbiome; SM: stool microbiome; BU: *Bacteroides uniformis*; BH: *Blautia hydrogenotrophica*; CA: *Collinsella aerofaciens*; DP: *Desulfovibrio piger*; PV: *Phocaeicola vulgatus*



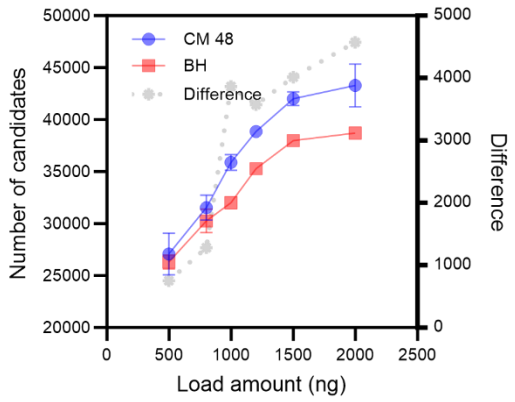
Supplementary Figure 3.1 The quality of data

(a) Peptide identification rate and the number of identified peptide. (b) The number of MS1 sampled in each raw data. Single-species samples were searched against corresponding assembled genomes using Maxquant version 1.5.2.8. microbiome samples were searched against the human gut microbial integrated gene catalogue database using Maxquant workflow in Metalab version 2.2. CM: cultured microbiome; SM: stool microbiome; BU: *Bacteroides uniformis*; BH: *Blautia hydrogenotrophica*; CA: *Collinsella aerofaciens*; DP: *Desulfovibrio piger*; PV: *Phocaeicola vulgatus*



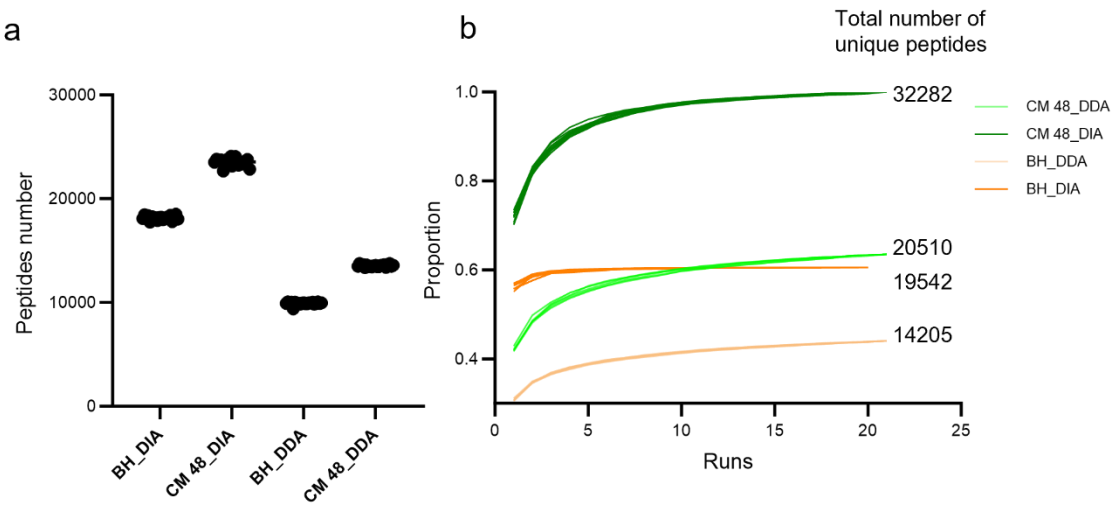
Supplementary Figure 3.2 The impact of loading amount on spectra complexity.

(a) Peaks number on MS1. (b) Spectral entropy of each MS1. (c) Normalized spectral entropy of each MS1. (d) The total ion current of each MS1. Each sample was injected three times under identical conditions. The loading amount was shown on x axis in ng. Three replicates were shown in different colors. The black line represents the median values of the parameters within each sample. The gray line represents the boundary between BH and CM 48. The equations to calculate spectral entropy and normalized spectral entropy was shown in method. CM: cultured microbiome; BH: *Blautia hydrogenotrophica*



Supplementary Figure 3.3 The number of peptide candidates

The number of peptide candidates detected along different loading amount (Blue and red). The discrepancy in mean candidate numbers between the two types of samples (grey). CM: cultured microbiome; BH: *Blautia hydrogenotrophica*



Supplementary Figure 3.4 Comparison between DDA and DIA on single species and microbiome.

(a) The number of identified peptides in each sample. (b) Cumulative number of unique peptides identified across 21 runs with randomly selected sample orders. The curves in each group are plotted ten times in randomly sample orders. The increasing curve is normalized by the number of total unique peptides (32282) in DIA microbiome experiment. The maximum number of identified peptides for each dataset is labeled. CM: cultured microbiome; BH: *Blautia hydrogenotrophica*

Chapter 4. MetaDIA: A Novel Database Reduction Strategy for DIA of Human Gut Metaproteomics

Preface

In Chapter 2, we substantiated that the detection depth of DDA-based metaproteomics is constrained. In Chapter 3, we discerned that the insufficiency in depth is not attributable to the lack of sensitivity in mass spectrometry instruments. On the contrary, the mass spectra generated from microbiome samples are replete with information awaiting interpretation. DIA, as a collection mode capable of unbiasedly capturing signals of both high and low abundance, holds promise for achieving greater depth in metaproteomics. The challenge in applying DIA to metaproteomics lies in the vast theoretical search space. In this chapter, we will develop a novel library reduction strategy to facilitate the application of DIA in metaproteomics.

The following chapter consists of a research article for which I am the primary author. The article was submitted to *Microbiome* and is under review. The full text is also available at bioRxiv

H. Duan, Z. Ning, Z. Sun, T. Guo, Y. Sun, D. Figeys, MetaDIA: A Novel Database Reduction Strategy for DIA Human Gut Metaproteomics, bioRxiv (2024) 2024.03.14.585104.

HD, ZN, ZS, TG, YS and DF designed the study. HD performed the experiments. HD, ZN, DF analyzed the data. TG and YS shared the diaPASEF raw data. HD and DF wrote the manuscript.

4.1 Abstract

Background: Microbiomes, especially within the gut, are complex and may comprise hundreds of species. The identification of peptides in metaproteomics presents a significant challenge, as it involves matching peptides to mass spectra within an enormous search space for complex and unknown samples. This poses difficulties for both the accuracy and the speed of identification. Specifically, analysis of data-independent acquisition (DIA) datasets has relied on libraries constructed from prior data-dependent acquisition (DDA) results. This approach requires running the samples in DDA mode to construct a library from the identified results, which can then be used for the DIA data. However, this method is resource-intensive, consumes samples, and limits identification to peptides previously identified by DDA. These limitations restrict the application of DIA in metaproteomics research.

Results: We introduced a novel strategy to reduce the search space by utilizing species abundance and functional abundance information from the microbiome to score each peptide and prioritize those most likely to be detected. Employing this strategy, we have developed and optimized a workflow called MetaDIA for analysis of microbiome DIA data, which operates independently of DDA assistance. Our method demonstrated strong consistency with the traditional DDA-based library approach at both protein and functional levels.

Conclusion: Our approach successfully created a smaller, yet sufficient database for DIA data search requirements in metaproteomics, showing high consistency with results from

the conventional DDA-based library. We believe this method can facilitate the application of DIA in metaproteomics.

4.2 Introduction

The microbiome encompasses a diverse array of microorganisms residing in different organisms, ecosystems, and environmental settings such as the human body, animals, plants, soil, water bodies, and various ecological niches[88, 101]. Metaproteomics serves as a tool for understanding the roles of proteins within these microbial communities[102]. Mass spectrometry-based proteomics aims to study all proteins in a sample. However, applying these techniques to the microbiome is challenged by its complexity. Without prior knowledge of the microbes present in a sample, metaproteomics relies on searching mass spectra against a large database, making the task of matching peptides and spectra notably challenging. Employing an iterative search strategy significantly reduces the search complexity in which the final search is against a database generated from previous searching results [62, 63]. The iterative strategy has been successfully used but only for the data acquired by data-dependent acquisition (DDA) mode[65, 67], Unfortunately, in DDA mode, only the most abundant precursor ions are selected for further inquiry, and lower abundant ones are overlooked[91].

In contrast, data-independent acquisition (DIA) uses a set of precursor isolation windows to collect all the fragments ions indiscriminately[49]. It has shown remarkable robustness, sensitivity, and reproducibility with fewer missing values[51]. DIA can be coupled with microLC enabling high-throughput analysis[52]. This makes it particularly suitable for conducting large-scale analyses. The DIA-PASEF[53] method integrates ion mobility

separation with the DIA workflow, adding a fourth dimension of analyzing ion mobility to the traditional three-dimensional data set. This not only enriches the structural information of analytes but also enhances ion utilization efficiency leveraging the linear relation between ion mobility and mass-to-charge ratio. Another improvement in mass spectrometer scanning speed enables the utilization of smaller isolation windows in DIA, termed as narrow-window DIA[54]. This approach achieves comprehensive peptide precursor coverage and high quantitative precision and accuracy. In bioinformatics, the development of prediction software for peptide properties (theoretically predicted spectrum[103, 104], retention times[105-107]) enables the querying of DIA datasets without dependence on libraries generated by DDA. Those predicted libraries even showed better performance than the measured libraries[108]. Moreover, DIA-specific searching software such as DIA-NN[79, 109], MaxDIA[110], and Spectronaut[111] have shown reliable results for the identification and quantification of peptides. The above advantages make DIA increasingly popular in proteomics. However, it is noteworthy that the benefits conferred by these techniques have not yet been fully extended to the field of metaproteomics. The main reason is that the inherent complexity of DIA data requires a much more constrained searching space compared with DDA data. To date, only a few metaproteomics studies have been done, and they were all compelled to use a spectral library derived from DDA data[70-72]. The DDA-derived method involves creating a spectral library from DDA runs for each sample, which is then used to interpret complex mass spectra from subsequent analyses. This approach requires multiple sample aliquots, extensive mass spectrometry resources and is limited to detecting peptides previously identified by DDA. Gladiator[76] uses DIA-Umpire[74] to assemble pseudo-DDA spectra

from DIA data for microbiome samples. The method does not require a DDA-based spectral library for its operation, however, it still relies on spectrum-centric algorithms and does not fully exploit the potential advantages of DIA data.

Therefore, to leverage the benefits of DIA in metaproteomics, the searching space needs to be further reduced. In the previous DDA iterative strategy[66, 67], the high-abundant proteins (HAP) were used for the first search to infer the species that exist in the sample then all the proteins belonging to those species were then used for the subsequent search. However, this database remains overly extensive when compared to the number of identified peptides. Since the abundance of species within the microbiome shows significant disparity[5], the species identified should not be considered equally. The same applies to proteins and peptides. Proteins with high abundance and peptides with high detectability[112] or shared among various species are more likely to be detected. Here we report on a DIA workflow for metaproteomics, called MetaDIA, that relies on an annotated peptide database. This database comprises peptides that are anticipated to be detected, leveraging information on species abundance and protein abundance to score each peptide. We conducted a proof-of-concept experiment on human gut microbiome data generated by diaPASEF mode[71]. The peptide identification number and quantitative results obtained through our peptide library are comparable to those from the DDA-based library. Moreover, the species and functional information obtained from both methods are highly consistent.

4.3 Materials and methods

4.3.1 Reference peptide sequence with detectability score for human gut microbiome.

The Unified Human Gastrointestinal Protein (UHGP) catalog, encompassing 4744 assembled genomes from the human gut microbiome, served as the reference database for this study[8]. Within this catalog, each protein sequence is uniquely associated with a distinct genome and is accompanied by detailed taxonomic and functional annotations. The detectability of peptides derived from these protein sequences was predicted using DeepDetect[112], a deep learning algorithm specifically designed for this purpose. This process involved in silico digestion of the protein sequences and subsequent assignment of a detectability score to each resultant peptide. Consequently, the peptide sequence reference database was enhanced by annotating each peptide with three key pieces of information: the genome identifier, the protein identifier, and the peptide's detectability score. Please note that the database is structured on an identifier-centric organization. This means that peptides with identical sequences may be present within the database; however, as long as they are not from same genome and protein, they are distinguished by unique identifiers.

4.3.2 Generation of FuncTax score

Firstly, identified peptides by MetaPep[59] are mapped to the UHGP database to establish peptide-genome associations. Subsequently, a greedy algorithm is employed to identify the minimal set of genomes that encompasses all peptide sequences, effectively reducing the complexity of the dataset. Following this, the intensity of each peptide is aggregated

to infer genome abundance. The relative genome abundance will be used as the taxonomic score. To address the assignment of shared peptides, a razor strategy is adopted, analogous to the MaxQuant approach for protein inference[92]. Specifically, when a peptide is found in multiple genomes, it is attributed to the genome with the greater number of associated peptides. However, this typically results in approximately 1,000 genomes remaining, with many containing only a single peptide. The number substantially larger than that is found in a typical human gut microbiome which is around 200[5]. So, we only choose the most abundant species for subsequent analysis. The selection of species for consideration is further explored in the optimization section of the study.

For the functional score, we constructed a fixed table from the MetaPep project [59]. While building the database Metapep, the peptide identification was performed by the software MetaLab MAG[67], which provides quantifications of protein abundance. Those proteins are well annotated. Subsequently, the relative abundance of each Clusters of Orthologous Groups (COG) accession was computed. Samples comprising fewer than 1000 COG accessions were considered to be of low quality and consequently were omitted from the analysis. A total of 1,031 high-quality samples were retained for further evaluation. The mean of non-zero relative abundance of the COG accessions was then determined across these 1,031 samples, establishing a metric referred to the functional score.

The FuncTax score was obtained by multiplying two scores. In the case of peptides with the same sequence, their FuncTax scores were combined to give higher priority to shared peptides; the highest detectability score among them was utilized to ensure the inclusion of all possible peptides.

4.3.3 Taxonomic and functional analysis

The taxonomic analysis is similar to the generation of genomic abundance score. The identified peptides are mapped to a database to establish peptide-genome associations. The database contains only the top 50 genomes. In our workflow, the peptide database was filtered out from the top 50 genomes. So, all the identified peptides were from the top genomes and thus can be used for the taxonomic analysis (the peptides added from MetaPep may not be used). In the DDA-based method, the peptide identified by the DDA library can be annotated to over 1,000 genomes even after using the greedy algorithm described above (method: generation of FuncTax score). However, we found that the top 50 genomes accounted for 79%-90% of peptides and 87%-92% of peptide intensity (Supplementary Figure 4.1). To simplify the comparison between the two methods, we discarded the small number of peptides that cannot be annotated to the top 50 genomes. Similarly, the razor strategy is used to process peptides shared by multiple genomes. Finally, the intensity of each peptide is aggregated to infer genome abundance.

For functional analysis, a protein abundance was firstly generated using the same strategy as taxonomic analysis. The proteins in the UHGP database have been extensively annotated thus the protein abundance can be further interpreted into functional abundance.

4.3.4 Deepdetect software configuration

Protein digestion was simulated using Trypsin with the following parameters: a maximum of two missed cleavages, and peptide lengths ranging from 7 to 50 amino acids. Default settings were applied for all other parameters.

4.3.5 DIA software configuration

DIA-NN (version 1.8.1) was used to process all the DIA data in this study. Maximum mass accuracy tolerances were set to 10 ppm for both MS1 and MS2 spectra. The --relaxed-prot-inf option was used for library-free searching. The --no-maxlfq option was used to disable the normalization for the quantification benchmark experiment. All other settings were left default. The precursor matrix containing the peptide information was used for taxonomic and functional analysis.

4.3.6 Metaproteomic datasets

The dataset used for optimizing workflow is sourced from a published study and shared by the authors[71]. The dataset for evaluating accuracy is from in-house samples. *Blautia hydrogenotrophica* (DSM 101114; Leibniz Institute DSMZ- German collection of microorganisms and cell cultures) was cultured in LB broth. The human stool was collected from a healthy adult volunteer at the University of Ottawa, Ottawa, ON, CAN. The protocol (# 20160585-01H) was approved by Ottawa Health Science Network Research Ethics. The protein extraction and digestion were performed as described previously[43]. Peptide concentrations were measured using Thermo Scientific Pierce Quantitative Colorimetric Peptide Assays according to the manufacturer's directions.

The in-house samples were then analysed using an UltiMate 3000 RSLCnano system (Thermo Fisher Scientific, USA) coupled to an Orbitrap Exploris 480 mass spectrometer (Thermo Fisher Scientific, USA). Peptides were loaded onto a tip column (75 µm inner diameter ×15 cm) packed with reverse phase beads (3 µm/120 Å ReproSil-Pur C18 resin, Dr. Maisch HPLC GmbH). A 60 min gradient of 5 to 35% (v/v) from buffer A (0.1% (v/v)

formic acid) to B (0.1% (v/v) formic acid with 80% (v/v) acetonitrile) at a flow rate of 300 µL/min was used. The mass spectrometer was in data-independent mode covering the mass range of 380–980 m/z with 10 m/z isolation windows.

4.4 Result

4.4.1 MetaDIA Workflow overview: Taxonomy- and function-guided construction of peptide database for metaproteomics

Here we propose a new workflow for DIA based metaproteomics called MetaDIA. MetaDIA is a multistep workflow that systematically reduces the search space for DIA searching. At its basis, it relies on a combination of taxonomic abundance, functional abundance as a proxy of protein levels, and peptide detectability ultimately enabling DIA searching without the need for DDA results. Briefly, in the first step, we created a new database of peptides, called MetaPepDetec, obtained by in silico digestion and detectability prediction of the Unified Human Gastrointestinal Protein (UHGP, 4744 genomes) database into peptides[8, 112]. Then each peptide is annotated with a FuncTax score (Figure 4.1). Both the FuncTax and the detectability scores are used to reduce the peptide database.

The FuncTax scores for each peptide in the MetaPepDetec are calculated using information from the MetaPep database [59]. MetaPep is a core peptide database compiling peptides previously identified in the published human gut metaproteomics studies. The information from MetaPep was used to create a static table of COG relative functional abundances and a sample-specific table of taxonomic relative abundances (method). We noted that despite significant differences in the species composition of gut

bacteria among different individuals, their functions are remarkably similar[10]. Therefore, functional abundance hierarchy information could act to estimate the likelihood of a protein being observed. We analyzed the search results used to construct the MetaPep database which contains 2,134 raw files and 415 individuals[59] (Method). The functional ranking among various samples exhibits a strong correlation (Supplementary Figure 4.2a and 4.2b). We observed a stable pattern in the functional hierarchy of human gut bacteria: abundant functions consistently remain high, while scarce functions persistently stay low across all samples (Supplementary Figure 4.2c). The sample-specific table of taxonomic relative abundances was generated by searching the DIA data against MetaPep[59]. The identified peptides and their quantitation were used to create the table (Method). The FuncTax score for each peptide is calculated by multiplying the taxonomic score for its taxonomic annotation and the functional score of its functional annotation. For peptides with identical sequences, their FuncTax scores were aggregated thereby leading to shared peptides having a higher ranking.

In the last step, sample-specific reduced peptide database is generated by filtering MetaPepDetec using the FuncTax score and the detectability score (Figure 4.1). The final search of the DIA data is done against the reduced peptide database. To validate the efficiency of our peptide ranking method, peptides were sorted by FuncTax score and partitioned into equal-sized subsets based on their percentile rank (e.g., top 0-5%, 5-10%, ..., 35-40%). Each subset was subjected to database searching with uniform parameters. We observed a decline in the number of peptides identified as the percentile ranking of the subsets decreased (Figure 4.2). The decreasing trend suggested our ranking method effectively prioritizes peptides with a higher probability of detection.

4.4.2 Optimized MetaDIA parameters reduces the database size

We explored whether the number of microbes in the reduced peptide database, the threshold for FuncTax score and the threshold for detectability score influenced the identification of peptides. We explored the impacts of the parameters using the DIA data from 10 different human gut microbiome samples previously reported[71].

In particular, we first tested the effect of the number of microbes (genomes) ranging from 50 to 150 and FuncTax score ranging from top 1% to 40% (Figure 4.3). We kept the detectability threshold at the top 40% in this experiment which is suggested by the author of Deepdetect[112]. Interestingly, no matter how many genomes we choose, the size of the reduced peptide database had the strongest effect on the number of identified peptides. The identification number plateaued once the reduced peptide database size reached around 1.6 million entries (Figure 4.3a and 4.3b, Supplementary Figure 4.3 and 4.4), corresponding to a FuncTax score threshold of 40% for 50 genomes, 20% for 100 genomes and 15% for 150 genomes respectively. We compared the three different reduced peptide databases, which led to consistent peptide identification results (Figure 4.2c and 4.2d, Supplementary Figure 4.5 and 4.6). In our previous studies, we observed that low-abundance species were underrepresented[91]. In this context, we chose to focus on the top 50 genomes to prioritize high-abundance genomes. It is important to note that this cut-off is a variable parameter that can be adjusted according to the specific objectives of different studies.

Subsequently, we explored whether the detectability threshold impacted the number of peptides identified. While the recommended threshold by the author of Deepdetect is 40%,

we explored thresholds ranging from 40% to 10%. We observed that a threshold of 25% was the point at which the number of identifications began to decrease significantly (Figure 4.4a). However, both the database size and the search time decreased substantially (Figure 4.4b). Comparing the identification results at thresholds of 25% and 40%, we found a substantial overlap (Figure 4.4c). Therefore, we selected a 25% threshold for detectability. Based on this analysis, we proceeded with peptides ranking in the top 40% by FuncTax score and the top 25% by detectability. Given that these two scores are entirely uncorrelated, applying both filters effectively reduced the database to one-tenth of its original size (25% times 40%, Supplementary Figure 4.7). After applying these optimized parameters, approximately 1 million peptide sequences remain in the reduced database.

4.4.3 MetaDIA maintains accuracy in DIA peptide identification.

We next explored whether the enrichment of high abundant and highly detectable peptides in our reduced database impacted the accuracy of peptide identification when applying the false discovery rate strategy. To evaluate this, we conducted a benchmark experiment using three samples: a human gut microbiome sample (A), a *Blautia hydrogenotrophica* sample (C), and a 50:50 mixed sample of the two (B) (Figure 4.5a). *Blautia hydrogenotrophica* was selected due to its absence in the microbiome sample used here and its minimal peptide overlap with the microbiome sample. Each sample was subjected to triplicate DIA measurements. Sample A and B were analyzed using the reduced peptide database generated by our workflow with optimized parameters of top 40% FuncTax score and top 25% detectability score, whereas sample B and C were searched against species-specific protein databases derived from NCBI (Genome

assembly ASM15797v1). In the first search against the peptide database, 32,624 unique peptides were identified. Of these, 1,952 peptides also present in the *Blautia hydrogenotrophica* database were excluded. Further, peptides unique to each sample were removed, leaving 27,830 peptides identified in both sample A and B. Ideally, the peptide abundance ratio between samples A and B should approximate 2. In the second search, 14,821 unique peptides were identified. Among these, 10,995 peptides were unique to the *Blautia hydrogenotrophica* database and were found in both samples B and C. The expected ratio between samples B and C should be around 0.5. We found whether using a protein database or a peptide database, the ratios of peptides identified in both searches closely aligned with the expected values (Figure 4.5b). This suggests that the employment of our reduced peptide database does not significantly affect the accuracy of peptide identification, thereby supporting its use in peptide identification workflows with a controlled FDR.

4.4.4 MetaDIA yields consistent peptide and protein identification results with DDA based strategies.

We next evaluated whether MetaDIA performed similarly to a conventional DDA-based library for DIA data analysis. The DIA data and corresponding DDA-based library were obtained from a published study[71]. We found that the MetaDIA provided identification numbers comparable to those obtained through the DDA-based library (Figure 4.6a). Notably, in certain instances, such as with samples 8 and 9, the MetaDIA surpassed DDA library in the number of identifications. The initial step in our workflow involves searching the raw data against MetaPep, which leverages the results from an open search strategy, thereby encompassing modified peptides not included in the original database.

Subsequent integration of peptides identified by MetaPep into a refined peptide database resulted in a marked increase in identification rates (Figure 4.6a).

Over 50% of peptides identified from the DDA-based library were also identified by MetaDIA (Figure 4.6b and Supplementary Figure 4.8). The divergence in unique identifications between the two methods may be attributed to inherent differences between DDA acquisition and DIA acquisition. Upon examining the quantification results of those peptides found by both methods, we observed a significant consistency in the outcomes, with a Pearson coefficient above 0.9 (Figure 4.6d and Supplementary Figure 4.9). It is worth noting that the fragment ions used for quantification in the DDA-based library correspond to actual DDA acquisitions. In contrast, MetaDIA uses theoretical spectra that are predicted from peptide sequences. The high degree of agreement between the quantification results underscores the reliability of the MS-Simulator algorithm which is employed by DIA-NN for spectra prediction [103].

At the protein level, our findings revealed greater consistency in identification compared to the peptide level (Figure 4.6c). Around 70% of proteins found by the DDA-based library can be found by MetaDIA. The overlap on protein level reinforces the reliability of the identifications and indicates that a significant subset of proteins is consistently identified by both methods despite differences at the peptide level (Supplementary Figure 4.10). Proteins like MGYG000002545_00035 had greater sequence coverage and higher detection intensity with the DDA library, while others like MGYG000002272_00452 showed higher coverage and intensity with MetaDIA. Given that the quantification of a protein is derived from different subsets of peptides in these two methods, we observed reduced consistency of quantification in the protein level between the methods, as

reflected by Pearson correlation coefficients of approximately 0.7 (Figure 4.6e and Supplementary Figure 4.11). However, it is important to note that in most proteomic studies, the primary interest lies in the differential abundance of the same protein across various samples. Therefore, it is crucial that we use the same fragment ions to quantify a protein. In this regard, the inconsistencies in protein quantification between the two methods do not undermine the utility of either approach. The substantial overlap in peptide and protein identification by both methods suggests a robust cross-validation of both methods. Then we annotated the proteins using COG accessions and calculated their relative abundances. Our analysis revealed that approximately 90% of the COG accessions identified by the DDA-based method were also covered by our MetaDIA (Figure 4.6c). Furthermore, the Pearson correlation coefficient for the relative abundance of COG accessions exceeded 0.9, with a stronger correlation for those COG accessions that were highly abundant (Figure 4.6f and Supplementary Figure 4.12).

4.4.5 MetaDIA provides taxonomic profiles highly similar to those obtained from searching DDA-libraries.

We verified whether both methods had a high degree of similarity in the taxonomic composition. We did comparative analysis of microbiome composition across different taxonomic levels using the result from both methods. Our findings indicate that there is a significant linear correlation between the compositions identified by both methods, with the degree of correlation strengthening at higher taxonomic levels (Figure 4.7a and 4.7b, Supplementary Figure 4.13). The two methods showed remarkably consistent taxonomic composition at the genus level with a Pearson coefficient above 0.98 across all the samples tested. Even Sample 9, which displayed the lowest correlation, demonstrated a

substantial degree of consistency between the two methods. To underscore the consistency, we have provided a detailed visualization of the taxonomic composition for Sample 9 (Figure 4.7c and d, Supplementary Figure 4.14)

The species compositions observed by MetaDIA in these ten samples differed significantly as expected, indicating that our database and taxonomic analysis have the capability to identify a diverse range of microbiota (Supplementary Figure 4.14). The most abundant species identified in the ten samples have been previously reported as high-abundance species in the human gut microbiome[113-117]. Except for *Phocaeicola dorei* which were identified as the top species in sample 2, 5 and 10, the other top species were all unique to each sample.

4.4.6 MetaDIA is universally applicable to different types of DIA, including DIA-PASEF

To further validate the versatility and applicability of our proposed metaproteomic workflow, we extended our analysis to a diverse set of 79 DIA datasets obtained from a published study[71]. This dataset encompasses samples from 62 individuals, featuring replicate injections, quality control (QC) samples, and pooled samples. We applied MetaDIA to this extensive dataset and compared the results with the conventional DDA-based approach. Remarkably, the number of peptides identified by both methods demonstrated a close equivalence, reinforcing the robustness and universal applicability of our metaproteomic workflow (Figure 4.8). Validating our method across diverse samples enhances confidence in its effectiveness and consistency, demonstrating its potential for widespread adoption in metaproteomics research.

4.5 Discussion

We propose a novel workflow for DIA data analysis from human gut microbiome called MetaDIA. The approach aims at prioritizing peptides with a higher likelihood of detection based on their detectability, taxonomic and functional scores.

MetaDIA is entirely devoid of DDA, thereby circumventing the drawbacks of DDA-based methods. Not only does this approach save time and resources, but it also enables the creation of a tailored database for each sample. In contrast, DDA-based methods typically rely on a single pooled sample to generate a library. For instance, Gomez *et al*.[70] used a pooled sample to represent 12 individual mice, while Sun *et al*.[71] did so for a cohort of 62 individuals. However, such a pooled sample may not effectively represent every sample. In our study, the ten samples showed highly diverse taxonomic composition (Supplementary Figure 4.13). To increase the sampling depth for the pooled sample, Sun *et al.*[71] had to fractionate the pooled sample into 30 portions and Gomez *et al*.[70] repeatedly injected the pooled sample 10 times. Moreover, utilizing a static library to search various samples may potentially compromise the accuracy of peptide identification, as it includes peptides from the pooled samples that are absent in the specific sample under investigation.

In MetaDIA, we pre-defined the range of genomes for each microbiome sample (50 genomes in this study). This approach not only enabled us to narrow the search space but also to mitigate the issues associated with protein inference that arise from common peptides. When assigning peptides to proteins, we confined our consideration to the

genomes within the predefined range rather than the entire dataset. This strategy significantly reduced the incidence of common peptides.

Although MetaDIA is currently focused on the human gut microbiome, we foresee that it can be extended to other types of microbiomes, such as those in animal intestines, environmental microbiomes when using an appropriate bait database. A database similar to MetaPep could be constructed for other microbiomes.

4.6 Conclusion

In conclusion, we introduced a new strategy to prioritize peptides with a high probability of detection. This strategy simulates protein digestion procedures in silico and uses taxonomic and functional information to infer the peptide abundance. MetaDIA is a fully DDA-free workflow and provides a user interface to change the different parameters. We compared the performance of MetaDIA with the DDA-based library and observed a high degree of consistency. We further validated our method across a DIA-PASEF dataset with 79 samples, thereby confirming its wide applicability. We believe that our approach will help the application of DIA in metaproteomics.

Availability of data and materials

The whole pipeline is available for use at https://github.com/northomics/MetaDIA

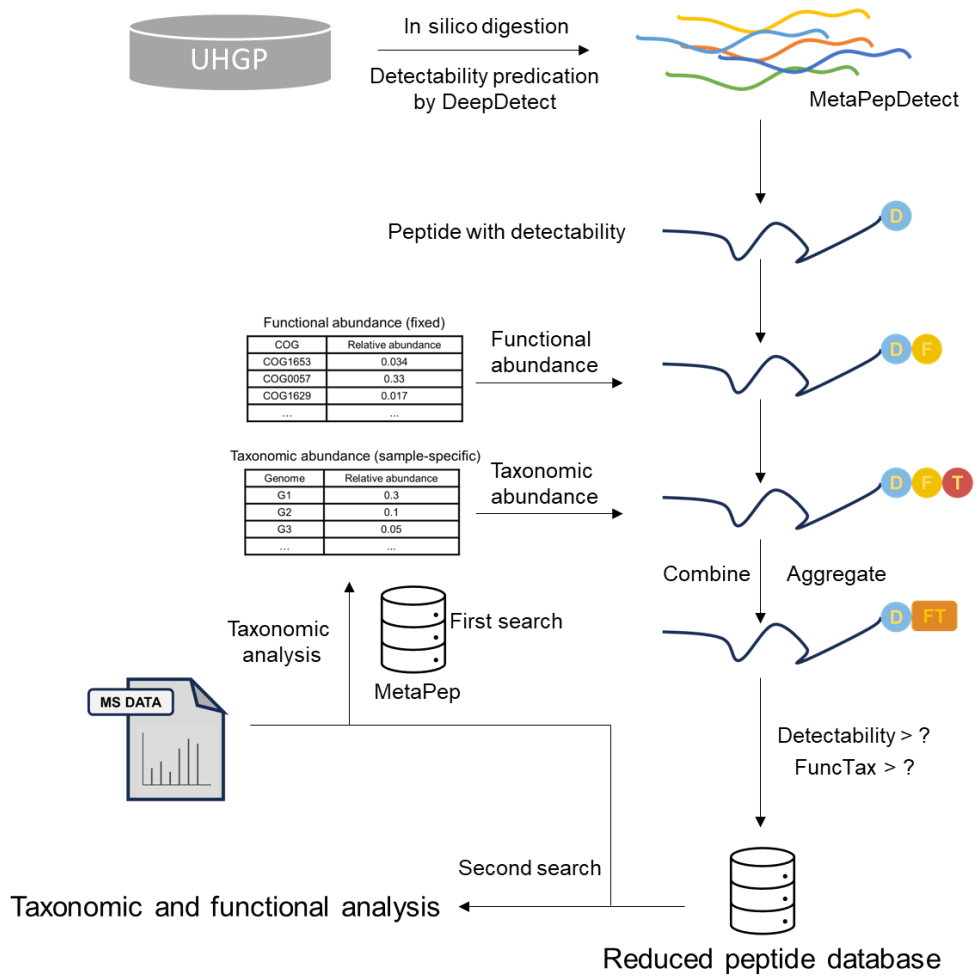


Figure 4.1 The flowchart for the MetaDIA.

All proteins in Unified Human Gastrointestinal Protein (UHGP) database were firstly in silico digested into peptides. The detectability of each peptide was predicated by DeepDetect algorithm. Following this prediction, each peptide was assigned a functional score and a taxonomic score, derived from a predetermined functional relative abundance table and a sample-specific taxonomic relative abundance table, respectively (method). The FuncTax score was calculated by multiplying the two scores. For peptides with identical sequence, their FuncTax scores were aggregated to prioritize shared peptides; the maximum of their detectability scores was used to ensure the inclusion of all potential

peptides. The detectability and FuncTac scores are both used for filtering peptides. The reduced peptide database was used for a second search.

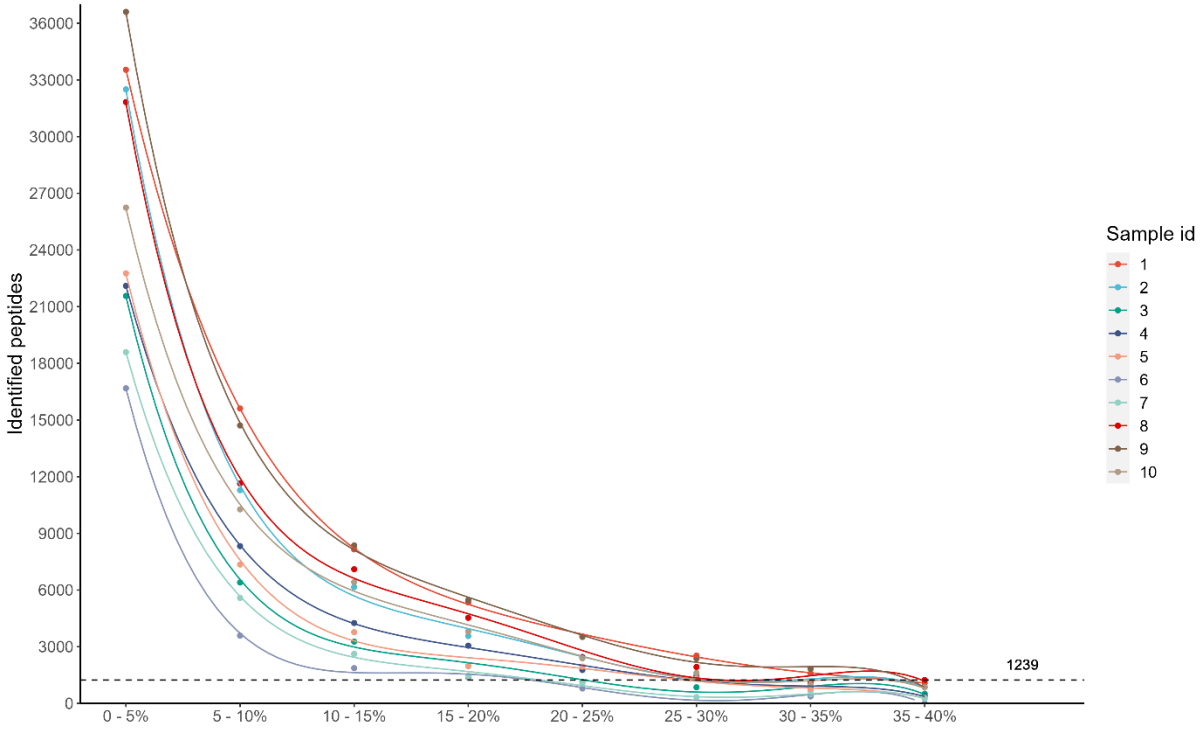


Figure 4.2 The number of peptides identified from each subset.

Ten samples were tested in the experiment. For constructing the peptide database, the top 100 genomes were considered; the detectability threshold was set at 40%. Each subset contains around 400,000 peptides. Peptide identification was performed by DIA-NN under same conditions. The maximum identification from the last subset was heighted in the figure.

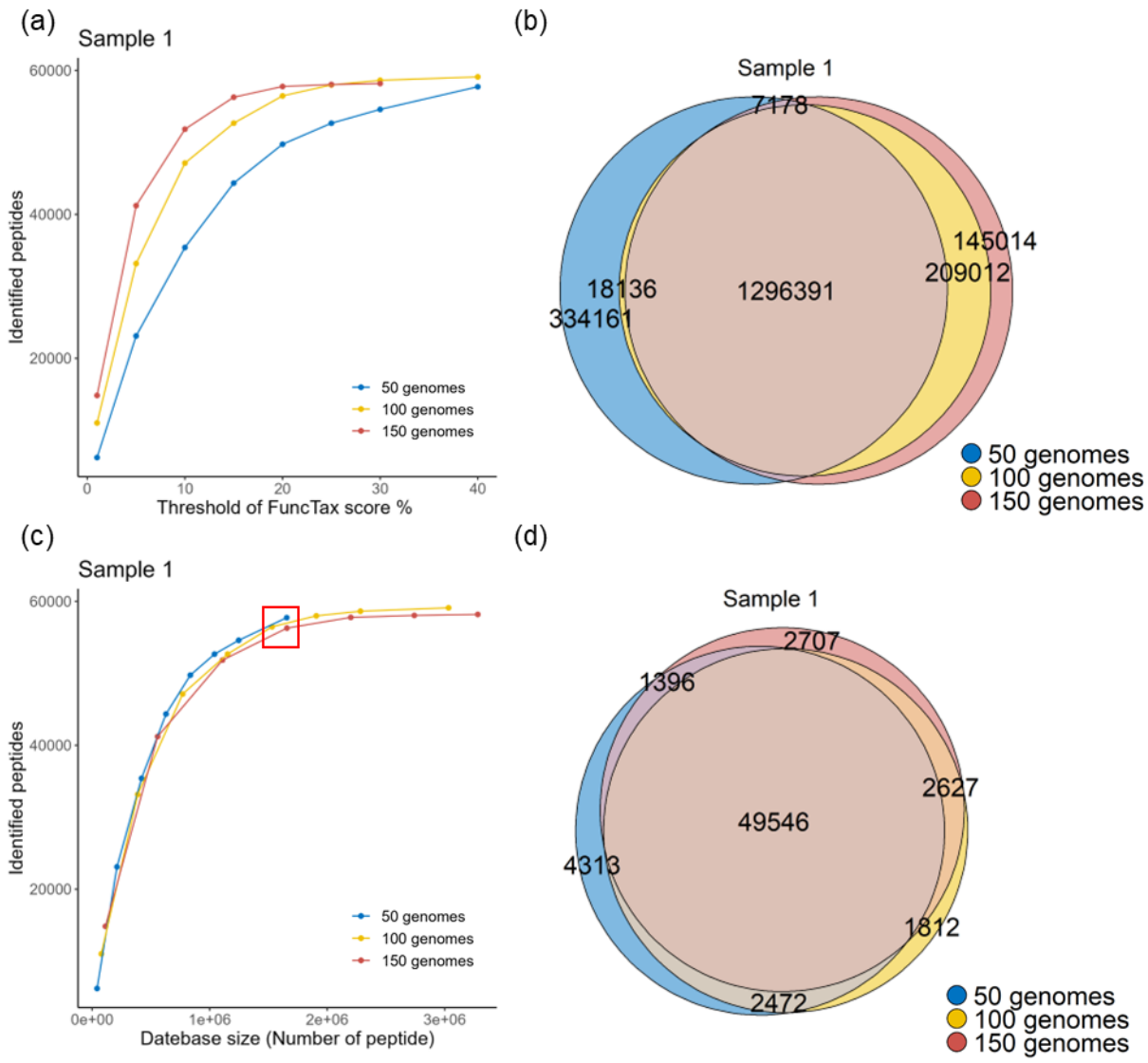


Figure 4.3 Optimization for genome number and FuncTax score.

(Sample 1 was shown. For the other samples, please see Supplementary figures). Peptides from n (50, 100, 150) genomes were ranked by the FuncTax score and top x% (1-40 for 50 and 100 genomes; 1–35 for 150 genomes) peptides was used as database. (a) Number of identified peptides against database percent. (b) Number of identified peptides against database size. The inflection point has been highlighted with a red box. (c) The overlap of the reduced peptide database and (d) identified peptide when taking

top 40% peptides for 50 genomes, top 20% for 100 genomes and top 15% for 150 genomes as database. Peptide identification was performed by DIA-NN under same conditions.

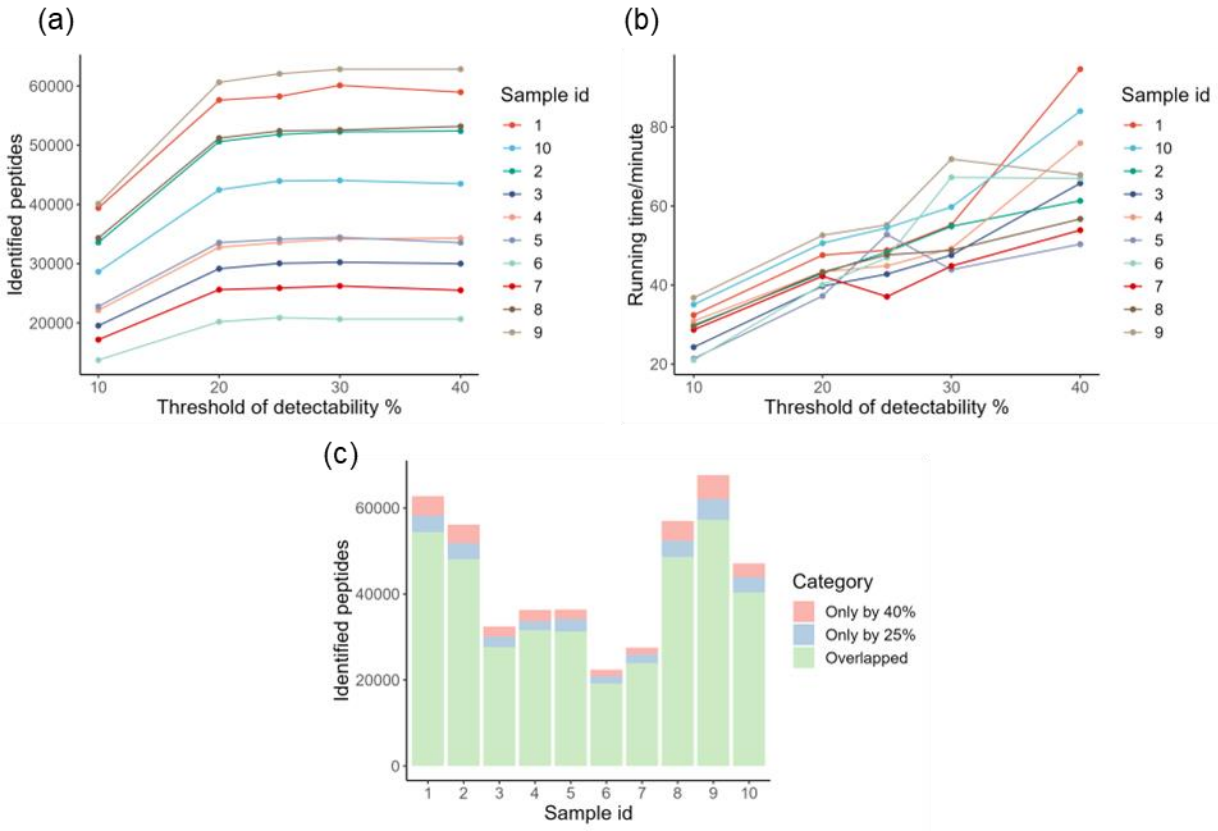


Figure 4.4 Optimization for detectability threshold.

(a) The number of peptides identified and (b) the searching time under detectability threshold from 10% to 40%. (c) The overlap of peptides identified by top 25% and top 40% of the database. Peptide identification was performed by DIA-NN under same conditions.

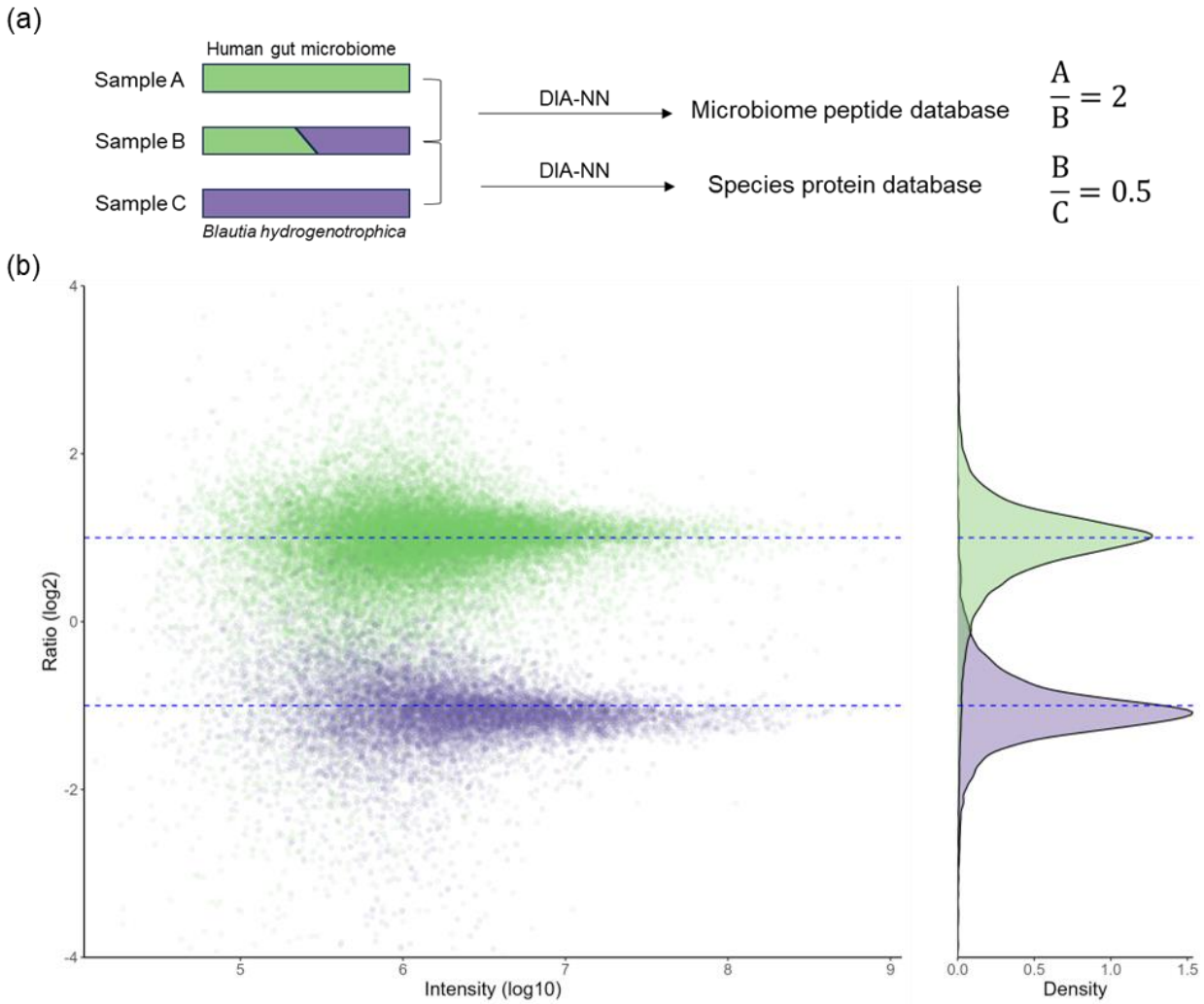


Figure 4.5 Benchmark experiment for peptide identification.

(a) The experimental design. Each sample was subjected to triple-run measurements (b) Log-transformed ratios are plotted as a function of peptide intensity for n = 27,830 (green) microbiome peptides and n = 10,995 (purple) *Blautia hydrogenotrophica* peptides. The point density for ratio was plotted at right. Dashed lines indicate the expected ratio. Peptide identification was performed by DIA-NN under same conditions. The intensities derived from various charge states of the same peptide were aggregated.

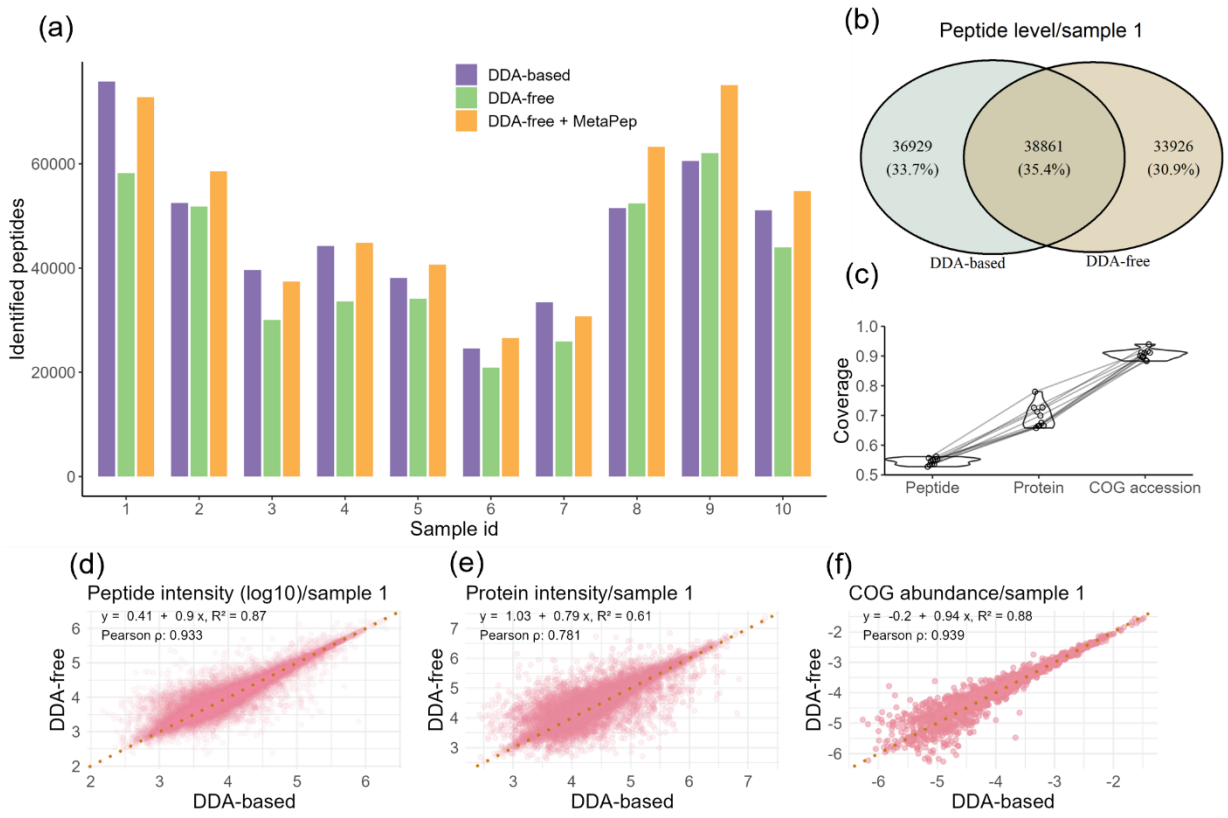


Figure 4.6 Comparison between the DDA-based method and DDA-free method.

(a) The peptide identified by each method. (b) The overlap of peptide identified in sample 1 by each method. (c) Coverage of peptides, proteins and cog accessions identified by DDA-based method with those found using DDA-free method. The intensity correlation of the overlapped peptides (d), proteins (e) and COG accessions in sample 1. The dashed line indicates y = x. For DDA-based method, the peptides identified as derived from human proteins are removed.

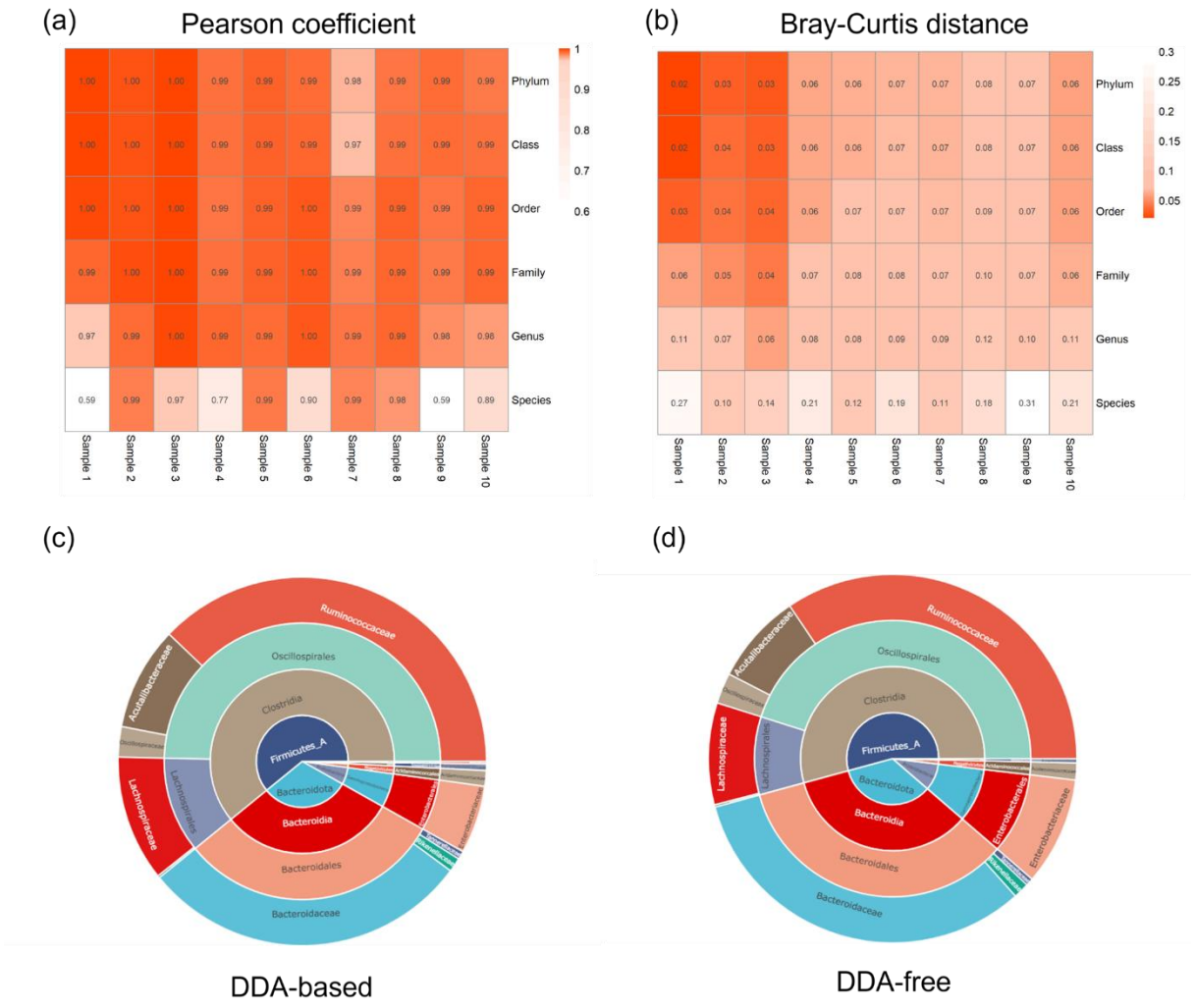


Figure 4.7 Comparison of the taxonomic composition between the DDA-based method and DDA-free method.

Pearson correlation (a) and Bray-Curtis distance (b) analysis between DDA-based method and DDA-free method on different taxonomic levels from Phylum to Species. The relative taxonomic abundance was used for the analysis. In the correlation analysis, taxonomic categories that were unique to one method were imputed with a value of zero. The taxonomic composition (Phylum to Family) of sample 9 derived from DDA-based method (c) and DDA-free method (d).

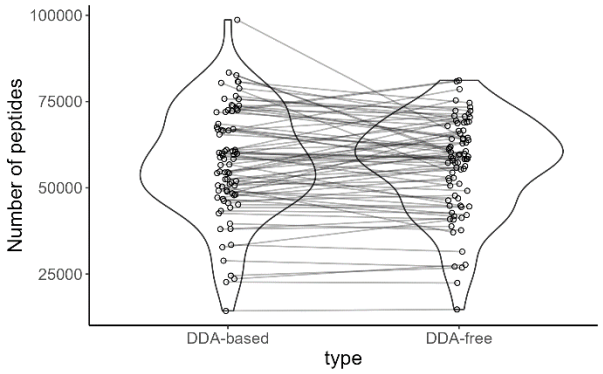
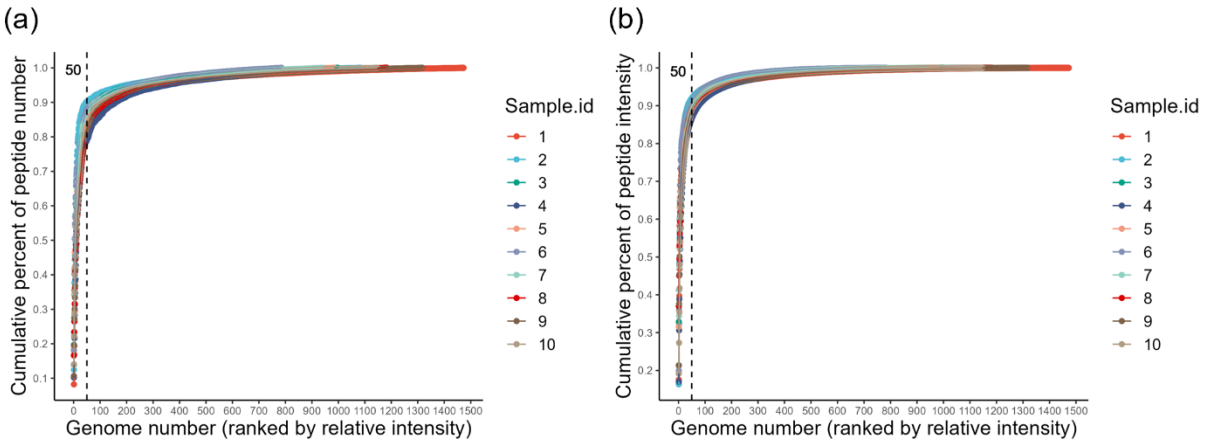


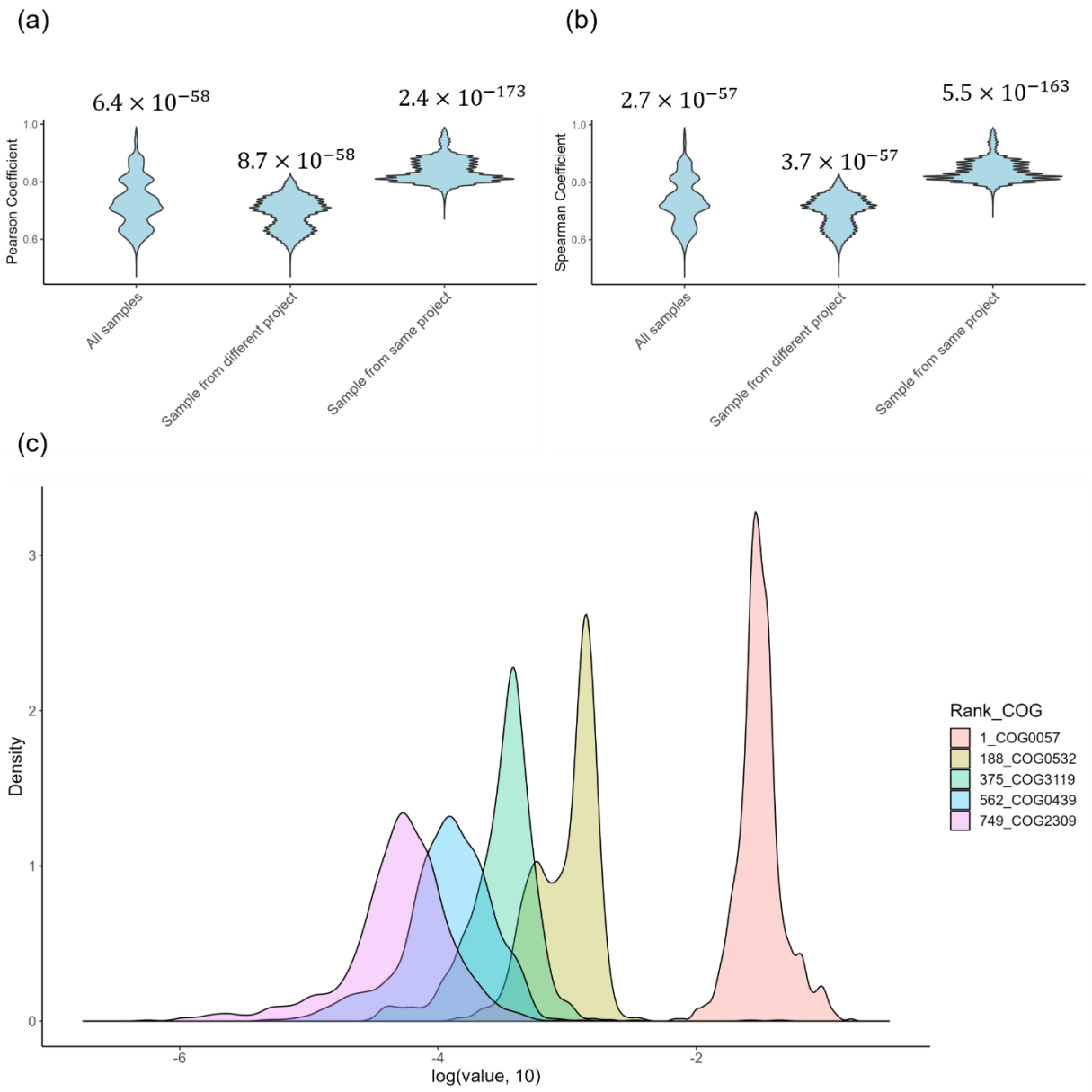
Figure 4.8 Number of peptides identified by both methods.

Number of peptides identified by both methods with mean value from 79 diaPASEF samples. For DDA-based method, the peptides identified as derived from human proteins are removed.



Supplementary Figure 4.1 Cumulative contribution of identified peptides and intensity.

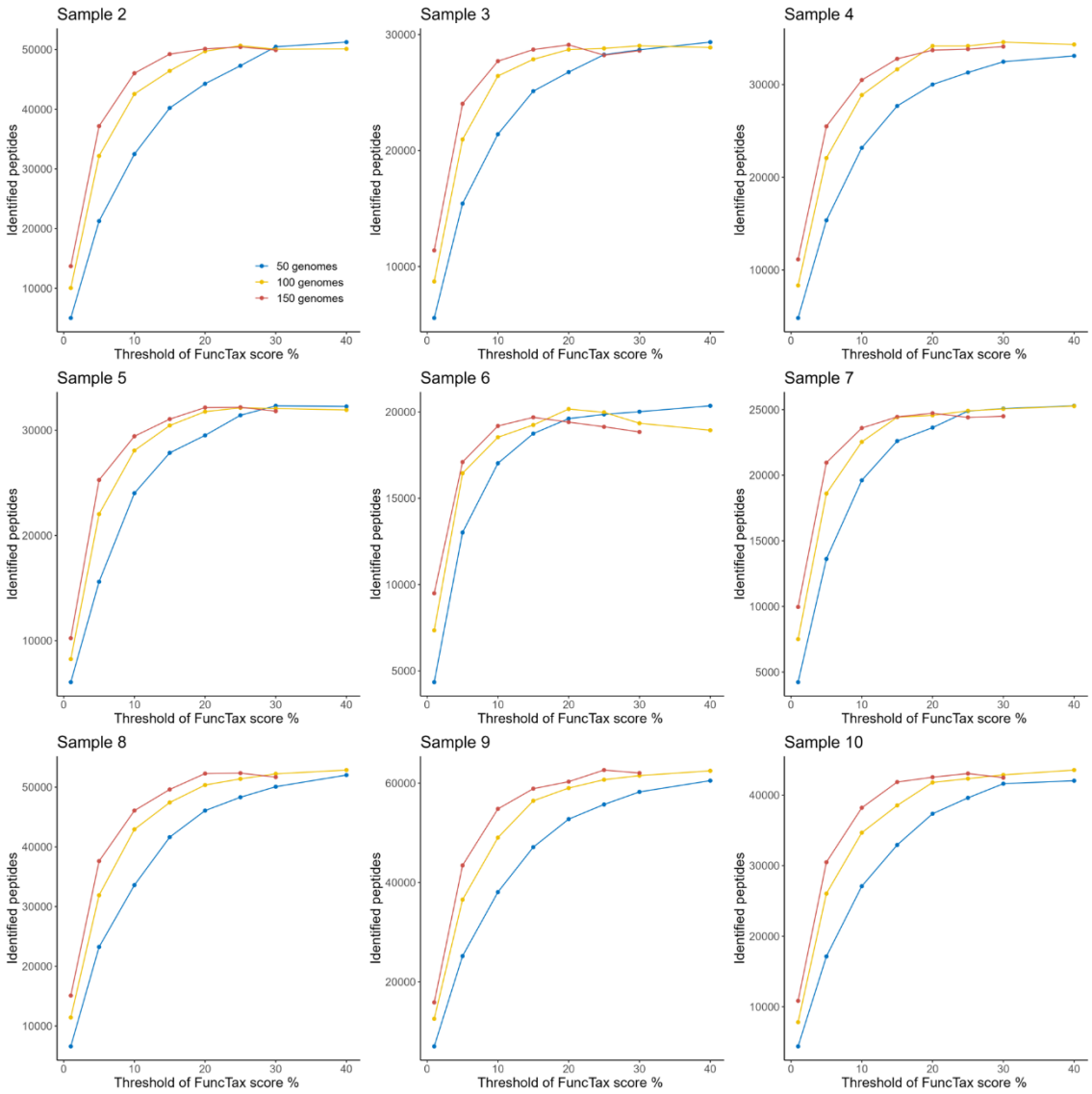
Cumulative contribution of identified peptides (a) and intensity (b). Figure were plotted against the number of genomes. Genomes are ordered by decreasing peptide count.



Supplementary Figure 4.2 The functional correlation across samples in MetaPep.

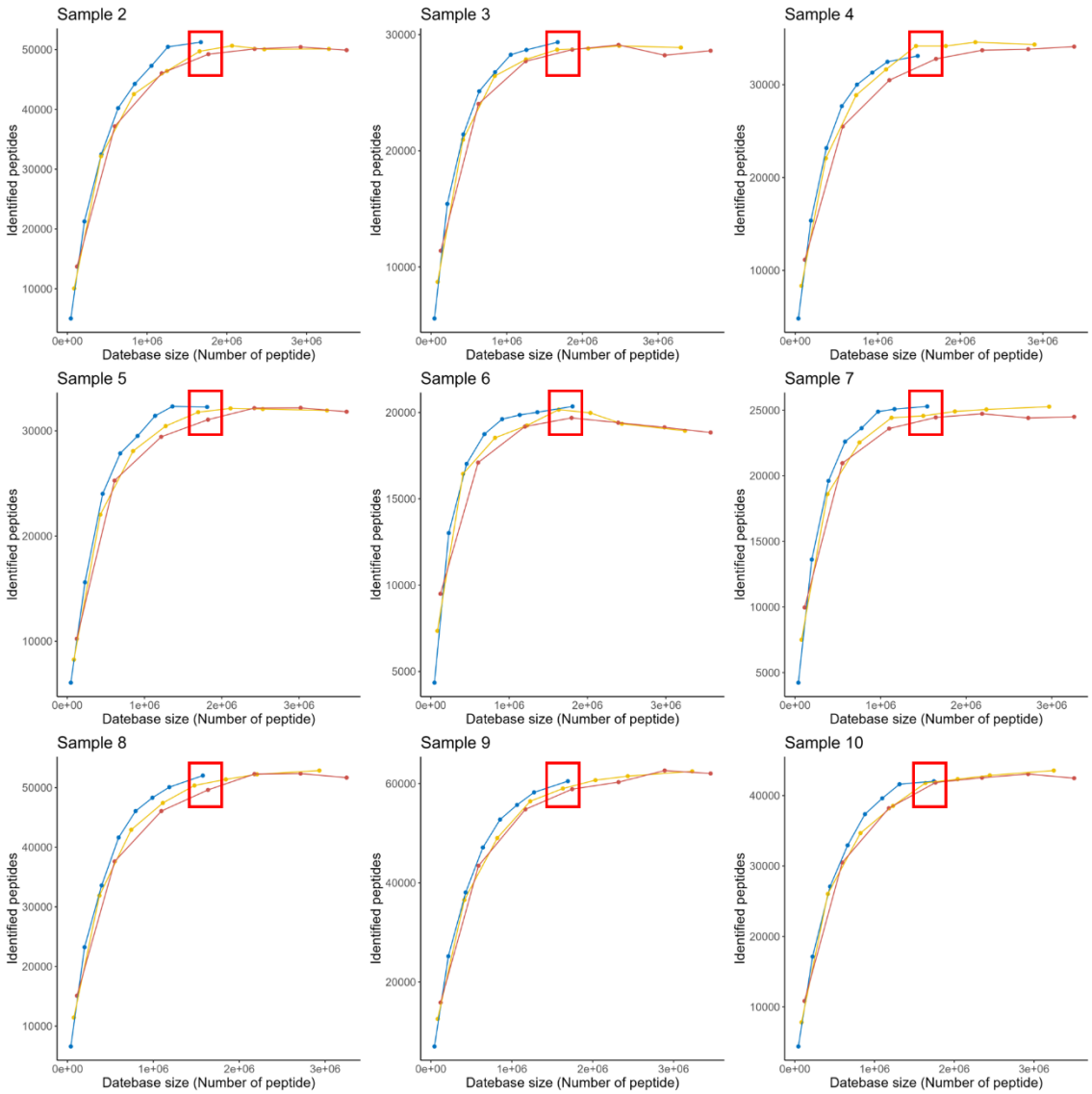
The Pearson coefficient (a) and Spearman coefficient (b) on COG accessions. Sample pairs from the same project are likely from the same individuals and plotted in different groups. The average p-values are plotted above each groups. (c) The distribution of relative abundance of COG accession across samples in MetaPep. The legends show

the rank and name of COG accessions. COG accessions present in more than 95% of the samples were retained. A total of 750 COG accessions remained and were subsequently sorted based on their functional abundance scores. Five COG accessions were selected for density plot at evenly spaced intervals from this ordered list.



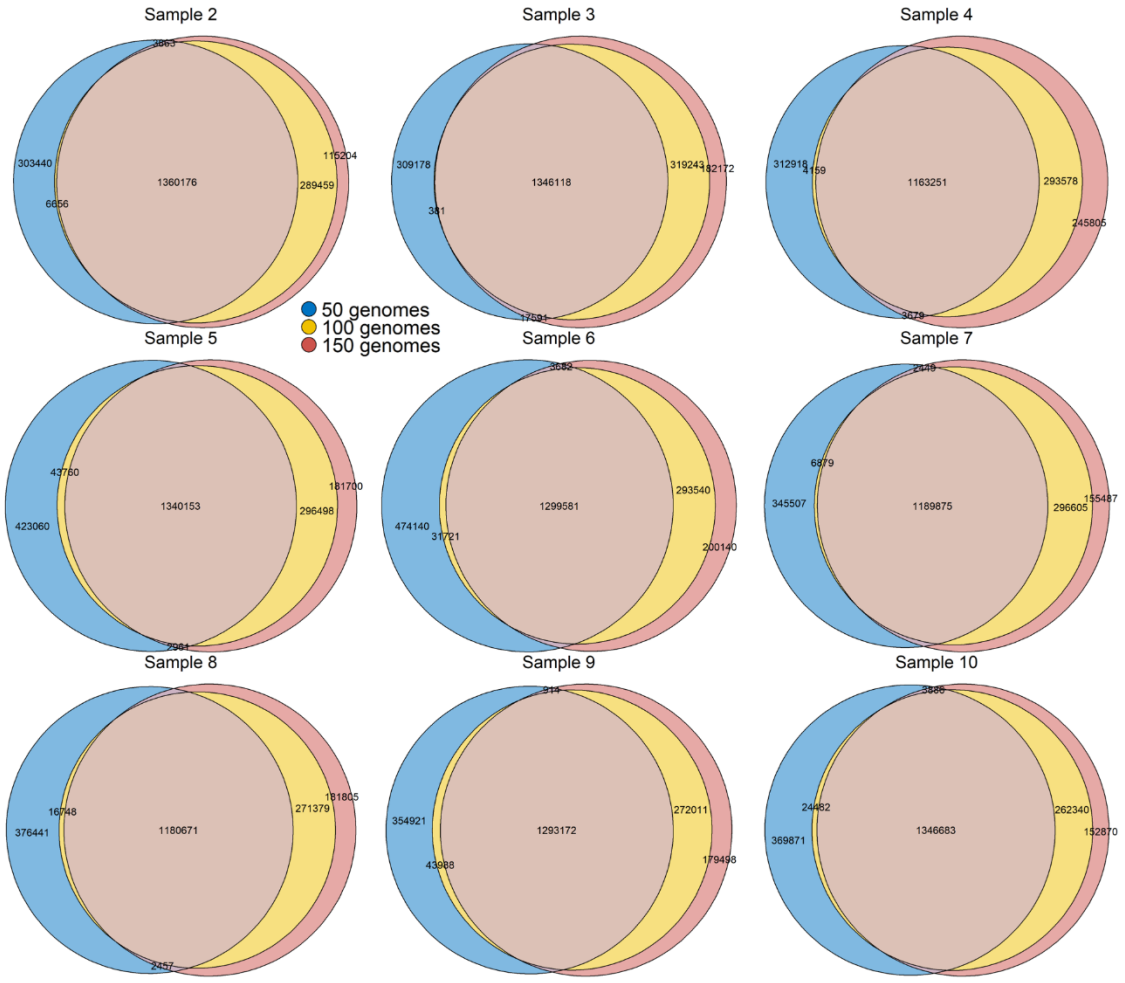
Supplementary Figure 4.3 Number of identified peptides against database percent.

Optimization for genome number and FuncTax score (Sample 2-10). Peptides from n (50, 100, 150) genomes were ranked by the FuncTax score and top x% (1-40 for 50 and 100 genomes; 1–35 for 150 genomes) peptides was used as database. Peptide identification was performed by DIA-NN under same conditions.



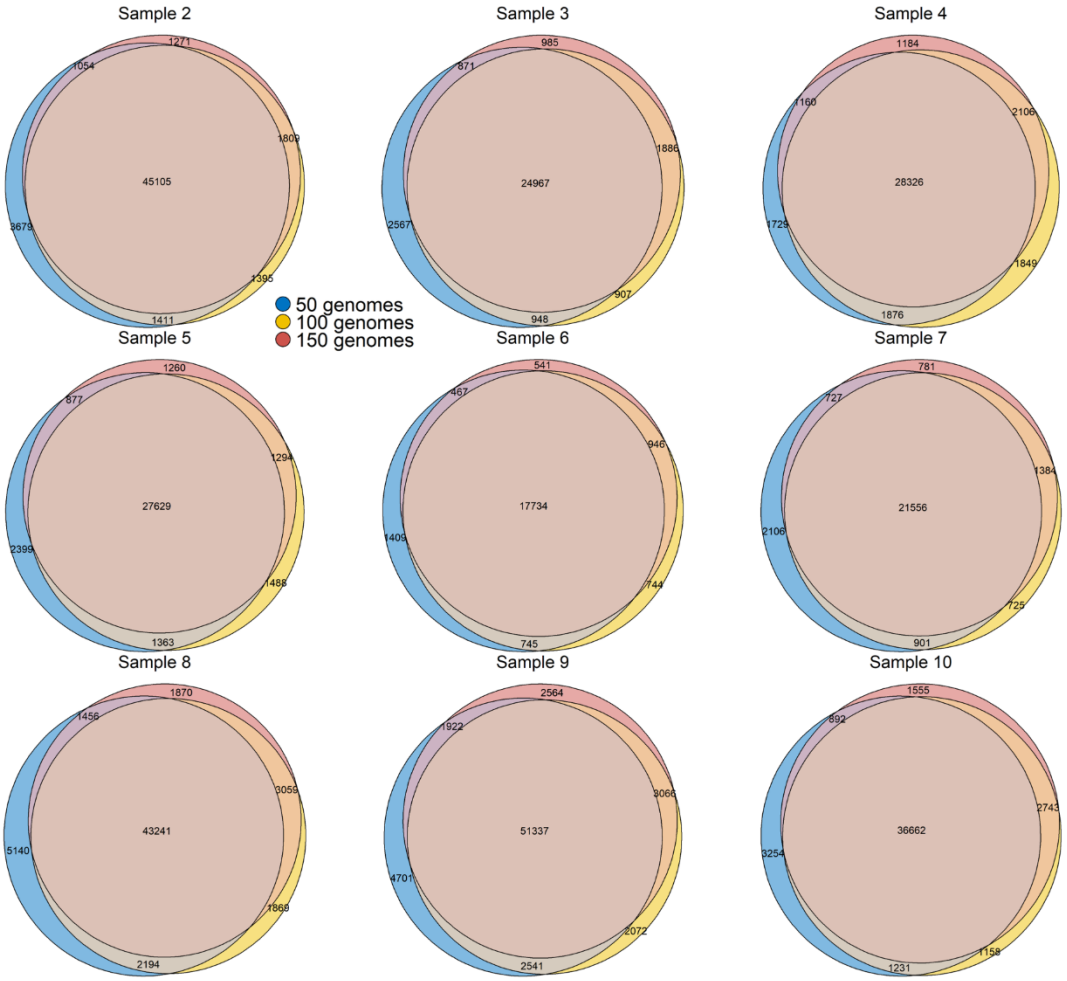
Supplementary Figure 4.4 Number of identified peptides against database size.

Optimization for genome number and FuncTax score (Sample 2-10). The inflection point has been highlighted with a red box. Peptides from n (50, 100, 150) genomes were ranked by the FuncTax score and top x% (1-40 for 50 and 100 genomes; 1–35 for 150 genomes) peptides was used as database. Peptide identification was performed by DIA-NN under same conditions.



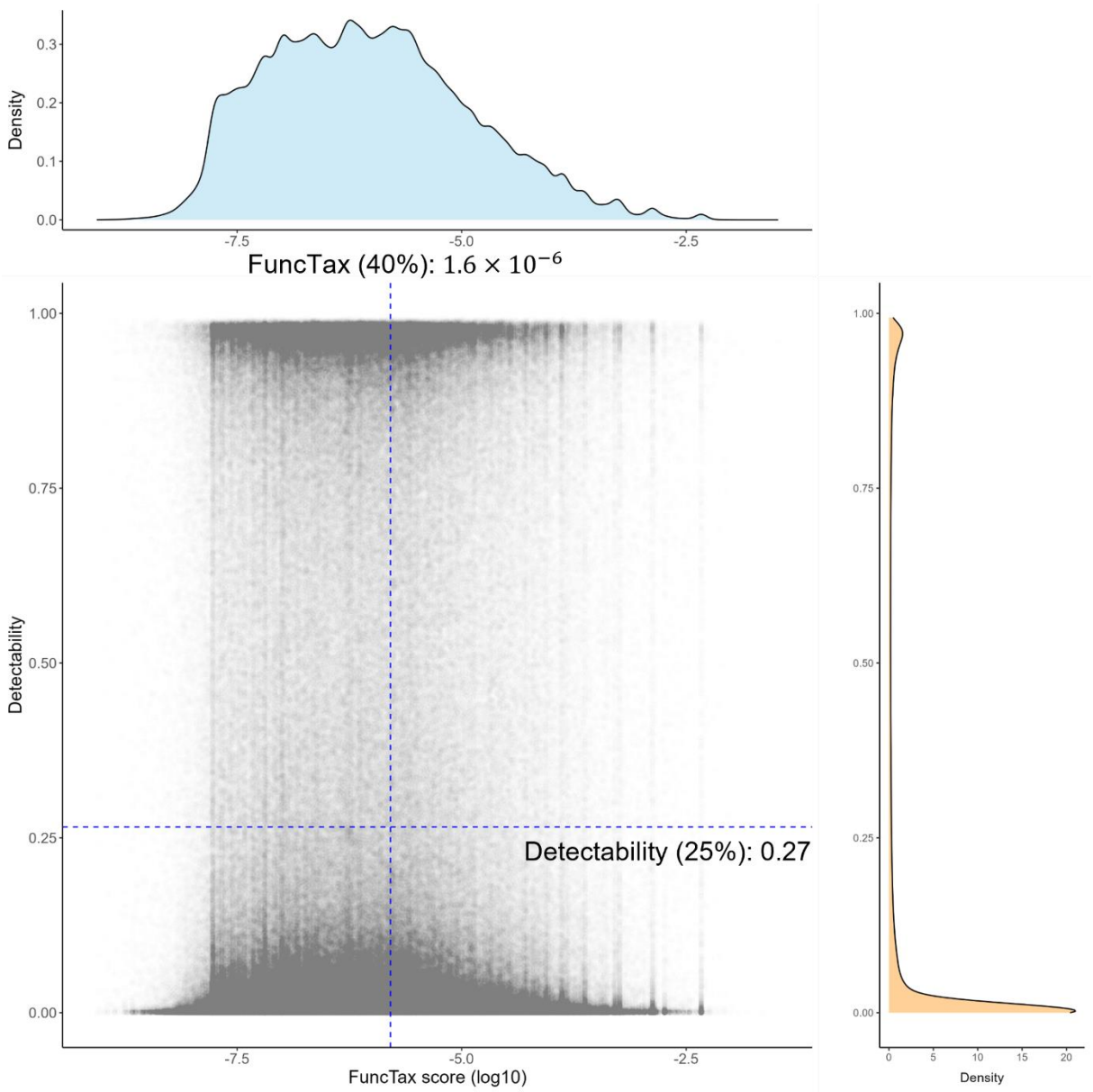
Supplementary Figure 4.5 The overlap of reduced peptide database

Optimization for genome number and FuncTax score (Sample 2-10). The overlap of reduced peptide database when taking top 40% peptides for 50 genomes, top 20% for 100 genomes and top 15% for 150 genomes as database.



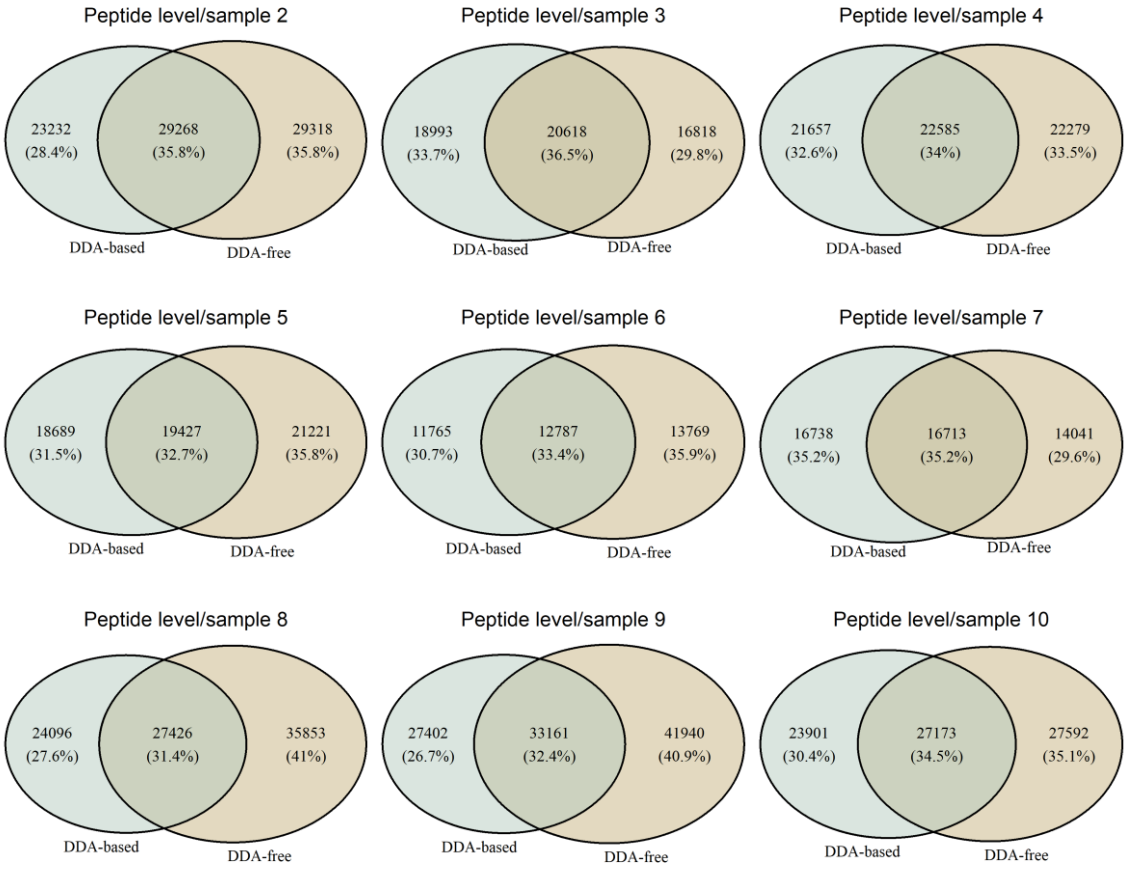
Supplementary Figure 4.6 The overlap of identified peptide.

Optimization for genome number and FuncTax score (Sample 2-10). The overlap of identified peptide when taking top 40% peptides for 50 genomes, top 20% for 100 genomes and top 15% for 150 genomes as database. Peptide identification was performed by DIA-NN under same conditions.

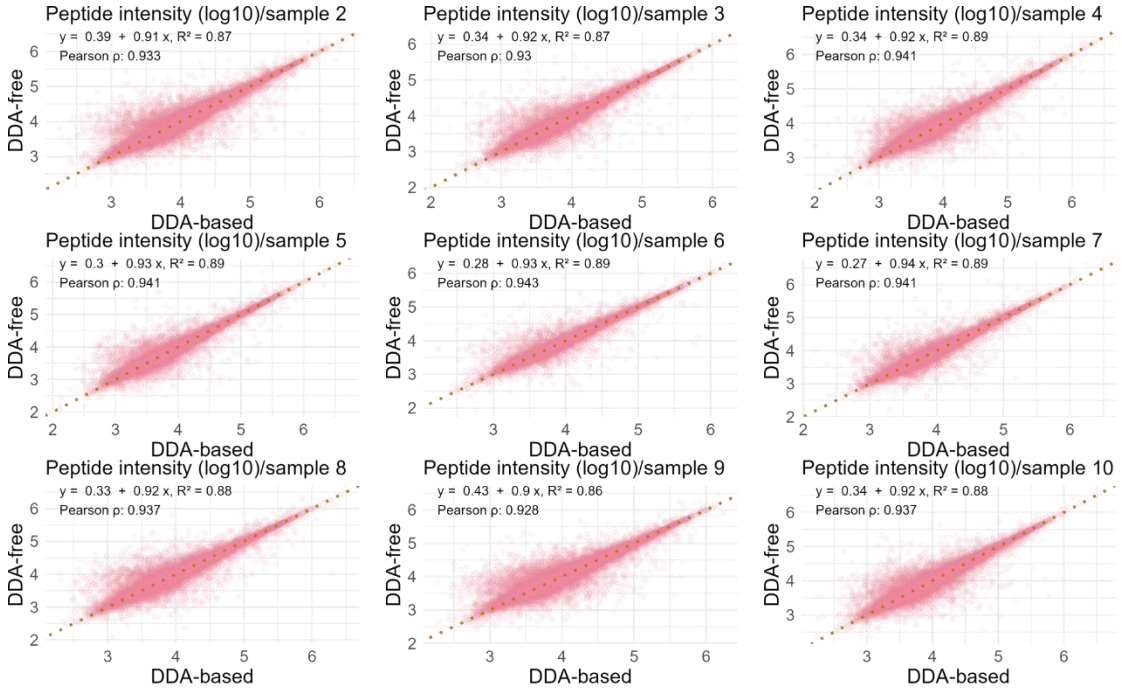


Supplementary Figure 4.7 The distribution of detectability score and FuncTax score.

(Sample 1, top 50 genomes). The dotted blue line shows the cutoff for detectability (top 25%) and FuncTax (top 40%).



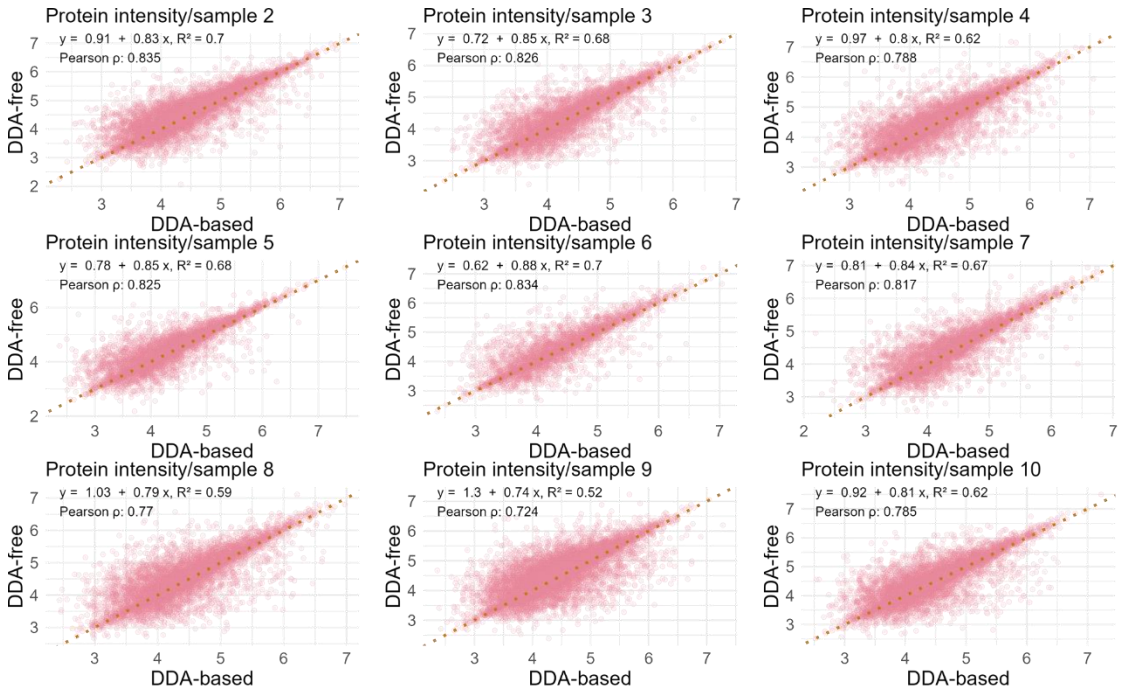
Supplementary Figure 4.8 The overlap of peptide identified by each DDA-based method and DDA-free method (Sample 2-10).



Supplementary Figure 4.9 The intensity correlation of the overlapped peptides found by both DDA-based and DDA-free method.

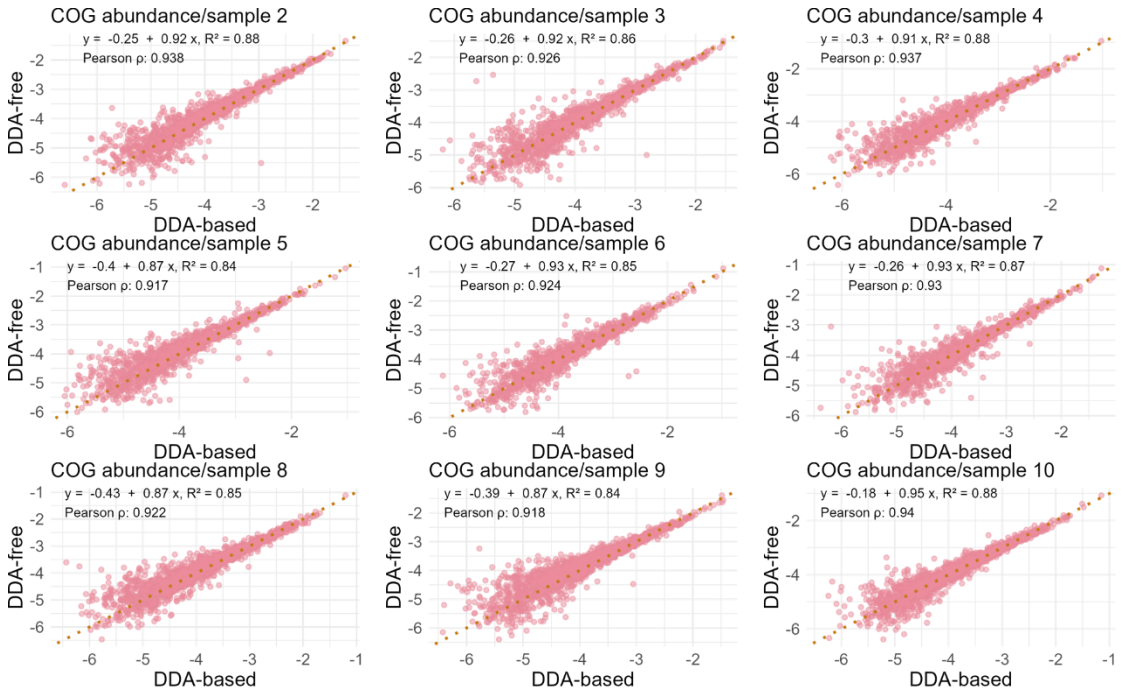
(Sample 2-10). The dashed line indicates y = x. For DDA-based method, the peptides identified as derived from human proteins are removed.

Supplementary Figure 4.10 The sequence coverage of representative proteins that exhibit the largest relative difference in intensity DDA-based and DDA-free method.



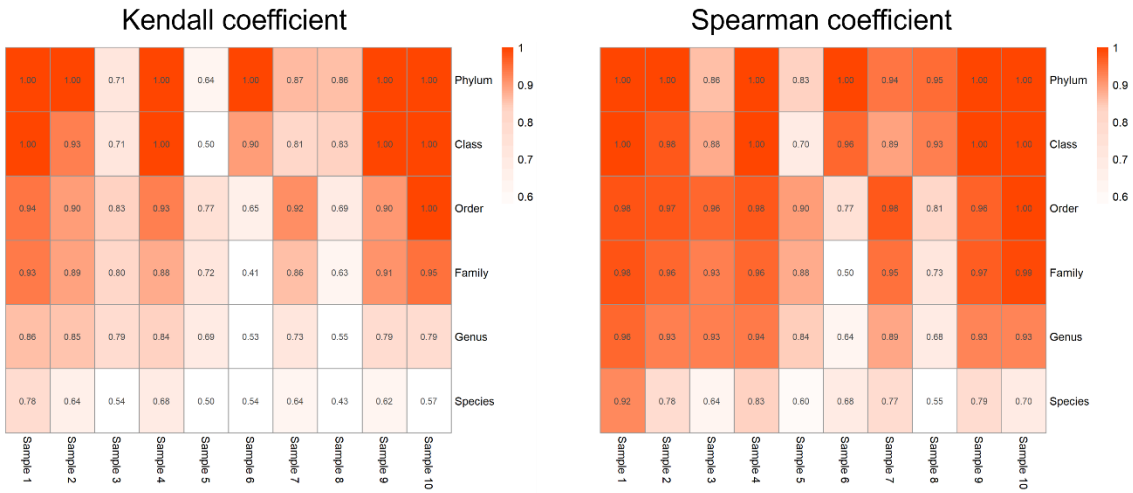
Supplementary Figure 4.11 The intensity correlation of the overlapped proteins found by both DDA-based and DDA-free method.

(Sample 2-10). The dashed line indicates y = x.



Supplementary Figure 4.12 The intensity correlation of the overlapped COG accessions found by both DDA-based and DDA-free method.

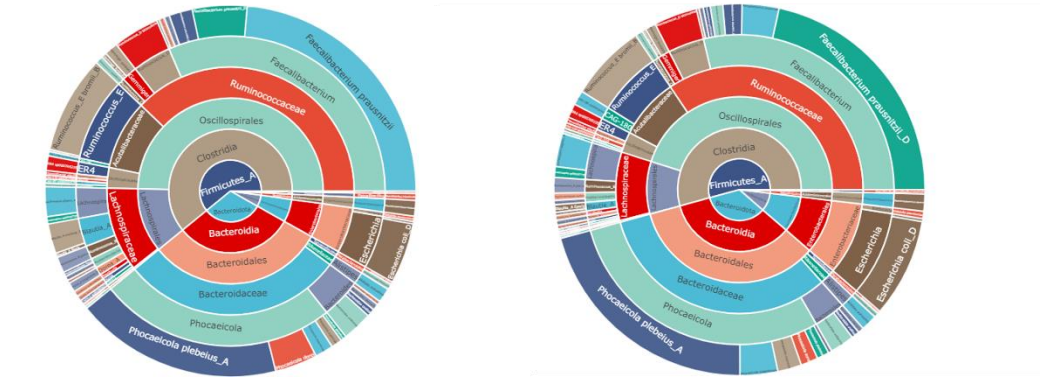
(Sample 2-10). The dashed line indicates y = x.



Supplementary Figure 4.13 Correlation analysis between DDA-based method and DDA-free method on different taxonomic levels from Species to Phylum.

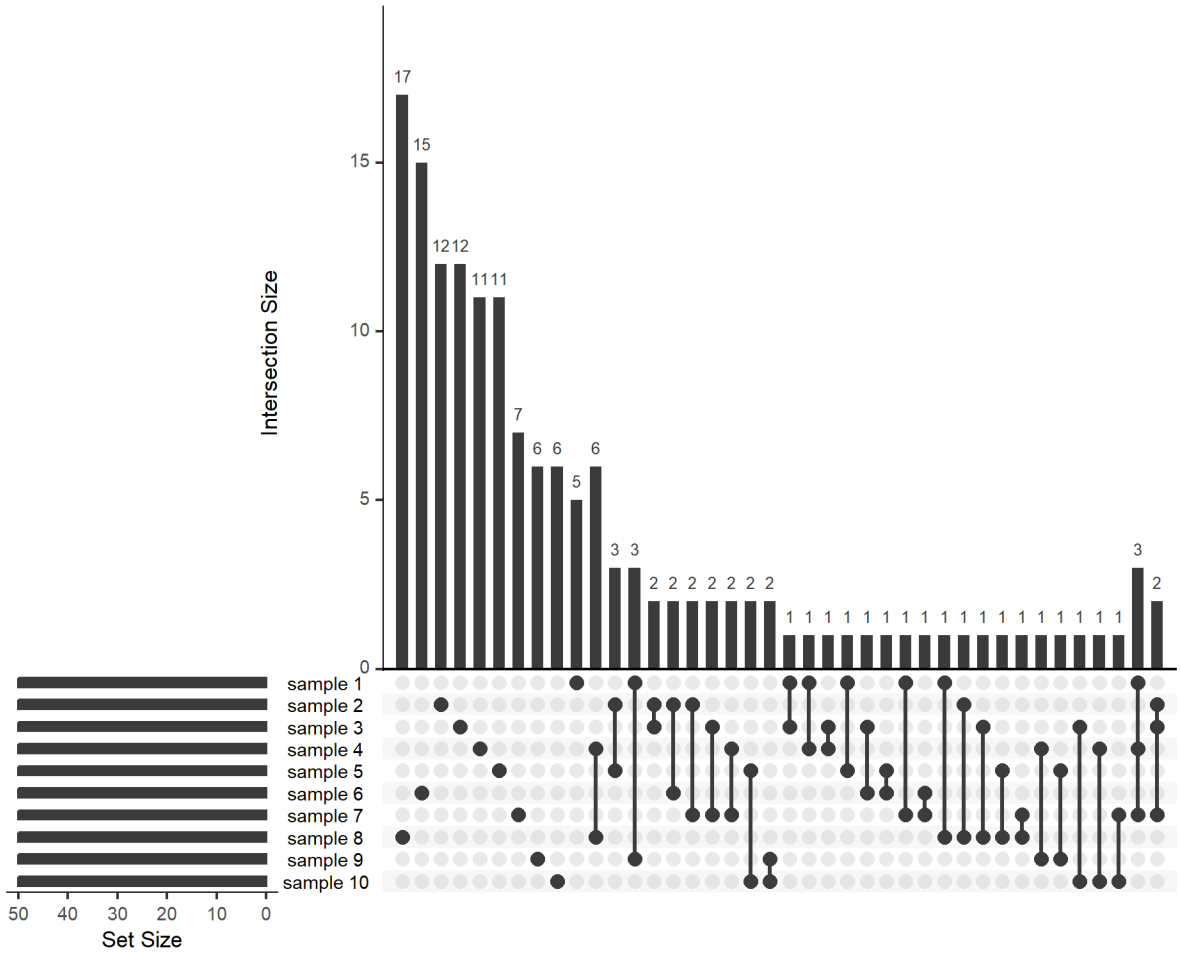
The relative abundance was used for the analysis. Taxonomic categories that were unique to one method were imputed with a value of zero.

(a) (b)



Supplementary Figure 4.14 Comparative Taxonomic Composition of the Microbiome at Levels from Phylum to Species.

(a) DDA-Based. (b) DDA-Free.



Supplementary Figure 4.15 UpSet plot illustrating the overlap in genomes identified by the DDA-free method across ten microbiome samples.

The top 40 intersections were plotted.

Chapter 5: General Discussion

The gastrointestinal tract represents a nutrient-rich environment inhabited by a vast array of microorganisms. Prior to the widespread adoption of high-throughput DNA sequencing technologies, the microorganisms present in the gut are largely unknown, let alone their functions. In 2005, the first application of DNA sequencing techniques on human gut microbiome revealed the high complexity of human gut microbiome[118]. Subsequently, with the rapid advancement of high-throughput DNA sequencing techniques, the human gut microbiome has been precisely identified and categorized. The question "Who are they?" has been well answered. But utilizing metagenomics techniques alone is not enough to address the question, "What do they do?". The gut microbiome has been identified as a critical factor influencing host health. The associations between the human gut microbiome and many diseases have been established and accepted. Beyond these associations, however, understanding the underlying mechanisms holds greater importance. Proteomics identifying and quantifying proteins is a powerful tool for addressing this question. In the past decades, both the mass spectrometry instruments and the bioinformatics in proteomics has been continuously improved. There have even been calls for the complete abandonment of traditional Western blotting techniques[119]. However, employing proteomics to study the human gut microbiome, termed metaproteomics still remains challenging. The unknown and complex nature of human gut microbiome samples poses a significant challenge to the depth of metaproteomic analysis.

In this study, we first evaluated the current depth of detection in metaproteomic analysis (Chapter 2). We mixed varying proportions of isotopically labeled *E. coli* into microbiome samples then attempted to detect *E. coli* from those samples. We observed a rapid decline in the number of identified *E. coli* peptides as the levels of *E. coli* decreased. And *E. coli* becomes undetectable when the mixing ratio falls below 0.5%. We demonstrated, for the first time, the depth of detection in metaproteomics. The intensity of identified peptides showed a good linear relationship with the proportion of mixing, which indirectly demonstrates the accuracy of the detection. In this study, we observe that the MBR technique, on one hand, facilitates the recovery of a lot of low-abundance peptides, yet on the other hand, it also introduces a significant amount of erroneous quantification data. In conventional proteomics, such inaccuracies are substantially mitigated at the protein level, primarily because LFQ algorithm necessitates the identification of at least two peptides per protein[47]. Given that false MBR events are random, the likelihood of two false MBRs occurring simultaneously within the same protein is markedly reduced. However, in metaproteomics, due to the prevalence of shared peptides, protein grouping becomes more complex, sometimes involving hundreds of proteins within a single protein group. Consequently, the quantitative inaccuracies caused by false MBR may still be pronounced. A more effective strategy to eliminate incorrect quantification must be considered when employing MBR in metaproteomics to ensure accuracy and reliability. Nevertheless, it is important to acknowledge by MBR, we have successfully recovered a substantial number of reliable peptides. This observation has drawn our attention to the vast amount of information inherent at the MS1 level. In Chapter 3, we conducted a comparison of the complexity at the MS1 level between single-species samples and

microbiomes. The complexity of microbiome detected by mass spectrometry is significantly higher than that in single species.

The benefit of MBR and the high complexity at the MS1 level in microbiome samples underscore the sensitivity of mass spectrometry in detecting low-abundance peptides. However, due to the inherent stochasticity of DDA, a vast amount of information is only recorded at the MS1 level. DIA, as an unbiased approach, emerges as the future direction for metaproteomics. Consequently, in Chapter 4, we explore the application of DIA within metaproteomics. We have developed a novel strategy (MetaDIA) to narrow down the search space, adapting to the peptide-centric strategy utilized in DIA data analysis. Testing confirmed the accuracy and consistency of our method with the traditional DDA-based method while our method is faster and less resource-intense. Furthermore, our strategy initially assigns a sample-specific taxonomic scope to each sample, significantly mitigating the issue of shared peptides and facilitating subsequent protein grouping. In this workflow, we achieved satisfactory performance by utilizing the top 50 genomes. However, this approach may not be optimal for every sample. Future optimizations could include the implementation of a method to assign a confidence score to each genome, thereby enhancing the accuracy and applicability of the workflow.

In the MetaDIA, an iterative search strategy is employed, wherein the search space expands with each iteration. The initial "seed" database is critical for this methodology. For our approach, we utilized MetaPep, which is specifically designed for the human gut microbiome. Consequently, MetaDIA can only be used for processing samples from the human gut microbiome. What’s more, although MetaPep encompasses identification results from thousands of samples, there remains a potential issue of incompleteness.

When dealing with certain unique samples, it is possible that some species may be overlooked. Since the MAG database for different kinds of samples (Human oral, Pig gut, Mouse gut *et.al*) have been released[120], future research could explore the implementation of this strategy independent of MetaPep. This adjustment would enable its application across a broader spectrum of microbiomes. A potential method could use proteins that are common to all organisms to infer the taxonomic composition of samples. Additionally, the functional abundance could be determined from a relatively smaller dataset.

In summary, our research represents the first assessment of the dark field of the metaproteome. We have shown the power of the MBR algorithm in retrieving lost data. Subsequently, we provided evidence at the level of mass spectrum that the microbiome exhibits greater complexity compared to single species. This information suggests that the current DDA approach does not effectively leverage the sensitivity of mass spectrometry. A significant number of signals detected by mass spectrometry are not utilized for peptide identification and quantification. Consequently, we have developed a workflow, MetaDIA, which enables the application of DIA to metaproteomics. I believe that this work will enhance our understanding of gut microbiome.

Curriculum Vitae

Academic Background

Ph.D. Biochemistry 2020-present

Supervisor: Dr. Daniel Figeys

Faculty of Medicine

University of Ottawa

M.Sc. Medicinal Chemistry 2017-2020

Supervisor: Dr. Jianhua Shen

Shanghai Institute of Materia Medica

University of Chinese Academy of Sciences

B.Sc. Medicinal Chemistry 2013-2017

School of Pharmacy

China Pharmaceutical University

Publications:

- Duan, H.; Cheng, K.; Ning, Z.; Li, L.; Mayne, J.; Sun, Z.; Figeys, D., Assessing the Dark Field of Metaproteome. Anal Chem 2022, 94 (45), 15648-15654.
- Sun, Z.; Ning, Z.; Cheng, K.; Duan, H.; Wu, Q.; Mayne, J.; Figeys, D., MetaPep: A core peptide database for faster human gut metaproteomics database searches. Comput Struct Biotechnol J 2023, 21, 4228-4237.
- Zheng, X.; Duan, H.; Lin, F.; Li, X.; Shen, J.; Han, F.; Huang, F.; Li, S.; Chang, L.; Xu, H.; Wang, K.; Liu, J., Quantification of microbiota-related phenols and aromatic acids in mouse feces of a diabetic nephropathy model by simultaneous BDAPE derivatization using ultra-performance liquid chromatography-tandem mass spectrometry. Anal Bioanal Chem 2020, 412 (13), 3241-3252.
- Duan, H.; Zhang, X.; Figeys, D., An emerging field: Post-translational modification in microbiome. Proteomics 2022, e2100389.

- Zhu, N.; Duan, H.; Feng, Y.; Xu, W.; Shen, J.; Wang, K.; Liu, J., Magnesium lithospermate B ameliorates diabetic nephropathy by suppressing the uremic toxin formation mediated by gut microbiota. Eur J Pharmacol 2023, 953, 175812.
- Sauve, M. F.; Feldman, F.; Sane, A. T.; Koudoufio, M.; Patey, N.; Spahis, S.; Butcher, J.; Duan, H.; Figeys, D.; Marcil, V.; Stintzi, A.; Levy, E., Glycomacropeptide as an Efficient Agent to Fight Pathophysiological Mechanisms of Metabolic Syndrome. Nutrients 2024, 16 (6).
- Fu, H.; Zhang, Q. L.; Huang, X. W.; Ma, Z. H.; Zheng, X. L.; Li, S. L.; Duan, H. N.; Sun, X. C.; Lin, F. F.; Zhao, L. J.; Teng, G. S.; Liu, J., A rapid and convenient derivatization method for quantitation of short-chain fatty acids in human feces by ultra-performance liquid chromatography/tandem mass spectrometry. Rapid Commun Mass Spectrom 2020, 34 (9), e8730.
- Li, S.; Zhu, N.; Tang, C.; Duan, H.; Wang, Y.; Zhao, G.; Liu, J.; Ye, Y., Differential distribution of characteristic constituents in root, stem and leaf tissues of Salvia miltiorrhiza using MALDI mass spectrometry imaging. Fitoterapia 2020, 146, 104679.
- Duan, H.; Ning, Z.; Sun, Z.; Guo, T.; Sun, Y.; Figeys, D., MetaDIA: A Novel Database Reduction Strategy for DIA Human Gut Metaproteomics. bioRxiv 2024, 2024.03. 14.585104. Submitted to Microbiome.
- H. Duan, Z. Ning, A. Zhang, D. Figeys, Spectral entropy as a measure of the metaproteome complexity. Submitted to Proteomics.

References

[1] R. Sender, S. Fuchs, R. Milo, Revised Estimates for the Number of Human and Bacteria Cells in the Body, PLoS Biol 14(8) (2016) e1002533.
[2] E. Thursby, N. Juge, Introduction to the human gut microbiota, Biochem J 474(11) (2017) 1823-1836.
[3] F. Baquero, C. Nombela, The microbiome as a human organ, Clin Microbiol Infect 18 Suppl 4 (2012) 2-4.
[4] M. Matijasic, T. Mestrovic, H.C. Paljetak, M. Peric, A. Baresic, D. Verbanac, Gut Microbiota beyond Bacteria-Mycobiome, Virome, Archaeome, and Eukaryotic Parasites in IBD, Int J Mol Sci 21(8) (2020).
[5] J. Yang, J. Pu, S. Lu, X. Bai, Y. Wu, D. Jin, Y. Cheng, G. Zhang, W. Zhu, X. Luo, R. Rossello-Mora, J. Xu, Species-Level Analysis of Human Gut Microbiota With Metataxonomics, Front Microbiol 11 (2020) 2029.
[6] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J.R. Kultima, E. Prifti, T. Nielsen, A.S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J.Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H.B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Dore, S.D. Ehrlich, H.I.T.C. Meta, P. Bork, J. Wang, H.I.T.C. Meta, An integrated catalog of reference genes in the human gut microbiome, Nat Biotechnol 32(8) (2014) 834-41.
[7] J. Lloyd-Price, A. Mahurkar, G. Rahnavard, J. Crabtree, J. Orvis, A.B. Hall, A. Brady, H.H. Creasy, C. McCracken, M.G. Giglio, D. McDonald, E.A. Franzosa, R. Knight, O. White, C. Huttenhower, Strains, functions and dynamics in the expanded Human Microbiome Project, Nature 550(7674) (2017) 61-66.
[8] A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z.J. Shi, K.S. Pollard, E. Sakharova, D.H. Parks, P. Hugenholtz, N. Segata, N.C. Kyrpides, R.D. Finn, A unified catalog of 204,938 reference genomes from the human gut microbiome, Nat Biotechnol 39(1) (2021) 105-114.
[9] A. Moya, M. Ferrer, Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance, Trends Microbiol 24(5) (2016) 402-413.
[10] L. Li, T. Wang, Z. Ning, X. Zhang, J. Butcher, J.M. Serrana, C.M.A. Simopoulos, J. Mayne, A. Stintzi, D.R. Mack, Y.Y. Liu, D. Figeys, Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics, Nat Commun 14(1) (2023) 3428.
[11] L. Tian, X.W. Wang, A.K. Wu, Y. Fan, J. Friedman, A. Dahlin, M.K. Waldor, G.M. Weinstock, S.T. Weiss, Y.Y. Liu, Deciphering functional redundancy in the human microbiome, Nat Commun 11(1) (2020) 6217.
[12] Y. Fan, O. Pedersen, Gut microbiota in human metabolic health and disease, Nat Rev Microbiol 19(1) (2021) 55-71.
[13] R.G. Armani, A. Ramezani, A. Yasir, S. Sharama, M.E.F. Canziani, D.S. Raj, Gut Microbiome in Chronic Kidney Disease, Curr Hypertens Rep 19(4) (2017) 29.
[14] M. Lee, E.B. Chang, Inflammatory Bowel Diseases (IBD) and the Microbiome-Searching the Crime Scene for Clues, Gastroenterology 160(2) (2021) 524-537.
[15] W.E. Roediger, Utilization of nutrients by isolated epithelial cells of the rat colon, Gastroenterology 83(2) (1982) 424-9.
[16] S.M. McNabney, T.M. Henagan, Short Chain Fatty Acids in the Colon and Peripheral Tissues: A Focus on Butyrate, Colon Cancer, Obesity and Insulin Resistance, Nutrients 9(12) (2017).

[17] T. Gryp, R. Vanholder, M. Vaneechoutte, G. Glorieux, p-Cresyl Sulfate, Toxins (Basel) 9(2) (2017).
[18] C.N. Hsu, Y.L. Tain, Developmental Programming and Reprogramming of Hypertension and Kidney Disease: Impact of Tryptophan Metabolism, Int J Mol Sci 21(22) (2020).
[19] J.M. Pickard, M.Y. Zeng, R. Caruso, G. Nunez, Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease, Immunol Rev 279(1) (2017) 70-89.
[20] P. Paone, P.D. Cani, Mucus barrier, mucins and gut microbiota: the expected slimy partners?, Gut 69(12) (2020) 2232-2243.
[21] N.M. Koropatkin, E.A. Cameron, E.C. Martens, How glycan metabolism shapes the human gut microbiota, Nat Rev Microbiol 10(5) (2012) 323-35.
[22] M.S. Desai, A.M. Seekatz, N.M. Koropatkin, N. Kamada, C.A. Hickey, M. Wolter, N.A. Pudlo, S. Kitamoto, N. Terrapon, A. Muller, V.B. Young, B. Henrissat, P. Wilmes, T.S. Stappenbeck, G. Nunez, E.C. Martens, A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility, Cell 167(5) (2016) 1339-1353 e21.
[23] M.E. Johansson, H.E. Jakobsson, J. Holmen-Larsson, A. Schutte, A. Ermund, A.M. Rodriguez-Pineiro, L. Arike, C. Wising, F. Svensson, F. Backhed, G.C. Hansson, Normalization of Host Intestinal Mucus Layers Requires Long-Term Microbial Colonization, Cell Host Microbe 18(5) (2015) 582-92.
[24] S. Johnson, V. Lavergne, A.M. Skinner, A.J. Gonzales-Luna, K.W. Garey, C.P. Kelly, M.H. Wilcox, Clinical Practice Guideline by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA): 2021 Focused Update Guidelines on Management of Clostridioides difficile Infection in Adults, Clin Infect Dis 73(5) (2021) e1029-e1044.
[25] B.J. Gawey, S. Khanna, Clostridioides difficile Infection: Landscape and Microbiome Therapeutics, Gastroenterol Hepatol (N Y) 19(6) (2023) 319-328.
[26] S. Khanna, M. Assi, C. Lee, D. Yoho, T. Louie, W. Knapple, H. Aguilar, J. Garcia-Diaz, G.P. Wang, S.M. Berry, J. Marion, X. Su, T. Braun, L. Bancke, P. Feuerstadt, Efficacy and Safety of RBX2660 in PUNCH CD3, a Phase III, Randomized, Double-Blind, Placebo-Controlled Trial with a Bayesian Primary Analysis for the Prevention of Recurrent Clostridioides difficile Infection, Drugs 82(15) (2022) 1527-1538.
[27] N. Zhu, H. Duan, Y. Feng, W. Xu, J. Shen, K. Wang, J. Liu, Magnesium lithospermate B ameliorates diabetic nephropathy by suppressing the uremic toxin formation mediated by gut microbiota, Eur J Pharmacol 953 (2023) 175812.
[28] S. Nayfach, Z.J. Shi, R. Seshadri, K.S. Pollard, N.C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome, Nature 568(7753) (2019) 505-510.
[29] Q. Tang, G. Jin, G. Wang, T. Liu, X. Liu, B. Wang, H. Cao, Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices, Front Cell Infect Microbiol 10 (2020) 151.
[30] J.A. McDonald, K. Schroeter, S. Fuentes, I. Heikamp-Dejong, C.M. Khursigara, W.M. de Vos, E. Allen-Vercoe, Evaluation of microbial community reproducibility, stability and composition in a human distal gut chemostat model, J Microbiol Methods 95(2) (2013) 167-74.
[31] J.M. Auchtung, C.D. Robinson, R.A. Britton, Cultivation of stable, reproducible microbial communities from different fecal donors using minibioreactor arrays (MBRAs), Microbiome 3 (2015) 42.
[32] L. Li, Z. Ning, X. Zhang, J. Mayne, K. Cheng, A. Stintzi, D. Figeys, RapidAIM: a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs, Microbiome 8(1) (2020) 33.
[33] P.D. Schloss, J. Handelsman, Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness, Appl Environ Microbiol 71(3) (2005) 1501-6.

[34] P.J. Turnbaugh, R.E. Ley, M. Hamady, C.M. Fraser-Liggett, R. Knight, J.I. Gordon, The human microbiome project, Nature 449(7164) (2007) 804-10.
[35] E.A. Franzosa, X.C. Morgan, N. Segata, L. Waldron, J. Reyes, A.M. Earl, G. Giannoukos, M.R. Boylan, D. Ciulla, D. Gevers, J. Izard, W.S. Garrett, A.T. Chan, C. Huttenhower, Relating the metatranscriptome and metagenome of the human gut, Proc Natl Acad Sci U S A 111(22) (2014) E2329-38.
[36] P. Vernocchi, F. Del Chierico, L. Putignani, Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health, Front Microbiol 7 (2016) 1144.
[37] D. Kumar, G. Bansal, A. Narang, T. Basak, T. Abbas, D. Dash, Integrating transcriptome and proteome profiling: Strategies and applications, Proteomics 16(19) (2016) 2533-2544.
[38] S.P. Gygi, Y. Rochon, B.R. Franza, R. Aebersold, Correlation between protein and mRNA abundance in yeast, Mol Cell Biol 19(3) (1999) 1720-30.
[39] G. den Besten, K. van Eunen, A.K. Groen, K. Venema, D.J. Reijngoud, B.M. Bakker, The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism, J Lipid Res 54(9) (2013) 2325-40.
[40] H. Zeng, S. Umar, B. Rust, D. Lazarova, M. Bordonaro, Secondary Bile Acids and Short Chain Fatty Acids in the Colon: A Focus on Colonic Microbiome, Cell Proliferation, Inflammation, and Cancer, Int J Mol Sci 20(5) (2019).
[41] A. Agus, J. Planchais, H. Sokol, Gut Microbiota Regulation of Tryptophan Metabolism in Health and Disease, Cell Host Microbe 23(6) (2018) 716-724.
[42] T.A. Costa-Silva, J.C. Flores-Santos, R.K.B. Freire, M. Vitolo, A. Pessoa-Jr, Microbial cell disruption methods for efficient release of enzyme L-asparaginase, Prep Biochem Biotechnol 48(8) (2018) 707-717.
[43] X. Zhang, L. Li, J. Mayne, Z. Ning, A. Stintzi, D. Figeys, Assessing the impact of protein extraction methods for human gut metaproteomics, J Proteomics 180 (2018) 120-127.
[44] J. Wang, X. Zhang, L. Li, Z. Ning, J. Mayne, C. Schmitt-Ulms, K. Walker, K. Cheng, D. Figeys, Differential Lysis Approach Enables Selective Extraction of Taxon-Specific Proteins for Gut Metaproteomics, Anal Chem 92(7) (2020) 5379-5386.
[45] X. Zhang, D. Figeys, Perspective and Guidelines for Metaproteomics in Microbiome Studies, J Proteome Res 18(6) (2019) 2370-2380.
[46] M. Bauer, E. Ahrne, A.P. Baron, T. Glatter, L.L. Fava, A. Santamaria, E.A. Nigg, A. Schmidt, Evaluation of data-dependent and -independent mass spectrometric workflows for sensitive quantification of proteins and phosphorylation sites, J Proteome Res 13(12) (2014) 5973-88.
[47] M.Y. Lim, J.A. Paulo, S.P. Gygi, Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model, J Proteome Res 18(11) (2019) 4020-4026.
[48] X. Zhang, W. Chen, Z. Ning, J. Mayne, D. Mack, A. Stintzi, R. Tian, D. Figeys, Deep Metaproteomics Approach for the Study of Human Microbiomes, Anal Chem 89(17) (2017) 9407-9415.
[49] L.C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis, Mol Cell Proteomics 11(6) (2012) O111 016717.
[50] J.D. Chapman, D.R. Goodlett, C.D. Masselon, Multiplexed and data-independent tandem mass spectrometry for global proteome profiling, Mass Spectrom Rev 33(6) (2014) 452-70.
[51] B.C. Collins, C.L. Hunter, Y. Liu, B. Schilling, G. Rosenberger, S.L. Bader, D.W. Chan, B.W. Gibson, A.C. Gingras, J.M. Held, M. Hirayama-Kurogi, G. Hou, C. Krisp, B. Larsen, L. Lin, S. Liu, M.P. Molloy, R.L. Moritz, S. Ohtsuki, R. Schlapbach, N. Selevsek, S.N. Thomas, S.C. Tzeng, H. Zhang, R. Aebersold, Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry, Nat Commun 8(1) (2017) 291.

[52] J. Vowinckel, A. Zelezniak, R. Bruderer, M. Mulleder, L. Reiter, M. Ralser, Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition, Sci Rep 8(1) (2018) 4346.
[53] F. Meier, A.D. Brunner, M. Frank, A. Ha, I. Bludau, E. Voytik, S. Kaspar-Schoenefeld, M. Lubeck, O. Raether, N. Bache, R. Aebersold, B.C. Collins, H.L. Rost, M. Mann, diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition, Nat Methods 17(12) (2020) 1229-1236.
[54] U.H. Guzman, A. Martinez-Val, Z. Ye, E. Damoc, T.N. Arrey, A. Pashkova, S. Renuse, E. Denisov, J. Petzoldt, A.C. Peterson, F. Harking, O. Ostergaard, R. Rydbirk, S. Aznar, H. Stewart, Y. Xuan, D. Hermanson, S. Horning, C. Hock, A. Makarov, V. Zabrouskov, J.V. Olsen, Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition, Nat Biotechnol (2024).
[55] A.R. Erickson, B.L. Cantarel, R. Lamendella, Y. Darzi, E.F. Mongodin, C. Pan, M. Shah, J. Halfvarson, C. Tysk, B. Henrissat, J. Raes, N.C. Verberkmoes, C.M. Fraser, R.L. Hettich, J.K. Jansson, Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease, PLoS One 7(11) (2012) e49138.
[56] H. Daniel, A.M. Gholami, D. Berry, C. Desmarchelier, H. Hahne, G. Loh, S. Mondot, P. Lepage, M. Rothballer, A. Walker, C. Bohm, M. Wenning, M. Wagner, M. Blaut, P. Schmitt-Kopplin, B. Kuster, D. Haller, T. Clavel, High-fat diet alters gut microbiota physiology in mice, ISME J 8(2) (2014) 295-308.
[57] J. Qin, R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D.R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H.B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, H.I.T.C. Meta, P. Bork, S.D. Ehrlich, J. Wang, A human gut microbial gene catalogue established by metagenomic sequencing, Nature 464(7285) (2010) 59-65.
[58] C. Human Microbiome Project, A framework for human microbiome research, Nature 486(7402) (2012) 215-21.
[59] Z. Sun, Z. Ning, K. Cheng, H. Duan, Q. Wu, J. Mayne, D. Figeys, MetaPep: A core peptide database for faster human gut metaproteomics database searches, Comput Struct Biotechnol J 21 (2023) 4228-4237.
[60] J.E. Elias, S.P. Gygi, Target-decoy search strategy for mass spectrometry-based proteomics, Methods Mol Biol 604 (2010) 55-71.
[61] P. Jagtap, T. McGowan, S. Bandhakavi, Z.J. Tu, S. Seymour, T.J. Griffin, J.D. Rudney, Deep metaproteomic analysis of human salivary supernatant, Proteomics 12(7) (2012) 992-1001.
[62] P. Jagtap, J. Goslinga, J.A. Kooren, T. McGowan, M.S. Wroblewski, S.L. Seymour, T.J. Griffin, A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies, Proteomics 13(8) (2013) 1352-7.
[63] X. Zhang, Z. Ning, J. Mayne, J.I. Moore, J. Li, J. Butcher, S.A. Deeke, R. Chen, C.K. Chiang, M. Wen, D. Mack, A. Stintzi, D. Figeys, MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota, Microbiome 4(1) (2016) 31.
[64] K. Cheng, Z. Ning, X. Zhang, L. Li, B. Liao, J. Mayne, A. Stintzi, D. Figeys, MetaLab: an automated pipeline for metaproteomic data analysis, Microbiome 5(1) (2017) 157.
[65] K. Cheng, Z. Ning, X. Zhang, L. Li, B. Liao, J. Mayne, D. Figeys, MetaLab 2.0 Enables Accurate Post-Translational Modifications Profiling in Metaproteomics, J Am Soc Mass Spectrom 31(7) (2020) 1473-1482.
[66] M. Stamboulian, S. Li, Y. Ye, Using high-abundance proteins as guides for fast and effective peptide/protein identification from human gut metaproteomic data, Microbiome 9(1) (2021).

[67] K. Cheng, Z. Ning, L. Li, X. Zhang, J.M. Serrana, J. Mayne, D. Figeys, MetaLab-MAG: A Metaproteomic Data Analysis Platform for Genome-Level Characterization of Microbiomes from the Metagenome-Assembled Genomes Database, J Proteome Res 22(2) (2023) 387-398.
[68] T. Yang, T. Ling, B. Sun, Z. Liang, F. Xu, X. Huang, L. Xie, Y. He, L. Li, F. He, Y. Wang, C. Chang, Introducing pi-HelixNovo for practical large-scale de novo peptide sequencing, Brief Bioinform 25(2) (2024).
[69] B. Behsaz, H. Mohimani, A. Gurevich, A. Prjibelski, M. Fisher, F. Vargas, L. Smarr, P.C. Dorrestein, J.S. Mylne, P.A. Pevzner, De Novo Peptide Sequencing Reveals Many Cyclopeptides in the Human Gut and Other Environments, Cell Syst 10(1) (2020) 99-108 e5.
[70] D. Gomez-Varela, F. Xian, S. Grundtner, J.R. Sondermann, G. Carta, M. Schmidt, Increasing taxonomic and functional characterization of host-microbiome interactions by DIA-PASEF metaproteomics, Front Microbiol 14 (2023) 1258703.
[71] Y. Sun, Z. Xing, S. Liang, Z. Miao, L.-b. Zhuo, W. Jiang, H. Zhao, H. Gao, Y. Xie, Y. Zhou, metaExpertPro: a computational workflow for metaproteomics spectral library construction and data-independent acquisition mass spectrometry data analysis, bioRxiv (2023) 2023.11. 29.569331.
[72] J. Aakko, S. Pietila, T. Suomi, M. Mahmoudian, R. Toivonen, P. Kouvonen, A. Rokka, A. Hanninen, L.L. Elo, Data-Independent Acquisition Mass Spectrometry in Metaproteomics of Gut Microbiota-Implementation and Computational Analysis, J Proteome Res 19(1) (2020) 432-436.
[73] Y.S. Ting, J.D. Egertson, S.H. Payne, S. Kim, B. MacLean, L. Kall, R. Aebersold, R.D. Smith, W.S. Noble, M.J. MacCoss, Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data, Mol Cell Proteomics 14(9) (2015) 2301-7.
[74] C.C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.C. Gingras, A.I. Nesvizhskii, DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics, Nat Methods 12(3) (2015) 258-64, 7 p following 264.
[75] Y. Li, C.Q. Zhong, X. Xu, S. Cai, X. Wu, Y. Zhang, J. Chen, J. Shi, S. Lin, J. Han, Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files, Nat Methods 12(12) (2015) 1105-6.
[76] S. Pietilä, T. Suomi, L.L. Elo, Introducing untargeted data-independent acquisition for metaproteomics of complex microbial samples, ISME Communications 2(1) (2022) 1-8.
[77] C.R. Weisbrod, J.K. Eng, M.R. Hoopmann, T. Baker, J.E. Bruce, Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification, J Proteome Res 11(3) (2012) 1621-32.
[78] Y.S. Ting, J.D. Egertson, J.G. Bollinger, B.C. Searle, S.H. Payne, W.S. Noble, M.J. MacCoss, PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data, Nat Methods 14(9) (2017) 903-908.
[79] V. Demichev, C.B. Messner, S.I. Vernardis, K.S. Lilley, M. Ralser, DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput, Nat Methods 17(1) (2020) 41-44.
[80] J.C. Clemente, L.K. Ursell, L.W. Parfrey, R. Knight, The impact of the gut microbiota on human health: an integrative view, Cell 148(6) (2012) 1258-70.
[81] R. Sender, S. Fuchs, R. Milo, Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans, Cell 164(3) (2016) 337-40.
[82] H. Chi, C. Liu, H. Yang, W.F. Zeng, L. Wu, W.J. Zhou, R.M. Wang, X.N. Niu, Y.H. Ding, Y. Zhang, Z.W. Wang, Z.L. Chen, R.X. Sun, T. Liu, G.M. Tan, M.Q. Dong, P. Xu, P.H. Zhang, S.M. He, Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine, Nat Biotechnol (2018).
[83] R.J. Millikin, S.K. Solntsev, M.R. Shortreed, L.M. Smith, Ultrafast Peptide Label-Free Quantification with FlashLFQ, J Proteome Res 17(1) (2018) 386-391.

[84] M.Y. Galperin, Y.I. Wolf, K.S. Makarova, R. Vera Alvarez, D. Landsman, E.V. Koonin, COG database update: focus on microbial diversity, model organisms, and widespread pathogens, Nucleic Acids Res 49(D1) (2021) D274-D281.
[85] M. Kleiner, E. Thorson, C.E. Sharp, X. Dong, D. Liu, C. Li, M. Strous, Assessing species biomass contributions in microbial communities via metaproteomics, Nat Commun 8(1) (2017) 1558.
[86] J. Cox, M.Y. Hein, C.A. Luber, I. Paron, N. Nagaraj, M. Mann, Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ, Mol Cell Proteomics 13(9) (2014) 2513-26.
[87] R.A. Maronna, R.D. Martin, V.J. Yohai, M. Salibián-Barrera, Robust statistics: theory and methods (with R), John Wiley & Sons2019.
[88] C. Human Microbiome Project, Structure, function and diversity of the healthy human microbiome, Nature 486(7402) (2012) 207-14.
[89] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nat Methods 4(3) (2007) 207-14.
[90] W.S. Noble, Mass spectrometrists should search only for peptides they care about, Nat Methods 12(7) (2015) 605-8.
[91] H. Duan, K. Cheng, Z. Ning, L. Li, J. Mayne, Z. Sun, D. Figeys, Assessing the Dark Field of Metaproteome, Anal Chem 94(45) (2022) 15648-15654.
[92] S. Tyanova, T. Temu, J. Cox, The MaxQuant computational platform for mass spectrometry-based shotgun proteomics, Nat Protoc 11(12) (2016) 2301-2319.
[93] H.L. Rost, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W.E. Wolski, O. Schilling, J.S. Choudhary, L. Malmstrom, R. Aebersold, K. Reinert, O. Kohlbacher, OpenMS: a flexible open-source software platform for mass spectrometry data analysis, Nat Methods 13(9) (2016) 741-8.
[94] H. Weisser, S. Nahnsen, J. Grossmann, L. Nilse, A. Quandt, H. Brauer, M. Sturm, E. Kenar, O. Kohlbacher, R. Aebersold, L. Malmstrom, An automated pipeline for high-throughput label-free quantitative proteomics, J Proteome Res 12(4) (2013) 1628-44.
[95] Y. Li, T. Kind, J. Folz, A. Vaniya, S.S. Mehta, O. Fiehn, Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification, Nat Methods 18(12) (2021) 1524-1531.
[96] Y. Li, O. Fiehn, Flash entropy search to query all mass spectral libraries in real time, Nat Methods 20(10) (2023) 1475-1478.
[97] A. Michalski, J. Cox, M. Mann, More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS, J Proteome Res 10(4) (2011) 1785-93.
[98] D.L. Tabb, L. Vega-Montoto, P.A. Rudnick, A.M. Variyath, A.J. Ham, D.M. Bunk, L.E. Kilpatrick, D.D. Billheimer, R.K. Blackman, H.L. Cardasis, S.A. Carr, K.R. Clauser, J.D. Jaffe, K.A. Kowalski, T.A. Neubert, F.E. Regnier, B. Schilling, T.J. Tegeler, M. Wang, P. Wang, J.R. Whiteaker, L.J. Zimmerman, S.J. Fisher, B.W. Gibson, C.R. Kinsinger, M. Mesri, H. Rodriguez, S.E. Stein, P. Tempst, A.G. Paulovich, D.C. Liebler, C. Spiegelman, Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry, J Proteome Res 9(2) (2010) 761-76.
[99] H. Liu, R.G. Sadygov, J.R. Yates, 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics, Anal Chem 76(14) (2004) 4193-201.
[100] C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B.C. Collins, R. Aebersold, Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial, Mol Syst Biol 14(8) (2018) e8126.

[101] L.R. Thompson, J.G. Sanders, D. McDonald, A. Amir, J. Ladau, K.J. Locey, R.J. Prill, A. Tripathi, S.M. Gibbons, G. Ackermann, J.A. Navas-Molina, S. Janssen, E. Kopylova, Y. Vazquez-Baeza, A. Gonzalez, J.T. Morton, S. Mirarab, Z. Zech Xu, L. Jiang, M.F. Haroon, J. Kanbar, Q. Zhu, S. Jin Song, T. Kosciolek, N.A. Bokulich, J. Lefler, C.J. Brislawn, G. Humphrey, S.M. Owens, J. Hampton-Marcell, D. Berg-Lyons, V. McKenzie, N. Fierer, J.A. Fuhrman, A. Clauset, R.L. Stevens, A. Shade, K.S. Pollard, K.D. Goodwin, J.K. Jansson, J.A. Gilbert, R. Knight, C. Earth Microbiome Project, A communal catalogue reveals Earth's multiscale microbial diversity, Nature 551(7681) (2017) 457-463.
[102] P. Wilmes, P.L. Bond, Metaproteomics: studying functional gene expression in microbial ecosystems, Trends Microbiol 14(2) (2006) 92-7.
[103] S. Sun, F. Yang, Q. Yang, H. Zhang, Y. Wang, D. Bu, B. Ma, MS-Simulator: predicting y-ion intensities for peptides with two charges based on the intensity ratio of neighboring ions, J Proteome Res 11(9) (2012) 4509-16.
[104] W.F. Zeng, X.X. Zhou, W.J. Zhou, H. Chi, J. Zhan, S.M. He, MS/MS Spectrum Prediction for Modified Peptides Using pDeep2 Trained by Transfer Learning, Anal Chem 91(15) (2019) 9724-9731.
[105] R. Bouwmeester, R. Gabriels, N. Hulstaert, L. Martens, S. Degroeve, DeepLC can predict retention times for peptides that carry as-yet unseen modifications, Nat Methods 18(11) (2021) 1363-1369.
[106] C. Ma, Y. Ren, J. Yang, Z. Ren, H. Yang, S. Liu, Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning, Anal Chem 90(18) (2018) 10881-10888.
[107] O. Al Musaimi, O.M.M. Valenzo, D.R. Williams, Prediction of peptides retention behavior in reversed-phase liquid chromatography based on their hydrophobicity, J Sep Sci 46(2) (2023) e2200743.
[108] J. Cox, Prediction of peptide mass spectral libraries with machine learning, Nat Biotechnol 41(1) (2023) 33-43.
[109] V. Demichev, L. Szyrwiel, F. Yu, G.C. Teo, G. Rosenberger, A. Niewienda, D. Ludwig, J. Decker, S. Kaspar-Schoenefeld, K.S. Lilley, M. Mulleder, A.I. Nesvizhskii, M. Ralser, dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts, Nat Commun 13(1) (2022) 3944.
[110] P. Sinitcyn, H. Hamzeiy, F. Salinas Soto, D. Itzhak, F. McCarthy, C. Wichmann, M. Steger, U. Ohmayer, U. Distler, S. Kaspar-Schoenefeld, N. Prianichnikov, S. Yilmaz, J.D. Rudolph, S. Tenzer, Y. Perez-Riverol, N. Nagaraj, S.J. Humphrey, J. Cox, MaxDIA enables library-based and library-free data-independent acquisition proteomics, Nat Biotechnol (2021).
[111] D.B. Bekker-Jensen, O.M. Bernhardt, A. Hogrebe, A. Martinez-Val, L. Verbeke, T. Gandhi, C.D. Kelstrup, L. Reiter, J.V. Olsen, Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries, Nat Commun 11(1) (2020) 787.
[112] J. Yang, Z. Cheng, F. Gong, Y. Fu, DeepDetect: Deep Learning of Peptide Detectability Enhanced by Peptide Digestibility and Its Application to DIA Library Reduction, Anal Chem 95(15) (2023) 6235-6243.
[113] A.G. Davis-Richardson, A.N. Ardissone, R. Dias, V. Simell, M.T. Leonard, K.M. Kemppainen, J.C. Drew, D. Schatz, M.A. Atkinson, B. Kolaczkowski, J. Ilonen, M. Knip, J. Toppari, N. Nurminen, H. Hyoty, R. Veijola, T. Simell, J. Mykkanen, O. Simell, E.W. Triplett, Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes, Front Microbiol 5 (2014) 678.
[114] K. Hosomi, M. Saito, J. Park, H. Murakami, N. Shibata, M. Ando, T. Nagatake, K. Konishi, H. Ohno, K. Tanisawa, A. Mohsen, Y.A. Chen, H. Kawashima, Y. Natsume-Kitatani, Y. Oka, H. Shimizu, M. Furuta, Y. Tojima, K. Sawane, A. Saika, S. Kondo, Y. Yonejima, H. Takeyama, A.

Matsutani, K. Mizuguchi, M. Miyachi, J. Kunisawa, Oral administration of Blautia wexlerae ameliorates obesity and type 2 diabetes via metabolic remodeling of the gut microbiota, Nat Commun 13(1) (2022) 4477.
[115] J.U. Scher, A. Sczesnak, R.S. Longman, N. Segata, C. Ubeda, C. Bielski, T. Rostron, V. Cerundolo, E.G. Pamer, S.B. Abramson, C. Huttenhower, D.R. Littman, Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis, Elife 2 (2013) e01202.
[116] P.T.N. Lan, M. Sakamoto, S. Sakata, Y. Benno, Bacteroides barnesiae sp. nov., Bacteroides salanitronis sp. nov. and Bacteroides gallinarum sp. nov., isolated from chicken caecum, Int J Syst Evol Microbiol 56(Pt 12) (2006) 2853-2859.
[117] C.V. Ferreira-Halder, A.V.S. Faria, S.S. Andrade, Action and function of Faecalibacterium prausnitzii in health and disease, Best Pract Res Clin Gastroenterol 31(6) (2017) 643-648.
[118] P.B. Eckburg, E.M. Bik, C.N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S.R. Gill, K.E. Nelson, D.A. Relman, Diversity of the human intestinal microbial flora, Science 308(5728) (2005) 1635-8.
[119] D. Mehta, A.H. Ahkami, J. Walley, S.L. Xu, R.G. Uhrig, The incongruity of validating quantitative proteomics using western blots, Nat Plants 8(12) (2022) 1320-1321.
[120] T.A. Gurbich, A. Almeida, M. Beracochea, T. Burdett, J. Burgin, G. Cochrane, S. Raj, L. Richardson, A.B. Rogers, E. Sakharova, G.A. Salazar, R.D. Finn, MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues, J Mol Biol 435(14) (2023) 168016.

Copyrights and Statement

As a non-native English speaker, I sought to enhance the clarity and coherence of my writing with the help of large language model. Throughout the process of composing this thesis, I utilized ChatGPT from OpenAI to obtain suggestions and assistance in refining the language. ChatGPT was employed solely to assist in polishing the language and not for generating sentences. I have meticulously scrutinized words and expressions in this thesis to ensure they accurately convey my intended meaning.