

DIVING DEEP INTO HUMAN GUT METAPROTEOMICS: INSIGHTS FROM PROTEIN TO PEPTIDE LEVEL

ZHONGZHI SUN

THESIS SUBMITTED TO THE UNIVERSITY OF OTTAWA IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE
DOCTORATE IN PHILOSOPHY DEGREE IN BIOCHEMISTRY

DEPARTMENT OF BIOCHEMISTRY, MICROBIOLOGY AND IMMUNOLOGY
FACULTY OF MEDICINE
UNIVERSITY OF OTTAWA



©ZHONGZHI SUN, OTTAWA, CANADA, 2025

ABSTRACT

The human gut microbiome has a complex and dynamic relationship with its host and is closely associated with human health and disease states. As an emerging high-throughput omics technique, metaproteomics measures the presence and abundance of proteins in the microbial community and serves as a powerful tool for investigating interactions between the human gut microbiome and the host. However, metaproteomics faces unique challenges due to the complexity of microbial communities. Limited proteome coverage in metaproteomics has recently been highlighted, suggesting it could obscure important biological discoveries. This issue is further compounded by information loss during protein inference. As mass spectrometry-based (meta)proteomics directly measures peptides rather than proteins, information loss occurs due to the difficulty in precisely attributing peptides to their parent proteins. Despite this, most current metaproteomic studies still rely on protein-level analyses, and there are fewer applicable tools for peptide-level analysis. To address these challenges, this thesis enhances traditional protein-level metaproteomics with peptide-level analysis, develops practical tools to reduce information loss.

First, we explored the potential of traditional protein-level metaproteomics analysis by investigating the effects of commonly used sugar substitute sweeteners on *ex vivo* human gut microbiome. Traditional protein-level metaproteomics analysis revealed that the functional profiles of the microbiomes grouped these sweeteners into two clusters, the noncaloric artificial sweetener (NAS) cluster and the carbohydrate (CHO) cluster. Notably, prebiotics fructooligosaccharide (FOS) and kestose (KES) clustered with the CHO cluster. Protein-level analysis further revealed that the sugar substitute sweeteners in the CHO cluster could modulate the metabolism of Clostridia.

Next, to facilitate peptide-level human gut metaproteomics analysis, we constructed a core peptide database by extracting and reanalyzing data from published human gut

metaproteomics studies. Raw files from fifteen publicly available human gut metaproteomics projects were reanalyzed and the identified peptide-spectrum matches (PSMs) were used to construct a core peptide database, MetaPep. The constructed MetaPep database enabled rapid and accurate identification of peptides for human gut metaproteomics. Additionally, annotation and reanalysis of peptides from the MetaPep database revealed the current landscape of human gut metaproteomics research.

Finally, to advance peptide-centric analysis in metaproteomics studies, we explored peptide abundance correlations to enhance taxonomic and functional characterization of the human gut microbiome. Specifically, we analyzed peptide abundance correlations in a metaproteomics dataset derived from six *in vitro* cultured human gut microbiomes exposed to over one hundred drug treatments. Peptides originating from the same protein or taxon exhibited correlated abundance changes. Peptides from the same taxon were clustered within a peptide correlation map visualized with t-SNE. Leveraging these correlations, we enabled genome-level taxonomic assignments for a larger number of peptides. In addition, by focusing on single-species subsets, we calculated taxon-based normalized peptide abundances (TNPA) and constructed peptide abundance correlation networks, which revealed functional linkages among peptides and provided novel insights into the roles of previously uncharacterized microbial proteins.

Altogether, this thesis demonstrated the utility of protein-level metaproteomics analysis by studying the functional effects of sugar-substitute sweeteners on the human gut microbiome. Furthermore, by constructing a core peptide database, analyzing peptides in this database, and studying peptide abundance correlations in a large metaproteomics dataset, this thesis provided practical tools for peptide-level analysis, which in turn advanced taxonomic and functional analysis in human gut metaproteomics.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr. Daniel Figeys, for giving me the invaluable opportunity to pursue my PhD and for providing continuous guidance and support throughout this journey. It is meaningful to see him begin a new chapter of his academic career in the UK as I conclude my own PhD journey, and I sincerely wish him continued success and fulfillment.

I sincerely thank all members of the NorthOmics Lab. I am especially grateful to Dr. Zhibin Ning for his detailed guidance and generous support in all aspects of my research. I would also like to thank Dr. Kai Cheng for providing essential support in metaproteomics data analysis and for his insightful discussions throughout my studies. I deeply appreciate the valuable and often interesting and inspiring conversations I had with both Dr. Ning and Dr. Cheng, during and beyond our lunch breaks. Special thanks to Dr. Janice Mayne for her excellent coordination and administrative support in the lab. I also wish to acknowledge our lab technicians, Alex Dewar and Sara Qin, whose work generated the high-quality experimental data that formed the foundation of many of our bioinformatics projects. In addition, I thank our former lab members, including Dr. Leyuan Li, Dr. Xu Zhang, Dr. Galen Guo, Dr. Joesselle Serrana, and Dr. Bo Xu, for their contributions to my research and their warm companionship during my time in the lab.

I would also like to express my heartfelt thanks to my fellow lab mates and close friends—Dr. Haonan Duan, Qing Wu, Ailing Zhang, and Adrian Beltran. I am truly grateful for your companionship and support, both academically and personally. The stimulating scientific discussions and shared moments of encouragement have been invaluable and will remain cherished memories. I wish each of you continued success in your careers and personal endeavors.

I extend my heartfelt thanks to my parents and family members in China. Although we were separated by great distance throughout my PhD journey, their unwavering spiritual, emotional, and financial support made this pursuit possible. I look forward to continuing my research while spending more time with them in the future.

I would like to acknowledge the members of my Thesis Advisory Committee—Dr. Alain Stintzi, Dr. Mathieu Lavallée-Adam, and Dr. Xuhua Xia—for their valuable suggestions and constructive feedback on my research work throughout my PhD.

I am also grateful to the University of Ottawa and the city of Ottawa for providing an inclusive, supportive, and inspiring environment throughout my doctoral studies. I would also like to thank all the friends I met here—your presence enriched my experience and made this chapter of my life truly memorable.

Finally, I would like to thank myself—a young man in his twenties who devoted much of his energy and time to scientific research and training. I hope this experience will serve as a solid foundation and a lasting source of strength, confidence, and inspiration in the next chapter of my life.

CONTRIBUTIONS OF COLLABORATORS

- **Dr. Daniel Figeys (University of Ottawa)**

Dr. Daniel Figeys provided overall supervision of my PhD research, contributed to funding acquisition and the experimental design of all projects, and offered continuous guidance and critical feedback throughout the preparation of this thesis.

- **Dr. Zhibin Ning (University of Ottawa)**

Dr. Zhibin Ning provided detailed guidance throughout my research, contributed to the study design, offered expert advice on bioinformatics analyses, and provided critical feedback during the development and writing of this thesis.

- **Dr. Kai Cheng (University of Ottawa)**

Dr. Kai Cheng developed tools and workflows for metaproteomics database searches and provided technical support and advice for bioinformatics analyses.

- **Dr. Janice Mayne (University of Ottawa)**

Dr. Janice Mayne assisted with lab organization and provided administrative support throughout the project.

- **Dr. Leyuan Li (University of Ottawa)**

Dr. Leyuan Li conducted wet lab experiments and performed the initial peptide identifications for the dataset used in the peptide abundance correlation study in Chapter 4.

- **Wenju Wang (University of Ottawa)**

M.Sc. student Wenju Wang conducted wet lab experiments for the sweetener project described in Chapter 2.

- **Dr. Andrew Doxey (University of Waterloo)**

Dr. Andrew Doxey contributed to the study design and interpretation of results in the peptide abundance correlation project described in Chapter 4.

- **Dr. Haonan Duan (University of Ottawa)**

Dr. Haonan Duan provided helpful discussions and feedback on my research projects.

- **Qing Wu (University of Ottawa)**

Ph.D. student Qing Wu provided technical support and advice for bioinformatics analyses, as well as helpful discussions on my research projects.

CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
CONTRIBUTIONS OF COLLABORATORS	vi
LIST OF FIGURES	xi
LIST OF TABLES	xiii
LIST OF ABBREVIATIONS	xiv
1 GENERAL INTRODUCTION	1
1.1 Human Gut Microbiome	2
1.1.1 General overview of the human gut microbiome	2
1.1.2 Close association between the human gut microbiome and human health	9
1.1.3 High-throughput research methods for the human gut microbiome	11
1.2 Metaproteomics and Its Application to Study the Human Gut Microbiome	12
1.2.1 Applying metaproteomics to study the human gut microbiome	12
1.2.2 The advantage of metaproteomics compared to other omics	15
1.2.3 Discoveries achieved through metaproteomics	16
1.3 Challenges Remaining in Metaproteomics Analysis	17
1.3.1 Limited proteome coverage in metaproteomics	17
1.3.2 Peptide-centric analysis to reduce information loss in the protein inference	18
1.4 Applying Metaproteomics to Study Protein Functional Interactions	20
1.5 Rationale and Hypothesis	21
2 PROTEIN-LEVEL METAPROTEOMICS ANALYSIS TO STUDY FUNCTIONAL EFFECTS OF SWEETENER ON HUMAN GUT MICROBIOME	22
2.1 Abstract	23
2.2 Introduction	25
2.3 Materials and Methods	26
2.3.1 Sugar substitute sweeteners and determination of concentrations	26
2.3.2 Human stool sample collection and culturing	28

2.3.3	Metaproteomic sample preparation	29
2.3.4	LC-MS/MS analysis	30
2.3.5	Protein identification, quantification, and profiles	30
2.3.6	Microbial taxonomic analysis	32
2.3.7	Microbial function analysis	32
2.3.8	Data availability	33
2.4	Results	34
2.4.1	Functional profiles of individual microbiomes were altered by sugar substitute sweeteners	34
2.4.2	Sugar substitute sweeteners induce genus-level protein biomass in microbiomes	37
2.4.3	Functional metaproteomic analysis segregates the sugar substitute sweeteners into two groups	39
2.4.4	CHO cluster sweeteners modulated the metabolism of Clostridia . .	41
2.5	Discussion	44
2.6	Conclusion	48
2.7	Supplementary Materials	49
2.7.1	Supplementary Figures	49
2.7.2	Supplementary Tables	54
3	A CORE PEPTIDE DATABASE FOR FASTER HUMAN GUT METAPROTEOMICS DATABASE SEARCHES	59
3.1	Abstract	60
3.2	Introduction	61
3.3	Materials and Methods	63
3.3.1	Raw file collection and peptide identification	63
3.3.2	Peptide sequence database and peptide spectral library construction	64
3.3.3	Evaluation of the search speed and peptide identification	65
3.3.4	Evaluation of the search specificity and sensitivity	66
3.3.5	Taxonomic profiling	67
3.3.6	Preliminary evaluation of applying MetaPep for DIA raw files . . .	68
3.4	Results	69
3.4.1	MetaPep from identified spectra and peptide sequences	70
3.4.2	Searching MetaPep achieved faster search speeds with comparable identifications for human gut metaproteomics	70
3.4.3	Searching MetaPep achieved peptide identification with high speci- ficity and sensitivity	73
3.4.4	MetaPep preserved the ability of taxonomic profiling	76
3.4.5	MetaPep was applicable for DIA metaproteomics	81
3.5	Discussion	81
3.6	Conclusion	85
3.7	Supplementary Materials	88
3.7.1	Supplementary Figures	88
3.7.2	Supplementary Tables	109
3.7.3	Supplementary Methods	111

4	PEPTIDE ABUNDANCE CORRELATIONS IN METAPROTEOMICS ENHANCE TAXONOMIC AND FUNCTIONAL ANALYSIS OF THE HUMAN GUT MICROBIOME	113
4.1	Abstract	114
4.2	Introduction	115
4.3	Materials and Methods	117
4.3.1	Metaproteomics dataset	117
4.3.2	Peptide abundance correlation calculation	118
4.3.3	Peptide annotations	119
4.3.4	Visualize peptide abundance correlations with global peptide correlation maps	120
4.3.5	Applying a machine learning model to predict peptide genome source	120
4.3.6	Single-species peptide abundance correlations	121
4.3.7	Data availability	123
4.4	Results	124
4.4.1	Peptide abundance correlations in the metaproteomics dataset . . .	124
4.4.2	A global peptide abundance correlation map	128
4.4.3	Peptide abundance correlations provide additional information on assigning peptide taxonomic source	130
4.4.4	Functionally related peptides were connected in single-species peptide abundance correlation networks	133
4.4.5	Applying single-species peptide correlation networks for predicting unknown protein functions	137
4.5	Discussion	140
4.6	Conclusion	143
4.7	Supplementary Materials	145
4.7.1	Supplementary Figures	145
4.7.2	Supplementary Tables	160
5	GENERAL DISCUSSION	161
	REFERENCES	167
	CURRICULUM VITAE	182
	PERMISSIONS AND COPYRIGHTS	185

LIST OF FIGURES

1.1	Conceptual composition of the microbiota and microbiome.	3
1.2	Environmental and microbial gradients along the gut.	4
1.3	Microbial taxa variation and functional stability in healthy individuals. . .	8
1.4	General workflow employed in metaproteomics experiments.	14
2.1	Sweeteners induce metaproteomic changes in the individual microbiomes. .	35
2.2	Sweeteners induced gut microbiome genus-level protein abundance changes.	39
2.3	Sweeteners induced functional changes in the gut microbiome.	40
2.4	Taxonomic and functional profiles of discriminative proteins for the CHO cluster and NAS cluster revealed by PLS-DA.	43
S2.1	Procedures used to determine sweetener concentrations.	49
S2.2	Individual PCA plots of other sweeteners.	50
S2.3	Total protein amount obtained compared to PBS control.	51
S2.4	Responses of other COG categories.	52
S2.5	Structure of sweeteners.	53
3.1	Workflow of the construction and evaluation of MetaPep.	69
3.2	The peptide identification by searching human gut metaproteomics raw files against MetaPep.	72
3.3	Evaluation of the specificity and the search sensitivity of searching MetaPep.	75
3.4	Application of genome-distinct peptides in MetaPep for taxonomic profiling.	78
S3.1	ID rate and peptide identification number across each project.	88
S3.2	Cumulative number of identified spectra and peptides.	89
S3.3	Characteristics of the collected PSMs in the peptide library.	90
S3.4	Peptide identification and overlaps in PXD033624.	91
S3.5	Peptide identification and overlaps in PXD017467.	92
S3.6	Peptide identification and overlaps in PXD032997.	93
S3.7	Venn diagram of identified peptides in PXD033624.	94
S3.8	Venn diagram of identified peptides in PXD017467.	95
S3.9	Venn diagram of identified peptides in PXD032997.	96
S3.10	Distribution of MaxQuant scores from searching MetaPep.	97
S3.11	PCA of peptide quantification from MetaLab and MetaPep.	98
S3.12	Intensity of uniquely quantified peptides in PXD033624.	99
S3.13	Intensity of uniquely quantified peptides in PXD017467.	100
S3.14	Peptide quantification correlation in PXD033624 dataset.	101
S3.15	Peptide quantification correlation in PXD017467 dataset.	102

S3.16	The number of MetaPep peptides assigned to different taxonomy levels. . .	103
S3.17	Distribution of UHGP in-silico digested peptides by number of shared genomes.	104
S3.18	Taxonomic profiling of a SIHUMIx raw file.	105
S3.19	Taxonomic profiling of a human gut metaproteomics raw file.	106
S3.20	Peptide identification in DIA dataset PXD008738.	107
S3.21	The number of UHGP representative bacterial genomes identified from each project.	108
4.1	Experimental design and research framework.	124
4.2	Peptide identification and abundance correlation across individuals. . . .	126
4.3	Peptide abundance correlations in the metaproteomics dataset of V52. . . .	127
4.4	The global peptide correlation map of V52 generated using t-SNE.	129
4.5	Applying peptide abundance correlations for peptide taxonomic assignments in Bacteroidaceae.	132
4.6	Functionally related peptides were connected in V52 single-species peptide abundance correlation networks.	134
4.7	Predicting functions of proteins of unknown functions.	138
S4.1	Genus-level peptide abundance profile across individuals.	145
S4.2	Abundance profiles of peptides from different proteins.	146
S4.3	Global peptide abundance correlation network of V52.	146
S4.4	Peptide correlation maps colored at different taxa levels.	147
S4.5	Peptide correlation maps colored by functional categories.	148
S4.6	Zoomed-in peptide correlation maps with functional annotations.	149
S4.7	Peptide taxonomic prediction in Eggerthellaceae.	150
S4.8	Peptide taxonomic prediction in Lachnospiraceae.	151
S4.9	Peptide taxonomic prediction in Burkholderiaceae.	152
S4.10	Taxonomic prediction for non-Bacteroidaceae peptides.	153
S4.11	Abundance patterns of OPA and TNPA across treatments.	154
S4.12	SCC distribution between OPA and TNPA across genomes.	155
S4.13	Distribution of SCC between OPA and TNPA across genomes.	156
S4.14	Single-species peptide abundance correlation networks based on OPA. . . .	157
S4.15	Single-species peptide abundance correlation networks based on TNPA. . .	158
S4.16	Peptide correlation maps in dataset PXD012724.	159

LIST OF TABLES

S2.1 Sweetener information table.	54
S2.2 Functional annotation of identified protein groups.	58
S2.3 Change of protein group intensities under different treatments.	58
S2.4 Functional annotation of enriched discriminative proteins from Clostridia. .	58
S3.1 Details of collected metaproteomics studies.	109
S3.2 Number of identified peptides and ID rates for collected studies.	111
S3.3 Taxonomical annotation of MetaPep peptides.	111
S4.1 List of metaproteomics raw files analyzed.	160

LIST OF ABBREVIATIONS

ADI	Acceptable Daily Intake
CAZymes	Carbohydrate-Active Enzymes
CHO	Carbohydrates
COG	Clusters of Orthologous Groups
CRediT	Contributor Roles Taxonomy
DIA	Data-Independent Acquisition
DMSO	Dimethyl Sulfoxide
FC	Fold Change
FDR	False Discovery Rate
FOS	Fructooligosaccharide
FTMS	Fourier Transform Mass Spectrometry
GH	Glycoside Hydrolases
GI	Gastrointestinal
GLU	Glucose
HCD	Higher Energy Collisional Dissociation
IBD	Inflammatory Bowel Disease
IBS	Irritable Bowel Syndrome
IGC	Integrated Gene Catalog
KES	Kestose
LC	Liquid Chromatography
LCA	Lowest Common Ancestor
LFQ	Label-Free Quantification
MAG	Metagenome-Assembled Genomes

MS	Mass Spectrometry
MW	Molecular Weight
NAS	Noncaloric Artificial Sweetener
OPA	Original Peptide Abundance
PASEF	Parallel Accumulation-Serial Fragmentation
PBS	Phosphate Buffered Saline
PCA	Principal Component Analysis
PLS-DA	Partial Least Squares Discriminant Analysis
PRIDE	Proteomics Identifications Database
PSMs	Peptide-Spectrum Matches
PUFs	Proteins of Unknown Functions
RapidAIM	Rapid Assay of Individual Microbiome
SCC	Spearman Correlation Coefficient
SDS	Sequence Database Search
SIHUMIx	Simplified Human Intestinal Microbiota
SLS	Spectral Library Search
TNPA	Taxon-based Normalized Peptide Abundance
t-SNE	t-distributed Stochastic Neighbor Embedding
UHGG	Unified Human Gastrointestinal Genome
UHGP	Unified Human Gastrointestinal Protein
VIP	Variable Importance in Projection

GENERAL INTRODUCTION

1.1 HUMAN GUT MICROBIOME

1.1.1 GENERAL OVERVIEW OF THE HUMAN GUT MICROBIOME

Microbiome vs. Microbiota. The human gut microbiota—or more broadly, the human gut microbiome—has become one of the most extensively studied topics in biomedical research over the past two decades (Cani, 2018). Although the terms “microbiota” and “microbiome” are often used interchangeably, there are important distinctions between them (Berg et al., 2020). The term “microbiota” dates back to the early 1900s, originating from the combination of “micro” and “biota” (Hou et al., 2022). However, its modern usage—referring to the assemblage of living microorganisms including bacteria, archaea, fungi, and viruses in a specific environment—was first defined by Lederberg and McCray in the early 2000s (Lederberg & McCray, 2001; Marchesi & Ravel, 2015). In contrast, the term “microbiome”, derived from “micro” and “biome”, was first defined in 1988 by Whipps et al., as a “characteristic microbial community” in a “reasonably well-defined habitat which has distinct physio-chemical properties” serving as their “theatre of activity” (Berg et al., 2020; Whipps et al., 1988).

More recent definitions have attempted to clarify the concept of the microbiome and its scope. For instance, the microbiome has been described as the sum of the microbes and their genomic elements in a particular environment (Ho & Bunyavanich, 2018), or as encompassing the genes and genomes of the microbiota, as well as the products of the microbiota and the host environment (Whiteside et al., 2015). Importantly, although these definitions vary slightly, they commonly emphasize the host and the environment as an integral ecological component of the microbiome, rather than an independent entity. In contrast, the term microbiota refers only to the collection of living members forming the microbiome. The relationship between the microbiota and the microbiome is illustrated in Figure 1.1.

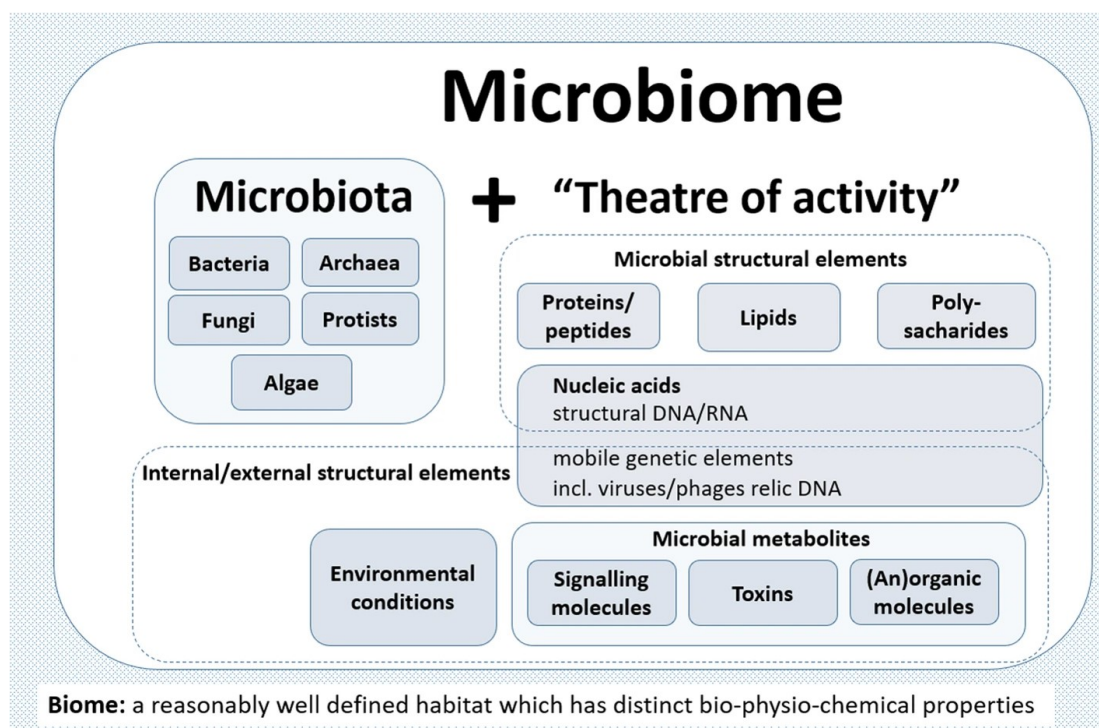


Figure 1.1: A schematic highlighting the composition of the term microbiome containing both the microbiota (community of microorganisms) and their “theatre of activity” (structural elements, metabolites/signal molecules, and the surrounding environmental conditions). Figure adapted from the reference (Berg et al., 2020) under the terms of the CC BY 4.0 license.

Human gut/Gastrointestinal (GI) tract. The term “gut” roughly refers to the human gastrointestinal (GI) tract, or more specifically, to the small and large intestines. In the context of human gut microbiome studies, however, “gut” often specifically denotes the large intestine or colon due to the widespread use of stool sampling, which primarily reflects the colonic microbiota (Levitan et al., 2023). The GI tract is a complex system in which each region performs locally specialized functions and maintains distinct physicochemical environments (Figure 1.2). Several critical factors—such as oxygen levels, pH, and mucus thickness—vary significantly along the length of the GI tract (Donaldson et al., 2016; Kennedy & Chang, 2020). The colon, a major component of the large intestine, begins at the cecum and extends approximately 150 cm to the rectum and anus, forming the terminal segment of the GI tract (Kennedy & Chang, 2020). It is highly anaerobic and exhibits lower acidity than both the stomach and the small intestine. The primary contents that

pass into the colon are indigestible dietary components such as complex polysaccharides and fibers, along with trace nutrients and residual bile acids that were not absorbed in the ileum (Donaldson et al., 2016). To facilitate fermentative metabolism of these fibrous substrates, the colon hosts a dense and diverse population of anaerobic bacteria (Hillman et al., 2017). Moreover, its slow motility—often exceeding 30 hours in transit—provides an extended window for microbial colonization and fermentation. (Metcalf et al., 1987).

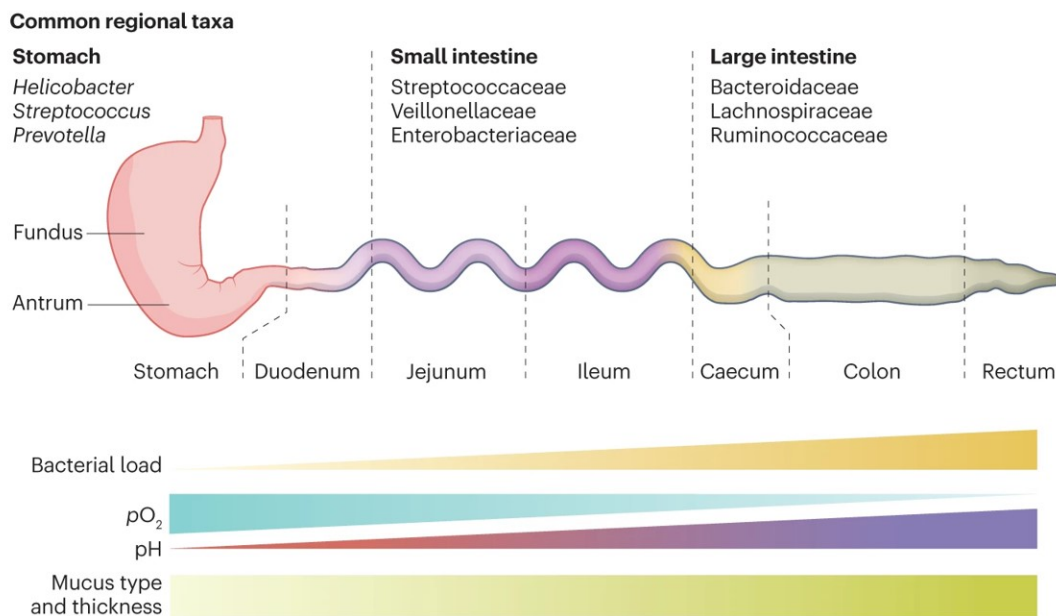


Figure 1.2: Environmental parameters and bacterial composition vary along the longitudinal axis of the gut. pO₂, partial oxygen pressure. Figure adapted from the reference (McCallum & Tropini, 2024) with permission from Springer Nature. Notably, some studies (Abuhelwa et al., 2016; Yamamura et al., 2023) have reported that the small intestine exhibits a higher pH than the colon, indicating it is less acidic, differing from the depiction in this figure.

Large bacterial load of the human gut microbiome. The number of bacteria in the human body is of the same order as the number of human cells, both reaching 10^{13} (Sender et al., 2016a), with the total mass of these bacteria estimated to be about 0.2 kg (Sender et al., 2016b). The human microbiota is closely associated with both health and disease and has even been considered the "last organ" of the human body (Baquero & Nombela, 2012; Hou et al., 2022). Among various body sites, the colon harbors the highest bacterial load—orders of magnitude greater than that of other locations such as the stomach (10^7)

and the small intestine (10^7 – 10^{11}) (Sender et al., 2016b).

Due to the extremely high microbial density in the colon, along with the practicality and noninvasiveness of fecal sampling—which predominantly reflects the colonic microbiota—the term “human gut microbiome” in most studies specifically refers to the microbiota residing in the colon and collected from stool samples (Levitan et al., 2023; Rode et al., 2024). Although spatial heterogeneity along the gastrointestinal tract has been demonstrated (Donaldson et al., 2016; Flynn et al., 2018; Zhang et al., 2014), and concerns have been raised about the limited representativeness of fecal samples (Levitan et al., 2023), fecal sampling remains the most commonly used method for studying the human gut microbiome due to its feasibility and accessibility.

Compositions of the human gut microbiome. The human gut microbiota serves multiple functions for the host, including vitamin production, stimulation of the immune system, fermentation of indigestible dietary components into short-chain fatty acids (SCFAs) and other metabolites (Bik, 2009; Hillman et al., 2017). It also provides protection against pathogens and maintains gut barrier integrity, contributes to the metabolism and detoxification of xenobiotics, and is involved in the synthesis of neurotransmitters (de Vos et al., 2022). These functions involve complex biological processes carried out by a highly diverse and functionally active bacterial community.

As a symbiotic system, the maternal microbial reservoir is a crucial source for the initial acquisition of microbial species and strains in the infant gut, and infants continue to acquire microbes from diverse maternal sources after birth (Ferretti et al., 2018). While an early study showed that the human gut microbiota is seeded during birth through contact with the maternal vaginal microbiota (Dominguez-Bello et al., 2010), a more recent study found that the composition of the maternal vaginal microbiome does not significantly affect the development of the infant gut microbiome in early life (Dos Santos et al., 2023), suggesting a more complex mode of acquisition for the infant microbiome. After initial acquisition and

formation, the gut microbiome is subsequently shaped by various environmental exposures, and the phylogenetic composition of the bacterial community typically evolves toward an adult-like configuration within the first three years of life (Yatsunenko et al., 2012).

Healthy gut microbial communities are generally characterized by both high taxonomic diversity and microbial gene richness, as well as stable functional cores (Fan & Pedersen, 2021; The Human Microbiome Project Consortium, 2012). Among these, Firmicutes and Bacteroidetes are the two dominant bacterial phyla, together accounting for more than 90% of the total microbial abundance in the human gut (The Human Microbiome Project Consortium, 2012). Other phyla, including Actinobacteria, Proteobacteria, and Verrucomicrobia are also commonly detected, but collectively account for only the remaining 10% (Lozupone et al., 2012; The Human Microbiome Project Consortium, 2012) (Figure 1.3).

At the genus level, the most prevalent and abundant taxa typically include *Prevotella*, *Bacteroides*, *Faecalibacterium*, *Ruminococcus*, *Blautia*, and *Clostridium* (Piquer-Esteban et al., 2022). However, due to differential exposure to environmental factors, the composition of human gut microbiota varies significantly across different geographical locations, dietary patterns, and age groups (Yatsunenko et al., 2012). Even among healthy individuals, the relative abundance of dominant taxa can differ greatly (The Human Microbiome Project Consortium, 2012).

In contrast, functional gene profiles tend to be more consistent across individuals (Lozupone et al., 2012). Core metabolic pathways—such as those related to ribosomal and translational machinery, ATP synthesis, and glycolysis—are commonly enriched across gut microbial communities from different individuals (Figure 1.3) (The Human Microbiome Project Consortium, 2012). Notably, several metabolic pathways are specifically enriched in the human gut microbiome, including mannan degradation, thiazole biosynthesis, and biosynthesis of L-glutamate and L-glutamine (Lloyd-Price et al., 2017), which align well with the primary microbial functions carried out in the colon.

Recently, increased research attention on the gut microbiome of individuals from low- and middle-income countries has revealed a broader spectrum of global microbial diversity. For example, a recent study of the human gut microbiome atlas of Africa identified more than 1,000 new prokaryotic metagenome-assembled genomes (MAGs) (Maghini et al., 2025), representing an increase over the Unified Human Gastrointestinal Genome (UHGG) catalog (Almeida et al., 2021), which was based on previously studied gut microbiome bacterial genomes. The composition of the human gut microbiome is expected to be further elucidated as more studies are conducted in diverse populations.

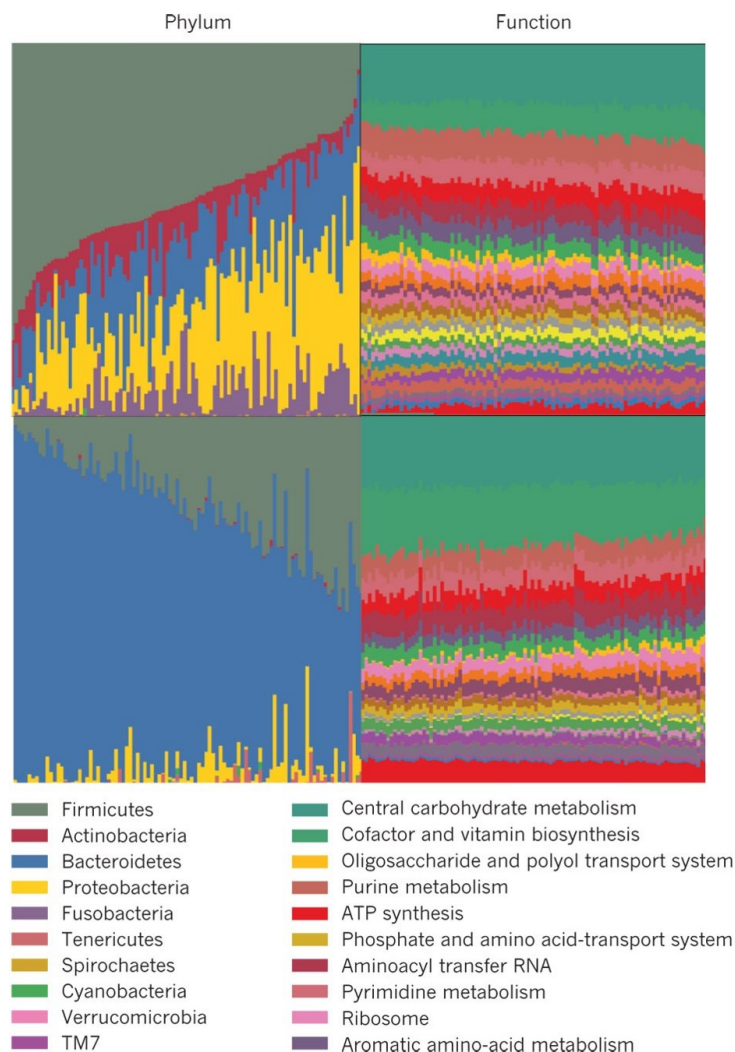


Figure 1.3: Carriage of microbial taxa varies while metabolic pathways remain stable within gut microbiomes of healthy individuals. Oral communities (top panels) and faecal communities (bottom panels) analysed using 16S rRNA (coloured by microbial phyla, left panels) show tremendous abundance diversity. The same samples analysed by shotgun metagenomics (panels on right) have remarkably similar functional profiles. Each column represents one individual, and samples were collected from 242 individuals in a published study (Lozupone et al., 2012; The Human Microbiome Project Consortium, 2012). Figure adapted from the reference (Lozupone et al., 2012) with permission from Springer Nature.

1.1.2 CLOSE ASSOCIATION BETWEEN THE HUMAN GUT MICROBIOME AND HUMAN HEALTH

Due to the complex interplay between the gut microbiota and the host, the human gut microbiome has been associated with a growing number of health issues, including both intestinal and extraintestinal disorders. Intestinal diseases such as inflammatory bowel diseases (IBD) (Lloyd-Price et al., 2019), irritable bowel syndrome (IBS) (Mars et al., 2020), and colorectal cancer (CRC) (Tilg et al., 2018) have been extensively investigated.

For example, patients with IBD exhibit altered microbial composition with reduced microbial diversity (Foppa et al., 2024). During active disease phases, there is an increase in facultative anaerobes and a decline in obligate anaerobes—indicating taxonomic shifts toward more aerotolerant species (Lloyd-Price et al., 2019). In the case of CRC, species such as *Streptococcus bovis* and *Fusobacterium nucleatum* (*F. nucleatum*) have been identified as risk factors associated with tumor formation (Kim & Lee, 2022). Specifically, in mouse models, *F. nucleatum* promoted a pro-inflammatory microenvironment that facilitated neoplastic progression in intestinal epithelial cells and recruited tumor-infiltrating immune cells (Kostic et al., 2013).

Beyond intestinal disorders, the gut microbiome has also been implicated in pancreatic disorders (Riquelme et al., 2019), metabolic, neurological, psychiatric, and immune-related diseases (Gomaa, 2020).

The gut microbial community plays a pivotal role in regulating host metabolism. Microbiota-produced metabolites interact with host receptors and modulate signaling pathways. Several metabolic disorders, including obesity (Deehan et al., 2024), diabetes (Heintz-Buschart et al., 2016; Qin et al., 2012), and cardiovascular diseases (CVD) (Brown & Hazen, 2018), have been closely associated with changes in the gut microbiota. For instance, in a mouse model of obesity, reduced ethanolamine-metabolizing capacity of the gut microbiota led to ethanolamine accumulation, which increased intestinal permeability and contributed

to inflammation and impaired glucose metabolism (Mishra et al., 2023).

The gut microbiota is also involved in both neurologic and psychiatric disorders, including autism spectrum disorder (ASD), Parkinson’s disease (PD), Alzheimer’s disease (AD), depression, and anxiety (Morais et al., 2021). These associations are mediated through the gut–microbiota–brain axis, a network of connections involving multiple biological systems that allow bidirectional communication between gut bacteria and the brain (Morais et al., 2021). For example, SCFAs produced by gut bacteria can modulate central nervous system (CNS) functions by influencing neuroplasticity, epigenetic regulation, gene expression, and immune responses (Dalile et al., 2019). Moreover, temporal changes in microbiome composition have been strongly correlated with ASD phenotypes, with characteristic microbial signatures involving genera such as *Prevotella*, *Bifidobacterium*, *Desulfovibrio*, and *Bacteroides* (Morton et al., 2023). These interactions indicate that the human gut microbiome is crucial for maintaining the homeostasis of gastrointestinal, neural, and microbial systems.

Gut microbiota and their metabolites also play essential roles in regulating both innate and adaptive immune responses (Pushalkar et al., 2018). For example, they stimulate host immune cells to produce immunoglobulin A (IgA) and antimicrobial peptides (AMPs), thereby protecting the host from pathogenic infections (Belkaid & Hand, 2014). Dysregulation of the gut microbiota is strongly associated with aberrant mucosal immune responses, intestinal inflammation, and tissue injury (Zheng et al., 2020). Moreover, microbial metabolites can translocate from the intestinal lumen to distant organs such as the liver, lungs, and brain via the circulatory system, inducing localized immune responses in peripheral tissues (Zheng et al., 2020).

Importantly, emerging evidence supports not only associations but also causal relationships between the gut microbiome and disease onset and progression. Germ-free animal models, which are raised in sterile environments and lack any microbial exposure, have been indispensable in demonstrating microbiome-mediated effects on host physiology (Cussotto

et al., 2019). For instance, the essential role of gut microbes in the development of colorectal cancer (Yang et al., 2022) and Alzheimer’s disease (Chen et al., 2022) has been validated using germ-free mice.

Beyond fundamental research, microbiome-based therapeutic strategies are increasingly being explored. Fecal microbiota transplantation (FMT), which involves transferring stool-derived microbiota from healthy donors to patients, has been approved for treating recurrent *Clostridioides difficile* infections (rCDI) (Mullard, 2023). In addition, live biotherapeutic products derived from gut microbes are currently under clinical evaluation (phase 2 trials) for conditions such as IBS, IBD, depression, and type 2 diabetes (T2D) (Pribyl et al., 2025). These emerging therapies represent a novel class of interventions harnessing the power of the human gut microbiome.

1.1.3 HIGH-THROUGHPUT RESEARCH METHODS FOR THE HUMAN GUT MICROBIOME

Recent technological advances have enabled in-depth investigation of the human gut microbiome beyond conventional methods. Traditional culture-based methods are limited by low throughput and poor culturability. To overcome these limitations, culture-independent omics approaches—including amplicon sequencing, metagenomics, metatranscriptomics, metaproteomics, and metabolomics—have become important tools for studying microbial communities in the gut (Franzosa et al., 2015). Each of these methods provides distinct layers of information, as outlined below.

Amplicon sequencing targets specific marker genes (predominantly the bacterial 16S ribosomal RNA gene), enabling taxonomic profiling of microbial communities but lacking functional insights.

Metagenomics captures millions of random genomic fragments from environmental DNA, allowing for both taxonomic and functional profiling of microbial communities.

1.2. METAPROTEOMICS AND ITS APPLICATION TO STUDY THE HUMAN GUT MICROBIOME

Metatranscriptomics sequences RNA transcripts produced by the microbiota, providing detailed insights into gene expression dynamics and complementing metagenomic analyses.

Metaproteomics directly quantifies protein abundance, offering a more accurate measure of microbial functional activity compared to gene- or transcript-based approaches. As well, metaproteomics has been used to identify and quantify post-translational modified peptides. These modifications are important for regulating the functions, localization, and interactions of proteins (Duan et al., 2023; Macek et al., 2019). In contrast, metagenomics and metatranscriptomics primarily reflect the functional potential of microbial communities, as the presence of genes or transcripts does not necessarily indicate active protein expression or function.

Metabolomics analyzes small molecules and metabolites derived from microbes and the host, offering further insights into metabolic functions and host-microbiome interactions.

Together, these high-throughput methods have provided complementary, multi-layered insights into the structure and function of the human gut microbiome, greatly advancing our understanding of its role in human health and disease.

1.2 METAPROTEOMICS AND ITS APPLICATION TO STUDY THE HUMAN GUT MICROBIOME

1.2.1 APPLYING METAPROTEOMICS TO STUDY THE HUMAN GUT MICROBIOME

The human gut microbiome has complex interactions with the host, and the taxonomic composition and functional activity of the gut microbiome have been closely associated with human health and diseases (de Vos et al., 2022). Proteins are the foundations of most biological processes, comprising about 50% of the dry mass of a cell across different species

(Milo, 2013). Their quantification is instrumental in assessing the biomass contributions of different bacterial species within a community (Kleiner et al., 2017). Consequently, a comprehensive analysis of the entire protein complement in the microbiome is critical for unraveling host-microbiome interactions.

Metaproteomics was first proposed by Wilmes and Bond in 2004, defined as the “Large-scale characterization of the entire protein complement of environmental microbiota at a given point in time” (Wilmes & Bond, 2004). Initially, metaproteomics involved separating proteins on a 2D gel and manually selecting individual protein spots for mass spectrometric analyses, which was experimentally demanding and low throughput (Wilmes & Bond, 2006). During this early stage of metaproteomics, approximately 2,000 proteins could be detected in a microbial community (Ram et al., 2005). Although still in its early stages, liquid chromatography and tandem mass spectrometry (LC-MS/MS)-based bottom-up metaproteomics is now able to detect approximately 50,000–70,000 protein groups in a single study using different techniques (Li et al., 2023; Zhao et al., 2023), showcasing the rapid technological advancement over the past 2 decades.

Currently, a typical metaproteomics workflow is illustrated in Figure 1.4 (Salvato et al., 2021). It involves collecting and preserving samples, followed by cell lysis and protein extraction. The extracted proteins are then digested, typically with trypsin, into peptides, which are subsequently separated by high-performance liquid chromatography (HPLC). The resulting peptides are analyzed by mass spectrometry (MS) to determine their mass-to-charge ratio (m/z), and by MS/MS to acquire fragmentation spectra for peptide identification and quantification. Identified peptides are then assembled into proteins or protein groups through protein inference, after which functional and taxonomic analyses are generally conducted at the protein or protein group level. A protein group refers to a set of similar, but not identical, protein sequences that share identified peptides and can not be distinguished due to the absence of protein unique peptides (Blakeley-Ruiz & Kleiner, 2022). For example, in the proteomics computational platform MaxQuant (Cox

1.2. METAPROTEOMICS AND ITS APPLICATION TO STUDY THE HUMAN GUT MICROBIOME

& Mann, 2008; Tyanova et al., 2016), whenever the set of identified peptides in one protein is equal to or completely contained within the set of identified peptides of another protein, these two proteins are joined into the same protein group.

It is noteworthy that most metaproteomics studies currently rely on data-dependent acquisition (DDA), which selects only the most abundant peptides in an MS1 scan for subsequent fragmentation and identification in MS2 (Ludwig et al., 2018). However, the emerging technique of data-independent acquisition (DIA) systematically fragments all peptides within predefined m/z windows in each MS1 scan (Ludwig et al., 2018), leading to higher reproducibility and improved proteome coverage (Aakko et al., 2020). This approach can identify 5–10 times more peptides in a single sample (Gómez-Varela et al., 2023b; Zhao et al., 2023). As a result, metaproteomics is moving toward more comprehensive profiling of both the composition and functional dynamics of microbial communities.

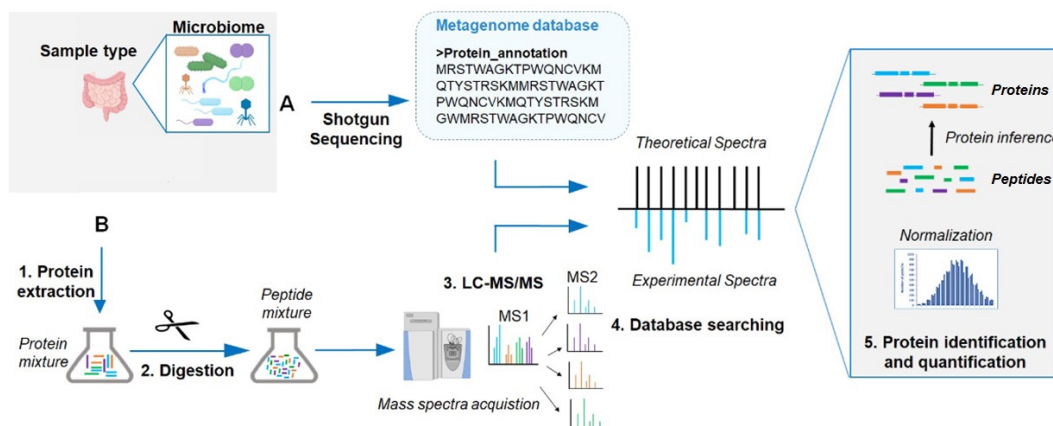


Figure 1.4: **(A)** A reference database is required for metaproteomics and can be generated using shotgun metagenomic sequencing. **(B)** The metaproteomics workflow includes protein extraction, digestion into peptides, high-resolution LC-MS/MS analysis, and subsequent database searching for protein identification and quantification based on the peptides identified and quantified by LC-MS/MS. LC, liquid chromatography; MS, mass spectrometry. Figure adapted from the reference (Salvato et al., 2021) under the terms of the CC0 1.0 license.

1.2.2 THE ADVANTAGE OF METAPROTEOMICS COMPARED TO OTHER OMICS

Metaproteomics measures the presence and abundance of proteins, thereby revealing gene expression dynamics within microbial communities (Salvato et al., 2021). Additionally, by assigning proteins to individual species or higher taxa, metaproteomics offers insights into the taxonomic composition of the microbiota (Kleiner, 2019; Salvato et al., 2021). However, metaproteomics encompasses much more than measuring gene expression and species biomass within microbial communities (Kleiner, 2019). When compared with other high-throughput omics techniques, including amplicon sequencing, metagenomics, metatranscriptomics, and metabolomics, metaproteomics has several unique advantages.

Firstly, metaproteomics is more likely to reveal the real functionality of the microbial community. This contrasts with DNA-based metagenomics and metatranscriptomics, which only suggest the potential functional capabilities of microbial communities. The reason for this disparity lies in the fact that not all DNA (genes) are transcribed into RNA, and not all RNA transcripts are subsequently translated into proteins. Secondly, metaproteomics enables the study of post-translation modifications (PTMs) in microbes. PTMs play a crucial role in modulating protein activity, structural conformation, and interactions, significantly influencing bacterial behavior within the microbiome and in interactions with the host (Duan et al., 2023). Such modifications, critical for understanding complex biological processes, remain elusive in other high-throughput omics analyses.

In addition, metaproteomics has advantages for studying host-microbe interactions, as both human proteins and microbial proteins are able to be identified and quantified. Human proteins, which constitute approximately 15% of the biomass in metaproteomic samples (Armengaud, 2023) are pivotal in mediating these interactions. Moreover, metaproteomics can also analyze isotope content, which helps determine carbon sources and provides a deeper understanding of metabolism in microbial communities (Kleiner et

al., 2018). Overall, metaproteomics distinctively enhances our comprehension of the human gut microbiome, offering insights that are not readily obtainable through other omics technologies.

1.2.3 DISCOVERIES ACHIEVED THROUGH METAPROTEOMICS

Metaproteomics has emerged as a powerful tool for unraveling the pathogenesis of various diseases and for identifying potential biomarkers. Its application ranges from unraveling the microbial contribution to oxidative stress in inflammatory bowel disease (Zhang et al., 2018b) to uncovering the interplay between gut microbiota and the development of type 1 diabetes (Gavin et al., 2018). Metaproteomics has also been applied to exploring host-microbiome interactions underlying other diseases, such as cancer (Long et al., 2020; Tanca et al., 2022), obesity (Calabrese et al., 2021; Kolmeder et al., 2015), and COVID-19 (Grega et al., 2022; He et al., 2021). A comprehensive overview of the clinical applications of metaproteomics has been provided by Wolf et al. (Wolf et al., 2023).

Recently, with the development of metaproteomics analysis platforms such as MetaLab (Cheng et al., 2023, 2017), MPA (Muth et al., 2018), and Galaxy-P (Jagtap et al., 2015), it has become easier to convert metaproteomics raw files into lists of protein groups annotated with functional and taxonomic information for downstream biological interpretation.

For example, MetaLab-MAG (Cheng et al., 2023) is a metaproteomics data analysis platform designed for genome-level characterization of microbiomes based on metagenome-assembled genomes (MAGs). It employs a two-step search strategy: first, it searches a high-abundance protein database to identify bacterial species present in the samples; then, it performs a refined search against a smaller database consisting of only these selected species. Metaproteomics analyses using MetaLab-MAG have successfully identified and quantified the components of simplified human gut microbiome samples whose composition was previously known (Cheng et al., 2023; Van den Bossche et al., 2021). Furthermore, it

has provided taxonomic profiling results that are broadly similar to, yet distinct from, those obtained from metagenomic read counts in real human gut microbiome samples (Cheng et al., 2023).

1.3 CHALLENGES REMAINING IN METAPROTEOMICS ANALYSIS

Unlike single-species proteomics, metaproteomics faces unique challenges due to the complexity of microbial communities. These challenges include database selection, statistical error estimation, protein inference, functional annotation, etc. (Schiebenhoefer et al., 2019). In the early life of metaproteomics research, it was difficult to identify enough peptides or proteins and to perform the downstream biological analysis (Muth et al., 2015; Schiebenhoefer et al., 2019). With the development of metaproteomics analyzing pipelines, such as MetaLab (Cheng et al., 2023), ProteoStorm (Beyter et al., 2018), and MPA (Muth et al., 2018), it is now able to identify around 30,000 proteins from each metaproteomics raw file with ultra-deep DDA (data-dependent acquisition) (Li et al., 2023) and even more with DIA (data-independent acquisition) (Zhao et al., 2023), making the biological interpretation of metaproteomics data more feasible.

Despite these advancements, metaproteomics still faces two major challenges that limit its biological resolution: (1) limited proteome coverage, and (2) information loss during protein inference. These will be discussed in the following sections.

1.3.1 LIMITED PROTEOME COVERAGE IN METAPROTEOMICS

In the metaproteomics community, it is widely acknowledged that there is still room for enhancing the depth of proteome coverage (Callister et al., 2018; Lohmann et al., 2020;

1.3. CHALLENGES REMAINING IN METAPROTEOMICS ANALYSIS

Van den Bossche et al., 2021). The limited proteome coverage means proteins from low-abundance taxa or with lower biomass in each taxon were left uncharacterized with the current metaproteomics workflow. This could hide some important biological discoveries. Although ultra-deep proteomics methods have shown promise in detecting nearly the entire proteome of single-species bacteria, identifying up to 75 to 77% of open-reading frames (Hou et al., 2023; Wang et al., 2021), the issue of limited proteome coverage becomes more significant when dealing with complex microbial communities composed of hundreds of species in metaproteomics. Previous research indicated that increased species diversity reduced the number of identified protein groups, and with the current state of metaproteomics technology, the estimated species and proteome coverage in a complex sample containing around 300 bacterial species is about 20% and 5%, respectively (Lohmann et al., 2020). From a more intuitive standpoint, the current stage ultra-deep metaproteomics techniques detected an average of 69,280 peptides, and 30,686 protein groups per microbiome sample (Li et al., 2023), these numbers represent a mere 2.34% of the theoretical microbiome proteome, based on an estimated 1,310,000 coding sequences (CDS) (Armengaud, 2023). These studies suggest that a significant portion of proteins and species remain undetected in metaproteomic analyses, aligning with our previous observations that bacteria with less than 0.5% biomass are difficult to detect using current metaproteomics workflows (Duan et al., 2022).

1.3.2 PEPTIDE-CENTRIC ANALYSIS TO REDUCE INFORMATION LOSS IN THE PROTEIN INFERENCE

Protein inference contributes substantially to information loss in metaproteomics and is also a cause for limited proteome coverage. It is due to the fact that within a complex microbial community, a single peptide could be shared by hundreds of different proteins, complicating the precise attribution of the peptide to its parent protein. In metaproteomics,

protein inference usually does not yield a list of proteins, but rather a list of protein groups. These protein groups can encompass proteins from various bacterial species with different functions, resulting in the loss of valuable information during protein inference (Plubell et al., 2022). In addition, different protein inference methods could result in different protein lists (Audain et al., 2017), further complicating the following analysis.

To address this problem, peptide-centric analysis has been proposed as a complementary strategy. The feasibility of peptide-centric taxonomic analysis was demonstrated with the introduction of the first tryptic peptide-based metaproteomics biodiversity analysis method, Unipept (Gurdeep Singh et al., 2019), in 2012. Since then, peptide-centric taxonomic analysis has gained widespread adoption in metaproteomics (Gavin et al., 2018; Long et al., 2020; Rechenberger et al., 2019), with subsequent developments in Unipept enabling functional analysis via GO terms and EC numbers (Gurdeep Singh et al., 2019). However, concerns have been raised about the limited sequence length of peptides and their capacity to convey meaningful functional information (Vicedomini et al., 2022; Xu et al., 2020). A proposed solution involves tailoring the protein sequence collection for in-silicon digestion, ensuring that each peptide correlates to a protein with a specific function. Customized or research-specific databases for peptide digestion have been shown to enhance functional resolution in peptide-centric metaproteomics analyses (Verschaffelt et al., 2023).

Several peptide-centric analysis tools, including Unipept (Gurdeep Singh et al., 2019; Mesuere et al., 2012), MetaGOmics (Riffle et al., 2017), and PepFunk (Simopoulos et al., 2020) have demonstrated the feasibility of the peptide-centric strategy for taxonomic and functional analysis in metaproteomics. The peptide-level analysis offers enhanced sensitivity, uncovering details that are masked at the protein level (Simopoulos et al., 2020). However, current implementations are incomplete, lacking full integration from peptide database searching to downstream functional and taxonomic analyses.

1.4 APPLYING METAPROTEOMICS TO STUDY PROTEIN FUNCTIONAL INTERACTIONS

Protein functional interactions, encompassing both direct (physical) protein–protein interactions (PPIs) and indirect (functional) interactions (Szkarczyk et al., 2017), are fundamental to a wide range of biological processes. Currently, both experimental and computational methods are employed to investigate PPIs, and proteomics has become an essential tool for studying these interactions (Kustatscher et al., 2019; Richards et al., 2021).

Experimental methods for inferring protein–protein interactions include *in vivo* techniques such as the traditional yeast two-hybrid (Y2H) assay and synthetic lethality screens, as well as *in vitro* approaches such as affinity chromatography, protein microarrays, and tandem affinity purification (Rao et al., 2014).

Computational approaches for predicting PPIs encompass several strategies. Similarity-based methods, such as interolog mapping, predict interactions by transferring experimentally confirmed PPIs between species based on homologous proteins (orthologs). Another similarity-based strategy involves domain–domain interactions (DDIs), which can predict PPIs even among proteins with low sequence similarity, given the critical role of domains in mediating protein functions. Structure-based approaches, meanwhile, predict PPIs if two proteins share similar primary, secondary, or tertiary structures (James & Muñoz-Muñoz, 2022). Other computational methods include genome context analyses, such as gene clustering and gene fusion events, as well as phylogenetic profiling, which leverages the observation that interacting protein pairs often exhibit similar patterns of conservation across multiple genomes (James & Muñoz-Muñoz, 2022).

A comprehensive summary of these methodologies is integrated into resources like the STRING database (Szkarczyk et al., 2017), which aims to provide quality-controlled protein–protein association networks for a broad range of organisms. The STRING database

integrates known and predicted functional associations from diverse sources, including: (i) systematic co-expression analysis, (ii) detection of shared selective signals across genomes, (iii) automated text mining of scientific literature, and (iv) computational transfer of interaction knowledge between organisms via gene orthology.

While these various approaches and proteomics have been widely applied to study protein functional interactions, such interactions have not yet been extensively investigated using metaproteomics datasets. In this thesis, we explore protein functional interactions within a metaproteomics dataset, leveraging taxonomically and functionally annotated peptides to gain novel insights.

1.5 RATIONALE AND HYPOTHESIS

To address the challenges described above, this thesis proposes a peptide-centric metaproteomics analysis framework designed to reduce information loss inherent in traditional protein-level human gut metaproteomics analyses. First, traditional protein-level metaproteomic analysis is assessed for its ability to reveal major functional changes of *ex vivo* human gut microbiomes in response to various sweetener treatments. This analysis also highlights the limited number of proteins identified and the smaller subset that could be taxonomically and functionally annotated. Second, a well-annotated, observation-based core peptide database is constructed to enable efficient peptide-centric searches and downstream taxonomic analyses. This database also offers an overview of existing human gut metaproteomics studies. Third, peptide abundance correlations are explored as a complementary strategy to enhance taxonomic and functional analyses. This approach enables genome-level taxonomic assignment of a larger number of peptides and facilitates the prediction of functions for previously uncharacterized proteins, thereby reducing information loss in traditional protein-level metaproteomics analysis.

PROTEIN-LEVEL METAPROTEOMICS
ANALYSIS TO STUDY FUNCTIONAL
EFFECTS OF SWEETENER ON HUMAN GUT
MICROBIOME

Preface

The following chapter consists of a research article in which I am listed as one of the two co-first authors. The article was published in *Microbiology Spectrum*.

Zhongzhi Sun, Wenju Wang, Leyuan Li, Xu Zhang, Zhibin Ning, Janice Mayne, Krystal Walker, Alain Stintzi, and Daniel Figeys. Comprehensive Assessment of Functional Effects of Commonly Used Sugar Substitute Sweeteners on *Ex Vivo* Human Gut Microbiome, *Microbiology Spectrum*, 2022.

Author contributions are described using the CRediT (Contributor Roles Taxonomy) system as listed below.

Zhongzhi Sun: Formal analysis, Investigation, Data Curation, Visualization, Writing – original draft, Writing – review & editing. **Wenju Wang:** Conceptualization, Formal analysis, Investigation, Data Curation, Visualization, Writing – original draft. **Leyuan Li:** Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Xu Zhang:** Formal analysis, Writing – original draft, Writing – review & editing. **Zhibin Ning:** Resources, Formal analysis, Writing – review & editing. **Janice Mayne:** Resources, Project administration, Writing - review & editing. **Krystal Walker:** Resources, Investigation. **Alain Stintzi:** Supervision, Writing – review & editing. **Daniel Figeys:** Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

2.1 ABSTRACT

The composition and function of the human gut microbiome are often associated with health and disease status. Sugar substitute sweeteners are widely used food additives, although many studies using animal models have linked sweetener consumption to gut microbial changes and health issues. Whether sugar substitute sweeteners directly change the human gut microbiome functionality remains largely unknown. In this study, we sys-

tematically investigated the responses of five human gut microbiomes to 21 common sugar substitute sweeteners, using an approach combining high-throughput *in vitro* microbiome culturing and metaproteomic analyses to quantify functional changes in different taxa. Hierarchical clustering based on metaproteomic responses of individual microbiomes resulted in two clusters. The noncaloric artificial sweetener (NAS) cluster was composed of NASs and two sugar alcohols with shorter carbon backbones (4 or 5 carbon atoms), and the carbohydrate (CHO) cluster was composed of the remaining sugar alcohols. The metaproteomic functional responses of the CHO cluster were clustered with those of the prebiotics fructooligosaccharides and kestose. The sugar substitute sweeteners in the CHO cluster showed the ability to modulate the metabolism of Clostridia. This study provides a comprehensive evaluation of the direct effects of commonly used sugar substitute sweeteners on the functions of the human gut microbiome using a functional metaproteomic approach, improving our understanding of the roles of sugar substitute sweeteners on microbiome-associated human health and disease issues.

Importance The human gut microbiome is closely related to human health. Sugar substitute sweeteners as commonly used food additives are increasingly consumed and have potential impacts on microbiome functionality. Although many studies have evaluated the effects of a few sweeteners on gut microbiomes using animal models, the direct effect of sugar substitute sweeteners on the human gut microbiome remains largely unknown. Our results revealed that the sweetener-induced metaproteomic responses of individual microbiomes had two major patterns, which were associated with the chemical properties of the sweeteners. This study provided a comprehensive evaluation of the effects of commonly used sugar substitute sweeteners on the human gut microbiome.

2.2 INTRODUCTION

Dietary components, which include carbohydrates (CHO), proteins, fats, minerals, food additives, and other compounds, have been shown to play major roles in shaping the composition and function of the gut microbiome and the associated health consequences (Kalantar-Zadeh et al., 2019; Zmora et al., 2019). Sugar substitute sweeteners are food additives used to increase the sweetness of food while contributing a modest amount of energy, compared to sugar. In the United States, 25% of children and 41% of adults consumed sugar substitute sweeteners, based on data collected from 2009 to 2012 (Sylvetsky et al., 2017), and the popularity of sugar substitute sweeteners has been continuously increasing (Sylvetsky & Rother, 2016). Sugar substitute sweeteners have been recommended as sugar replacements for better caloric and glycemic control while preserving sweetness (Gardner et al., 2012). Sugar substitute sweeteners can be categorized as noncaloric artificial sweeteners (NASs) of high sweetness intensity, which carry few calories, and sugar alcohols (nutritive sweeteners), which have sweetness comparable to that of sucrose but are indigestible by humans and thus contribute few calories (Liauchonak et al., 2019).

Despite the proposed health benefits of sugar substitute sweeteners, many studies have found associations between their consumption and the development of diseases and metabolic syndrome (Cohen et al., 2012; Fowler et al., 2008; Laska et al., 2012), some of which were initially intended to be prevented by the use of sugar substitute sweeteners. Sweeteners have been found by metagenomics to induce changes in the gut microbiome composition in animals and humans (Beards et al., 2010; Javier Ruiz-Ojeda et al., 2019; Suez et al., 2014). Suez et al. demonstrated saccharin (SAC)-induced glucose intolerance in healthy volunteers and showed that these effects were transferable to germfree mice through fecal transplantation (Suez et al., 2014). However, Serrano et al. found that gut microbial diversity and composition at any taxonomic level, as well as glucose intolerance, were not altered by high-dose SAC supplementation in healthy humans and mice using 16S

2.3. MATERIALS AND METHODS

rRNA sequencing (Serrano et al., 2021). To date, there has been no study that compared the effect of a comprehensive panel of approved sugar substitute sweeteners on the functionality of the *ex vivo* human gut microbiome. Most of the studies were conducted using animal models and focused on only a few sweeteners (Abou-Donia et al., 2008; Schiffman & Nagle, 2019; Suez et al., 2014). In addition, comparison across small-scale studies is challenging due to the variation of experimental approaches.

Here, we report a systematic study of the effects of sugar substitute sweeteners on individual human gut microbiomes. Briefly, 21 common sweeteners, covering all sweeteners approved by Health Canada (HC) (Health Canada, 2017), the U.S. Food and Drug Administration (FDA) (US FDA, 2018), and the European Food Safety Authority (EFSA) (EFSA, 2018) as food additives, were tested for their impact on the composition and function of five healthy adult microbiomes using the Rapid Assay of Individual Microbiome (RapidAIM) technology (Li et al., 2020), consisting of in vitro culturing and metaproteomic analysis (see Supplementary Table S2.1 for the 21 sweeteners). To the best of our knowledge, this is the first study using metaproteomic approaches to simultaneously examine the effects of 21 sugar substitute sweeteners on human gut microbiomes. Our results revealed that the sweetener-induced metaproteomic responses of individual microbiomes had two major patterns, which were associated with the chemical properties of the sweeteners.

2.3 MATERIALS AND METHODS

2.3.1 SUGAR SUBSTITUTE SWEETENERS AND DETERMINATION OF CONCENTRATIONS

The concentrations of the sugar substitute sweeteners used in the assay were determined based on their consumption levels in the general public, the acceptable daily intake (ADI), and the proportion of consumed sweeteners that could reach the colon (see Supplementary

2.3. MATERIALS AND METHODS

Figure S2.1). Twelve of the tested sweeteners had ADI data defined by FDA or EFSA. The culturing concentration that met the ADI (cADI) for each sweetener was calculated based on consumption from an average participant body weight of 70.3 kg, normalized to 200 g of colon content. Calculation of the cADI was conducted based on the following formula:

$$\begin{aligned} \text{cADI} = & \frac{\text{ADI (mg kg}^{-1} \text{ day}^{-1}) \times \text{average volunteers body weight 70.3 kg}}{(\text{weight of culture system 1 g})^{-1} \times \text{weight of colon content 200 g}} \\ & \times 1 \text{ mL}^{-1} \times \text{MW}^{-1} \times \text{proportion that reaches the colon (\%)} \end{aligned} \quad (2.1)$$

Sweeteners with cADI values much lower than 2 mg/mL were tested both at the cADI and at 2 mg/mL to facilitate the comparison of their effects on the microbiome with those of other sweeteners. For sweeteners without defined ADI, in the case of all sugar alcohols the concentration used was standardized to 2 mg/mL, which would represent consumption levels lower than those in the general public (Aguilar et al., 2009; Aguilar et al., 2015; Tennant, 2014). Since the cADI of advantame exceeded its solubility and the ADI would represent a consumption level much higher than that in the general public (Aguilar et al., 2013), advantame was tested at 2 mg/mL, about one-fifth of the cADI. In addition, aspartame (ASP), THA, and salt of aspartame-acesulfame (SAA) are known to be completely metabolized before reaching the colon (Caballero et al., 2003; Renwick, 1986), resulting in cADI values of 0. A study has shown that CHOs that can be absorbed by the small intestine can still enter the large intestine and be fermented by the colonic microbiota (Wong & Jenkins, 2007). We included ASP, THA, and SAA in this study, assuming that a portion of these sweeteners may also reach the colon.

2.3.2 HUMAN STOOL SAMPLE COLLECTION AND CULTURING

The study was approved by the Ottawa Health Science Network Research Ethics Board at the Ottawa Hospital (protocol number 20160585-01 H). Informed consent was obtained from all subjects involved in the study. Stool samples from five healthy volunteers (23 to 48 years of age; 3 males and 2 females) were included in this study. Exclusion criteria included the diagnosis of irritable bowel syndrome (IBS), inflammatory bowel disease (IBD), or diabetes mellitus; antibiotic use or gastroenteritis episode in the past 3 months; use of probiotic/prebiotic, laxative, or antidiarrheal drugs in the past month; or pregnancy. Volunteers included in this study were self-assessed as non-sweetener consumers, based on their food and nutritional supplement consumption. The stool samples were collected on site and immediately transferred into an anaerobic workstation (5% H₂, 5% CO₂, and 90% N₂, at 37°C). A 20% (wt/vol) stool slurry was made in sterile prereduced PBS (pH 7.6) containing 10% (vol/vol) glycerol and 1 g/L l-cysteine, vortex homogenized, and filtered through gauze. The filtered slurry aliquots were stored at −80°C until culturing.

The frozen fecal samples were thawed at 37°C and immediately transferred into the anaerobic workstation, vortex homogenized, and inoculated at 2% (wt/vol) into 96-deep-well plates containing prereduced, optimized microbiome culture medium (Li et al., 2019, 2020) and a sweetener (the manufacturers and concentrations used are shown in Supplementary Table S2.1), a positive control (2 mg/mL FOS or 2 mg/mL KES), or the negative control (PBS). The culture medium was composed of 2.0 g/L peptone water, 2.0 g/L yeast extract, 0.5 g/L L-cysteine hydrochloride, 2 mL/L Tween 80, 5 mg/L hemin, 10 µL/L vitamin K1, 1.0 g/L NaCl, 0.4 g/L K₂HPO₄, 0.4 g/L KH₂PO₄, 0.1 g/L MgSO₄·7H₂O, 0.1 g/L CaCl₂·2H₂O, 4.0 g/L NaHCO₃, 4.0 g/L porcine gastric mucin, 0.25 g/L sodium cholate, and 0.25 g/L sodium chenodeoxycholate. The plates were shaken at 500 rpm on shakers (MS3; IKA, Germany) in the anaerobic workstation at 37°C for 24 h. Following anaerobic culture, samples were processed to isolate bacterial pellets as described previ-

2.3. MATERIALS AND METHODS

ously (Li et al., 2020). Bacterial pellets were stored at -80°C until LC-MS/MS sample preparation.

2.3.3 METAPROTEOMIC SAMPLE PREPARATION

Proteins were extracted from bacterial pellets according to the method described by Li et al. (Li et al., 2020). Briefly, bacterial pellets were resuspended in bacterial lysis buffer containing 4% (wt/vol) sodium dodecyl sulfate (SDS), 8 M urea, 50 mM Tris-HCl (pH 8.0), cOmplete mini protease inhibitor tablets, and PhosSTOP inhibitor (MilliporeSigma). Samples were sonicated (Qsonica, USA) at 8°C , at 50% amplitude, for 10 min with a 10 s on/10 s off working cycle. Lysates were centrifuged at $16,000 \times g$ at 8°C for 10 min to remove cell debris. Supernatant protein concentrations were measured with the Bio-Rad DC protein assay reagent in triplicate and were used to calculate the total biomass of the microbiome after culturing.

Proteins from each sample were precipitated overnight at -20°C by mixing lysis supernatant with a protein precipitation buffer containing 50% (vol/vol) acetone, 50% (vol/vol) ethanol, and 0.1% (vol/vol) acetic acid, at 1:5 ratio (vol/vol). Precipitated proteins were collected by centrifugation at $16,000 \times g$ at 4°C for 25 min. Proteins were washed three times with 1 mL of -20°C acetone and pelleted by centrifugation at $16,000 \times g$ at 4°C for 25 min. Following acetone washes, proteins were dissolved in 6 M urea-50 mM ammonium bicarbonate (ABC) (pH 8.0). Protein concentrations were determined as described above. Protein aliquots (50 μg) were reduced with 10 mM dithiothreitol at 56°C for 30 min, with shaking. Protein alkylation was then performed with 20 mM iodoacetamide in the dark for 40 min at room temperature. Samples were diluted 10-fold with 50 mM ABC (pH 8.0). Trypsin (Worthington Biochemical Corp., Lakewood, NJ) was added to a trypsin/protein mass ratio of 1:50, and digestion was performed at 37°C for 19 h at 850 rpm. Trypsin digestion was stopped by adding 50 μL 10% (vol/vol) formic acid to a final pH of 2 to 3.

2.3. MATERIALS AND METHODS

Desalting was performed on in-house-made 96-channel filter tip plates packed with 5 mg 10- μ m C18 resin (Dr. Maisch GmbH, Ammerbuch, Germany). Desalted samples were freeze-dried and stored at -20°C .

2.3.4 LC-MS/MS ANALYSIS

An Eksigent NanoLC system (nano2D ultra) coupled with a Q Exactive mass spectrometer (Thermo Fisher Scientific Inc.) was used for analysis. Tryptic peptides were reconstituted in 0.1% (vol/vol) formic acid to approximately 0.25 $\mu\text{g}/\mu\text{L}$, and 1 μg of peptides was loaded. The column used for peptide separation was of 75- μm inner diameter and was 15 cm long, packed with reverse-phase C18 resin (1.9- $\mu\text{m}/120\text{-\AA}$ ReproSil-Pur C18 resin; Dr. Maisch GmbH). A 90-min gradient with acetonitrile changing from 5% to 30% (vol/vol) was used, at a flow rate of 300 nL/min. Solvent A was composed of 0.1% (vol/vol) formic acid, and solvent B was composed of 0.1% (vol/vol) formic acid and 80% (vol/vol) acetonitrile. MS analysis was performed with a Q Exactive mass spectrometer (Thermo Fisher Scientific Inc.). Full MS scans were performed from m/z 300 to m/z 1,800, and data-dependent MS/MS scans were performed for the 12 most intense ions. MS and MS/MS scans were performed with resolutions of 70,000 and 17,500, respectively. Samples were loaded in a randomized order. In this study, 197 samples were analyzed over a period of 23 days. All raw data from LC-MS/MS have been deposited with the ProteomeXchange Consortium (<http://www.proteomexchange.org>) via the PRIDE (Perez-Riverol et al., 2019) partner repository via the data set identifier PXD030458.

2.3.5 PROTEIN IDENTIFICATION, QUANTIFICATION, AND PROFILES

The MetaLab software (version 1.2.0) was used for peptide/protein identification and quantification, peptide taxonomic assignment, and protein functional annotation (Cheng et al., 2017). The searches were performed against a database based on the integrated gene cata-

2.3. MATERIALS AND METHODS

log (IGC), which included close-to-complete sets of genes for most gut microbes (Li et al., 2014). MaxQuant is used as the search engine in the MetaLab workflow for peptide/protein identification (Tyanova et al., 2016). Carbamidomethyl (C) was set as a fixed modification, and protein N-terminal acetylation (protein N-term) and oxidation (M) were set as variable modifications. The enzyme was set as trypsin, and two missed cleavages were allowed. The false-discovery rate (FDR) was set as 0.01 for both the peptide and protein levels. Because one peptide can be matched to multiple proteins, the database search engine adopts the protein inference method to determine the attribution of peptides (Cox & Mann, 2008). Basically, two proteins join in a protein group when one of the proteins has a set of peptides equal to or completely contained in the other protein. Shared peptides remain in all groups where they occur, but they are most parsimoniously associated with the group that has the highest number of identified peptides (“razor” peptides). Only unique and razor peptides are used for protein quantification. The taxonomic assignment of peptides can be found below in “Microbial taxonomic analysis.”

Analysis of changes in the gut metaproteome was based on the quantified protein groups. LFQ intensities of each protein group were first normalized by the estimated size factor calculated using the R package *DESeq2* (Love et al., 2014). Bray-Curtis distances between samples were calculated based on the normalized intensities using the R package *vegan* (Dixon, 2003). For PCA, protein groups were filtered based on the criteria that the protein group appears in at least one treatment in at least three of the five tested microbiomes (60%). The intensities were then log10 transformed, and PCA was performed using the R package *stats*. To reduce the effect of interindividual microbiome variation on data analysis, the distribution of each protein group among individual microbiomes was fitted to the same empirical distribution using an empirical Bayesian algorithm with ComBat (Johnson et al., 2007) on iMetaLab (Li et al., 2022). ComBat was also used for the intensities before log10 transform. ComBat-normalized intensities for each protein group were added to the absolute value of the smallest intensity of the protein group to avoid

2.3. MATERIALS AND METHODS

negative numbers. The intensities were then uploaded to the differential protein analyzer on iMetaLab with the default parameters. Under different treatments, protein groups with intensity fold changes of > 2 and P values of < 0.05 were considered significantly altered protein groups.

The CAZyme profile of the microbiome was acquired from the major protein of protein groups. The protein sequences of major proteins were extracted from the IGC database and then annotated with dbCAN meta server with two tools, HMMER (E value of $< 1e - 15$, with coverage of > 0.35) and DIAMOND (E value of $< 1e - 102$) (Zhang et al., 2018a).

2.3.6 MICROBIAL TAXONOMIC ANALYSIS

Taxonomic assignment of each peptide was performed based on the lowest common ancestor (LCA) algorithm. If one peptide is identified in more than one taxon, then the peptide is assigned to the LCA of those taxa. Using a previously constructed pep2tax database integrated with the MetaLab software (version 1.2.0), each identified peptide was assigned to a taxon (Cheng et al., 2017). The abundance of each taxon was calculated by summarizing the intensities of all distinctive peptides assigned to that taxon. The relative abundance of taxa in a specific taxonomic rank was calculated by normalization to the summed abundance of all taxa at that rank. For comparison of absolute taxon abundance between samples, relative abundances were multiplied by the total microbial biomass calculated using protein concentration data. Fold changes were calculated between sweetener-treated samples and PBS-treated control samples from the same microbiome.

2.3.7 MICROBIAL FUNCTION ANALYSIS

Functional annotation was carried out in the MetaLab software, and each identified protein group was assigned to a COG category. The relative abundance of each COG was calculated based on the summed LFQ intensities of all assigned protein groups. Clustering

2.3. MATERIALS AND METHODS

of sweeteners was based on the fold change of relative COG intensities, averaged across all tested sweeteners. The Euclidean distance between sweeteners was calculated, and the clustering was performed with the ward.D method, using the R package *stats*. Bootstrapping evaluation (Hennig, 2007) of the two major clusters was performed using the R package *fpc* (51), with the number of resampling runs being 100.

PLS-DA was performed to identify discriminative proteins that reveal the difference in the effects of the two clusters of sweeteners. PLS-DA was performed in MetaboAnalyst 5.0 (Pang et al., 2021), and discriminative proteins with VIP scores of > 2 were selected. The taxonomic sources of the discriminative proteins were obtained by the method described in a previous study (53). Briefly, among all of the peptides contained in the discriminative protein, the taxonomic source of the peptide with the most detailed taxonomic information was regarded as the taxonomic source of the protein. All of the peptides from a discriminative protein were acquired from the protein group table from the MetaLab output, and the taxonomic information of the peptides was obtained through the taxonomy table from MetaLab. Functional enrichment analysis of these discriminative proteins was performed using an online enrichment analysis tool (<https://shiny.imetalab.ca>), and the P value threshold was set at < 0.05 . Visualization of pathways was performed using COG accession numbers in iPath3 (Darzi et al., 2018).

2.3.8 DATA AVAILABILITY

All raw data from the LC-MS/MS analyses have been deposited with the ProteomeXchange Consortium (<http://www.proteomexchange.org>) via the PRIDE partner repository (PXD030458).

2.4 RESULTS

2.4.1 FUNCTIONAL PROFILES OF INDIVIDUAL MICROBIOMES WERE ALTERED BY SUGAR SUBSTITUTE SWEETENERS

In this study, fecal samples from five healthy volunteers were cultured individually in the presence of each sweetener in an anaerobic workstation for 24 h (Figure 2.1A). A total of 197 samples (including quality controls and technical replicates) were analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS). The average MS/MS identification rate was $33.8 \pm 6.7\%$. An average of $8,332 \pm 1,744$ peptides were identified, and $3,428 \pm 533$ protein groups were quantified from each sample. The Clusters of Orthologous Groups (COG) functional annotation and carbohydrate-active enzyme (CAZyme) annotation of identified protein groups are listed in Supplementary Table S2.2.

2.4. RESULTS

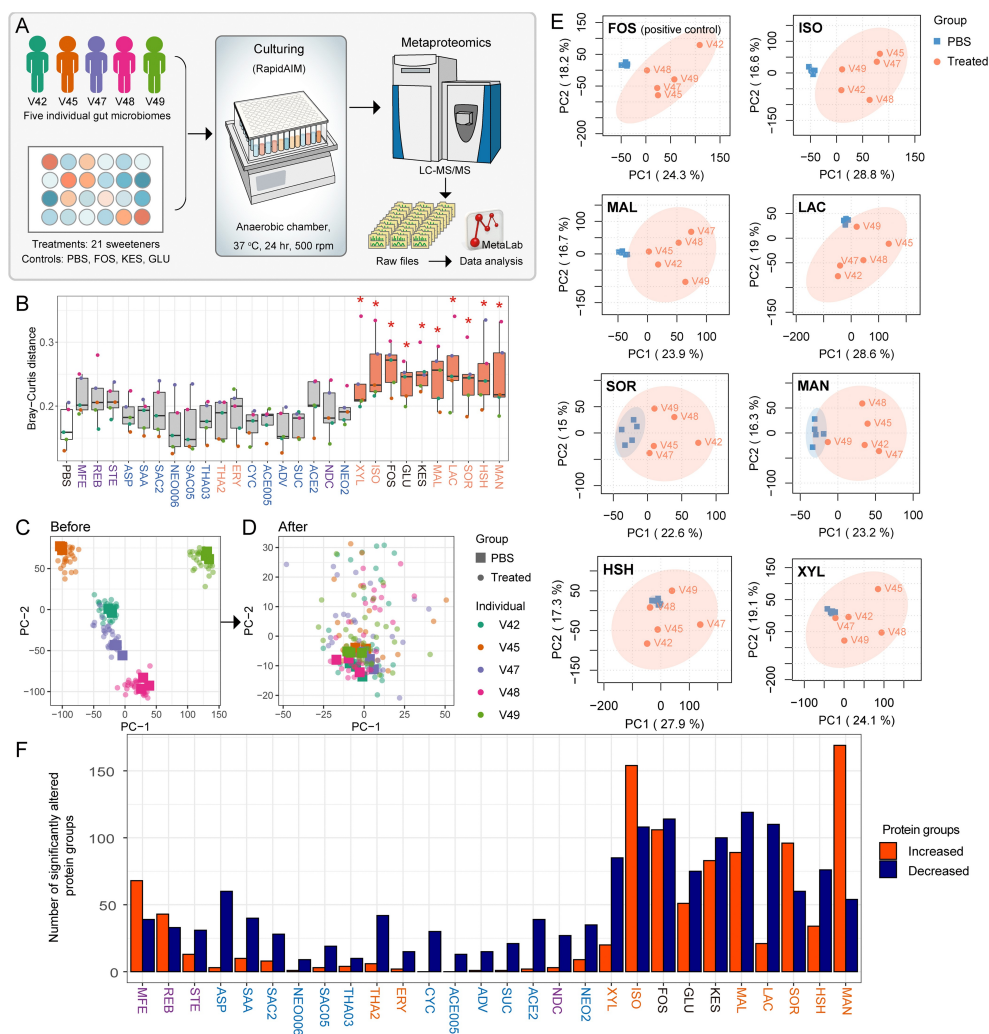


Figure 2.1: Sweeteners induce metaproteomic changes in the individual microbiomes. See detailed caption in the text below.

Figure 2.1, Detailed Caption. Twenty-one sweeteners were analyzed in the study, including four sweeteners tested at two different concentrations (2 mg/mL and cADI). Each sweetener is represented by a three-letter abbreviation. SAC2, NEO2, THA2, and ACE2 correspond to sweeteners at 2 mg/mL, and SAC05, NEO006, THA03, and ACE005 correspond to sweeteners at the cADI (see Supplementary Table S2.1 for abbreviations and specific concentrations). (A) Workflow combining in vitro culturing and metaproteomics to study the effects of common sweeteners on the gut microbiome. (B) Bray-Curtis distances of protein group LFQ intensities between sweetener-treated groups and the PBS treated

2.4. RESULTS

control for each microbiome. Boxes span the interquartile range; jitter colors indicate the microbiome number (same as Figures C and D). *, $P < 0.05$, Wilcoxon rank sum test between each group and the average distance among control sample triplicates. Gray boxes indicate non-significantly altered and orange boxes indicate significantly altered. Colors of sweetener abbreviations are as follows: orange, sugar alcohols; purple, glycoside-type NASs; blue, other NASs; black, controls. The average Bray-Curtis distances of protein group LFQ intensities between PBS triplicates from five individuals were also included (white box). (C) PCA score plot generated from protein group LFQ intensities of all samples. (D) PCA score plot after ComBat transformation to remove interindividual variances. (E) Individual PCA score plots of microbiomes treated with the positive-control FOS and a subset of sweeteners, showing separation from the PBS control (based on data after empirical Bayesian transformation). (F) Numbers of significantly altered proteins under different sweetener treatments. Protein groups with ComBat-normalized intensity fold changes of > 2 and P values of < 0.05 were considered significantly altered. See Supplementary Table S2.3 for the list of significantly altered proteins.

The comparison of Bray-Curtis distances between the sweetener-treated microbiome and nontreated microbiome metaproteomic profiles revealed seven sweeteners (xylitol [XYL], isomalt [ISO], maltitol [MAL], lactitol [LAC], sorbitol [SOR], hydrogenated starch hydrolysates [HSH], and mannitol [MAN]) that significantly altered the metaproteome across all five individual microbiomes (Figure 2.1B). Of the eight tested sugar alcohols, only erythritol (ERY) did not show significantly altered metaproteome. Principal-component analysis (PCA) of all samples based on protein group label-free quantification (LFQ) intensities showed, as expected, strong interindividual variations (Figure 2.1C); these were due to the nature of the mixture distribution of each protein group among different individuals. We used an empirical Bayesian algorithm (Johnson et al., 2007) to fit each mixture distribution to an empirical distribution to reduce the effect of individual variance on the data set. PCA of the data following transformation showed that the control samples of

2.4. RESULTS

different individuals now clustered together (Figure 2.1D). In agreement, the sweeteners mentioned above and controls showed better separation than other groups (Figure 2.1E; also see Supplementary Figure S2.2). In addition, monk fruit extract (MFE)-treated microbiomes showed a distinct cluster away from the phosphate-buffered saline (PBS) group (see Supplementary Figure S2.2). A total of 1,075 protein groups were significantly altered by at least one sweetener (see Supplementary Table S2.3). Consistent with Bray-Curtis distances and PCA, a larger number of protein groups were significantly altered by the seven sweeteners (XYL, ISO, MAL, LAC, SOR, HSH, and MAN) mentioned above (Figure 2.1F). An average of 54 ± 19 protein groups were identified as CAZyme from each sample. Most of the protein groups were glycoside hydrolases (GHs) (40 ± 16 from each sample). Only 25 of 1,075 significantly altered protein groups were annotated as CAZyme (see Supplementary Table S2.3). Two protein groups from GH97 (top protein identification numbers DOM005_GL0010478 and 272559.BF1158), a family composed of α -glucosidase and α -galactosidase, were significantly increased by HSH.

2.4.2 SUGAR SUBSTITUTE SWEETENERS INDUCE GENUS-LEVEL PROTEIN BIOMASS IN MICROBIOMES

We evaluated the effect of sugar substitute sweeteners on total microbial biomass by measuring the total proteins obtained in each sample using a detergent-compatible (DC) protein assay, as described in Materials and Methods. For most sweeteners, their effects on the individual microbiomes varied, while ISO and thaumatin at 2 mg/mL (THA2) increased the total biomass in all five microbiomes (see Supplementary Figure S2.3).

Genus-level protein abundance was then calculated from the distinctive peptide intensities of each genus measured by LC-MS/MS and the total microbial protein biomass of each sample. Most sweeteners showed significant effects on the abundance of at least one genus. Clustering based on the effects of genus-level biomass shows that the genus-level

2.4. RESULTS

protein biomass composition was affected in patterns by different classes of sugar substitute sweeteners (Figure 2.2). Glycoside-type NASs rebaudioside A (REB), neohesperidin dihydrochalcone (NDC), and MFE, as well as sugar alcohols MAL, LAC, XYL, and ISO, clustered together and formed a distinct cluster from other sweeteners and controls. This cluster showed significant increases in several genera, such as *Dialister*, *Parabacteroides*, *Ruminococcus*, *Phascolarctobacterium*, *Butyrivibrio*, *Blautia*, and *Marvinbryantia*. Sugar alcohols SOR, MAN, and HSH clustered with controls fructooligosaccharide (FOS), kestose (KES), and glucose. This cluster featured significant increases in Actinobacteria genera *Bifidobacterium* and *Collinsella* and decreases in *Dorea*, *Clostridium*, *Lachnoclostridium*, *Alistipes*, *Roseburia*, and *Flavonifractor*. Genera *Coprococcus*, *Oscillibacter*, *Anaerostipes*, and *Butyrivibrio* were increased by XYL but not by any other sugar alcohols.

2.4. RESULTS

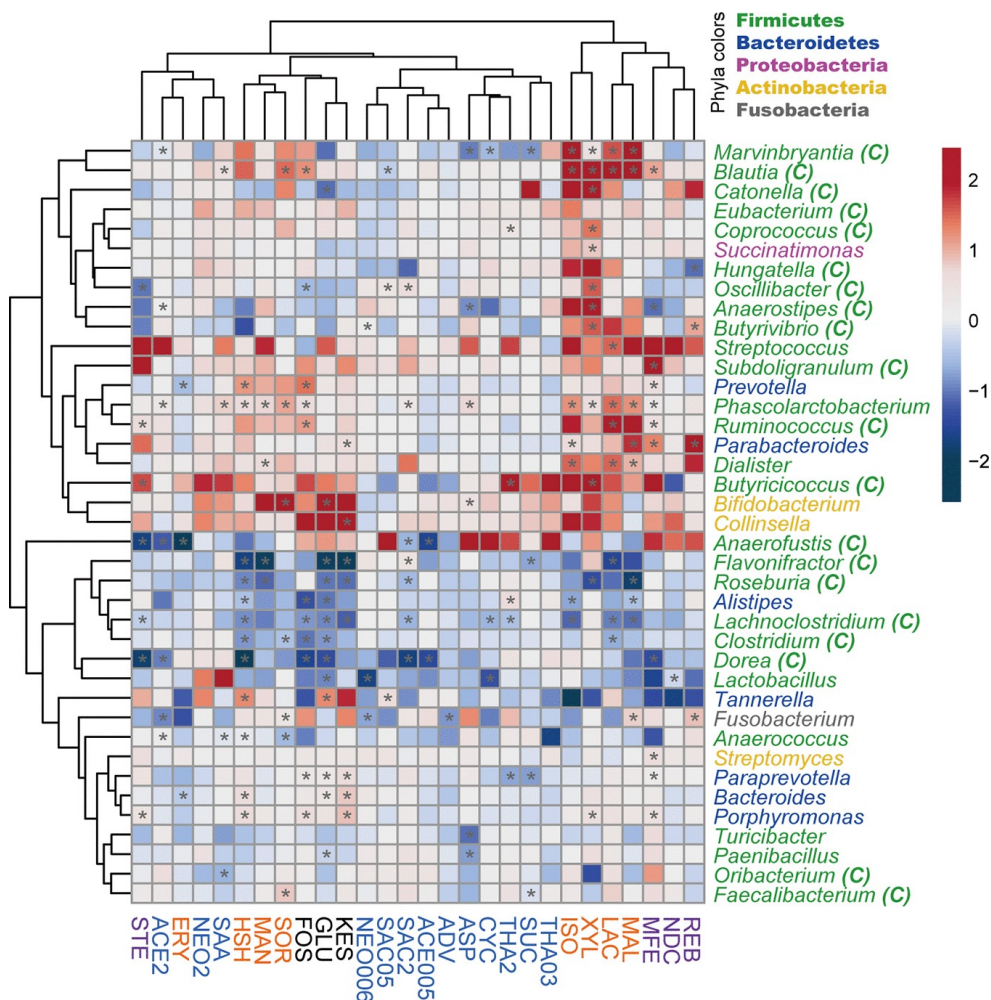


Figure 2.2: Sweeteners induced gut microbiome genus-level protein abundance changes. Sweeteners are named as in Figure 2.1. The heatmap shows log₂ fold changes in genus-level protein abundance of sweetener-treated samples versus the PBS control. For each treatment, the averaged genus-level protein abundance of all five microbiomes was used for coloring and clustering. *, $P < 0.05$, Wilcoxon rank sum test. Genera that were detected in PBS controls in at least four of the five microbiomes are shown. Genera from Clostridia are indicated with C in parentheses.

2.4.3 FUNCTIONAL METAPROTEOMIC ANALYSIS SEGREGATES THE SUGAR SUBSTITUTE SWEETENERS INTO TWO GROUPS

Of the identified protein groups, 93.5% had COG functional annotation. Sweeteners were categorized into two major clusters using COG abundances (Figure 2.3A; also see Supplementary Figure S2.4). Bootstrapping of the two major clusters gave scores of 0.983 and

2.4. RESULTS

0.956, respectively, indicating high clustering robustness (Figure 2.3A). We named the two clusters the NAS and CHO clusters according to the properties of the sweeteners. Statistical analysis at the COG category level identified 14 of 21 COG categories that were significantly altered by at least two compounds (Figure 2.3B to G; also see Supplementary Figure S2.4).

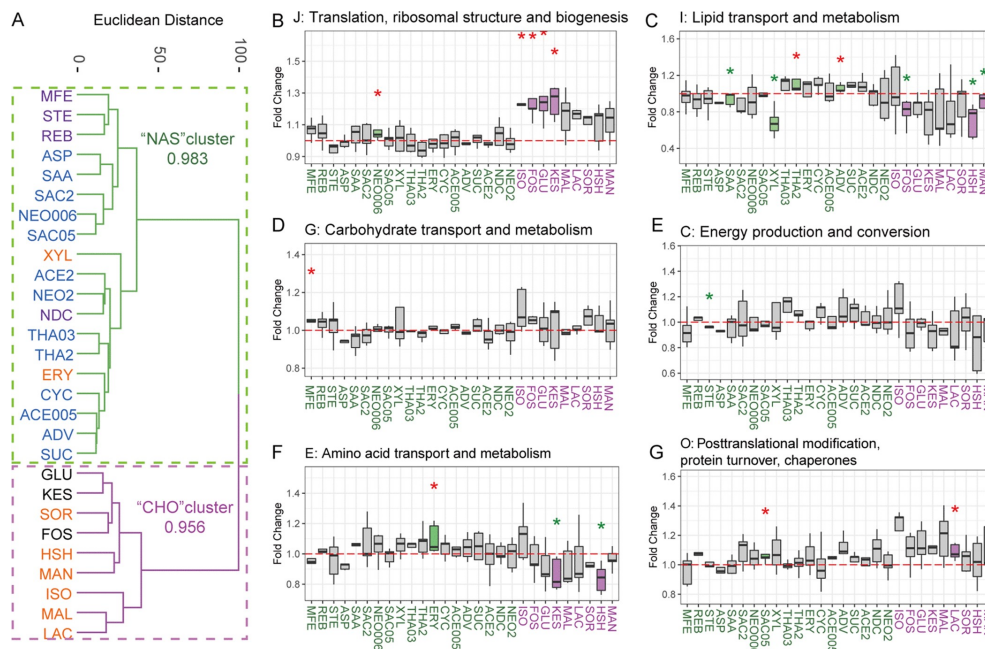


Figure 2.3: Sweeteners induced functional changes in the gut microbiome. Sweeteners are named as in Figure 2.1. (A) Clustering of sweeteners based on induced functional responses. Euclidean distances between sweeteners were calculated based on averaged log2 fold changes of COG abundances of sweetener-treated samples versus the PBS-treated control. Bootstrapping scores of the two major clusters are shown. (B to G) Fold changes between the treated group and the PBS-treated control for several COG categories. Colored boxes indicate significantly changed COG categories. Red and green asterisks indicate significant increases and decreases, respectively. *, $P < 0.05$, Wilcoxon rank sum test. Responses of all other COG categories are shown in Supplementary Figure S2.4.

In the NAS cluster, all NASs were included, plus sugar alcohols ERY and XYL, which have shorter carbon backbones (see Supplementary Figure S2.5). Interestingly, although four sugar alcohols, MAL, lactitol monohydrate, XYL, and ISO, were clustered with three of the NASs, i.e., REB, NDC, and MFE, in the taxonomic analysis in Figure 2.2, the functional responses of the microbiome to these two clusters were different. While XYL

still belonged to the NAS cluster, the three other sugar alcohols (i.e., MAL, lactitol monohydrate, and ISO) belonged to the CHO cluster. XYL showed marked effects on the metaproteome, including significantly decreased lipid transport and metabolism and cell motility and significantly increased coenzyme transport and metabolism (Figure 2.3C; also see Supplementary Figure S2.4). Extracellular structures were significantly promoted by the MFE-stevioside (STE)-REB subcluster (see Supplementary Figure S2.4). The CHO cluster included all remaining sugar alcohols and positive controls, all of which are CHOs. In the CHO cluster, sugar alcohols SOR, MAN, LAC, MAL, ISO, and HSH cannot be digested by the human body. Therefore, despite incomplete absorption of sugar alcohols in the small intestine (Livesey, 2003), they can reach the colon intact, serving as substrates for microbial fermentation to produce hydrogen gas, carbon dioxide, methane, and short-chain fatty acids (SCFAs) (Vakil, 2018). Compounds in this cluster were found to induce marked responses in the metaproteomes in a similar pattern, including an increase in translation, ribosomal structure, and biogenesis and a decrease in lipid transport and metabolism (Figure 2.3B and C). In addition, proteins with general function prediction only (COG category R) were significantly increased by a subset of the CHO cluster (Supplementary Figure S2.4). Sugar alcohols have been shown to induce gastrointestinal symptoms, including bloating, laxative effects, and abdominal pain (Lenhart & Chey, 2017). Accordingly, we showed that some sugar alcohols, such as ISO, significantly reduced cell motility, which has been reported to be associated with increased susceptibility to intestinal expulsion and larger fluctuation in absolute abundance (Wiles et al., 2020).

2.4.4 CHO CLUSTER SWEETENERS MODULATED THE METABOLISM OF CLOSTRIDIA

Partial least-squares discriminant analysis (PLS-DA) was performed to identify the most important differences in functional effects between the CHO cluster and the NAS cluster

2.4. RESULTS

(Figure 2.4A and B). A total of 214 of 3,608 protein groups had a variable importance in projection (VIP) score of > 2 in the first five components, which indicates that these proteins explain the most important differences in functional profiles between the two clusters. These proteins were referred to as discriminative proteins. The intensity profiles of these 214 discriminative proteins in response to the treatment with different sweeteners were segregated into two well-defined groups (Figure 2.4C). A total of 106 of the discriminative proteins that had greater abundance in the CHO cluster were referred to as the CHO elevated group. Sweeteners of the NAS cluster were clustered with the PBS group, indicating that proteins enriched in the NAS cluster were decreased in the CHO cluster. Therefore, 108 of the discriminative proteins that had greater abundance in the NAS cluster were referred as the CHO depleted group.

2.4. RESULTS

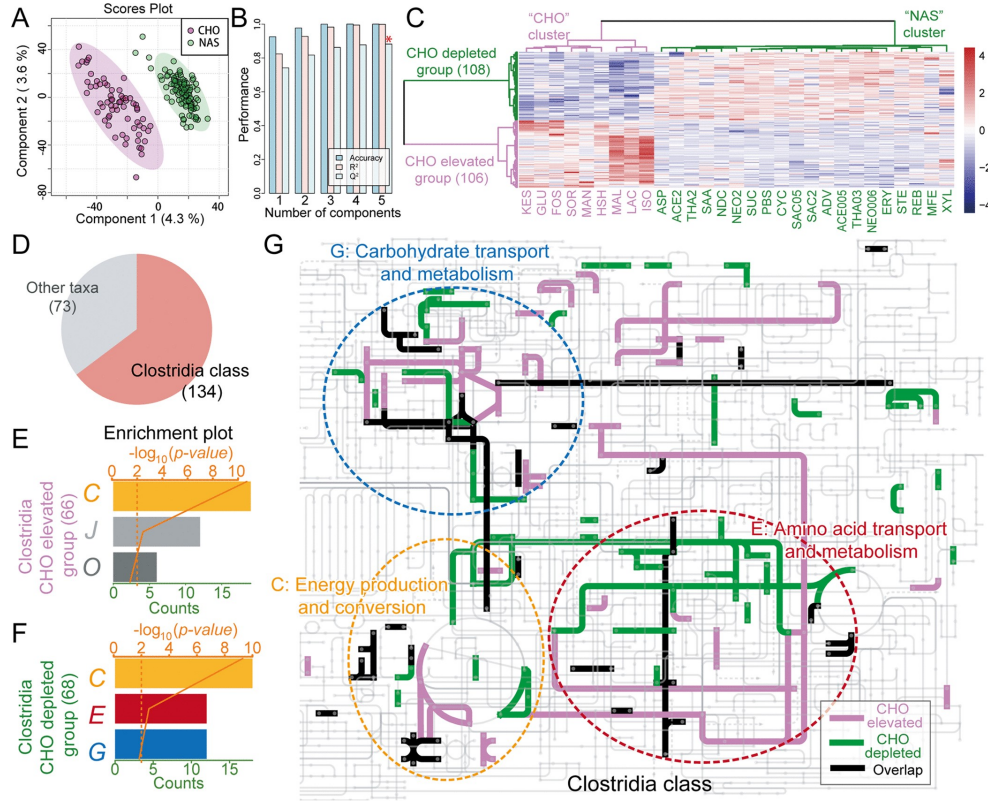


Figure 2.4: Taxonomic and functional profiles of discriminative proteins for the CHO cluster and NAS cluster revealed by PLS-DA. (A) PLS-DA score plot for differential protein profiles across the CHO cluster and NAS cluster sweeteners. (B) PLS-DA cross-validation results. (C) Heatmap of the intensity of 214 discriminative proteins under treatment with different sweeteners (sweeteners are named as in Figure 2.1). (D) Taxonomic sources of discriminative proteins. (E and F) Enriched COG categories of discriminative proteins in Clostridia in the CHO elevated group (E) and the CHO depleted group (F). (G) Pathways of discriminative proteins in Clostridia in the CHO elevated group and the CHO depleted group. Proteins related to COG categories C, G, and E are framed in the dashed circles.

Because not all of the identified peptides had taxonomic information from the MetaLab, 207 of the 214 discriminative proteins were assigned taxonomic information from peptides, as described in Materials and Methods, and 62.6% of the discriminative proteins (134 proteins) were from class Clostridia (Figure 2.4D). Forty-six of the 134 proteins could be assigned to the genus level, as they were from 11 genera, i.e., *Roseburia* (21 proteins), *Clostridium* (6 proteins), *Blautia* (4 proteins), *Eubacterium* (4 proteins), *Lachnoclostridium* (3 proteins), *Ruminococcus* (3 proteins), *Anaerostipes* (1 protein), *Coproccoccus* (1 protein), *Faecalibacterium* (1 protein), *Fusicatenibacter* (1 protein), and *Oscillibacter* (1 protein).

Functional enrichment analysis of each group of discriminative proteins from *Clostridia* was performed. The CHO elevated group was enriched in COG categories C (energy production and conversion), J (translation), and O (posttranslational modification, protein turnover, and chaperone functions), while the CHO depleted group was enriched in COG categories C, E (amino acid metabolism and transport), and G (CHO metabolism and transport) (Figure 2.4E and F). Although both the CHO increased group and the CHO depleted group from *Clostridia* were enriched in COG category C, the enriched proteins from the two groups had different functions (see Supplementary Table S2.4). Pathway analysis also showed that proteins from the two groups were from different metabolic pathways (Figure 2.4G). In particular, iron-dependent proteins that are involved in the tricarboxylic acid (TCA) cycle between succinyl-coenzyme A (CoA) and malate and its linkage to oxidative phosphorylation (COG0427, COG0479, COG1145, COG1838, COG1951, and COG2048) were highly enriched. The CHO cluster sweeteners had effects on the protein abundance of discriminative proteins similar to those of FOS and KES, which are commonly used prebiotics.

2.5 DISCUSSION

Although there have been studies of the effects of sweeteners on the human gut microbiome (Beards et al., 2010; Suez et al., 2014), a comparison of the effects of a large number of different sugar substitute sweeteners on human gut microbes has not been reported. In this study, we investigated the taxonomic and functional responses of five individuals' human gut microbiomes cultured *in vitro* to 21 common sugar substitute sweeteners. Among the 21 sweeteners, 13 are NASSs, with diverse chemical properties and high sweetness intensities, and 8 are sugar alcohols, which are CHOs with low digestibility (Lenhart & Chey, 2017). We used a previously developed *in vitro* model that was shown to maintain microbial taxon and function (Li et al., 2019, 2020). We tried to use this *in vitro* model to reveal

2.5. DISCUSSION

the real functional change in the human gut microbiome. This *in vitro* model has been shown to recapitulate microbiome changes observed *in vivo* (Li et al., 2019). The purpose of the *in vitro* model is to better understand the direct effects of compounds on individual microbiomes and, as with any assays, would need *in vivo* confirmation.

We observed that seven sugar substitute sweeteners (XYL, ISO, MAL, LAC, SOR, HSH, and MAN) significantly altered the metaproteome across the five gut microbiomes (Figure 2.1B). The remaining sweeteners had no global metaproteome effects. Although MFE showed a clear separation between control and treatment by PCA, the alteration was not significant using Bray-Curtis distance. A previous study reported that SAC alters the gut microbiota and induces glucose intolerance in a mouse model (Suez et al., 2014). While we did not observe any global effects of SAC on the metaproteome using our *in vitro* assay (Figure 2.1B), we observed changes at the individual level, indicating an individual-specific response to SAC (see Supplementary Figure S2.2). A similar individual-specific response was observed for sucralose (SUC), which has been shown by others to alter the composition of the gut microbiota in a rat model (Abou-Donia et al., 2008; Schiffman & Nagle, 2019). From the PCA plots, we observed that, compared with PBS controls, there were larger variations between replicates with the treatment with NAS cluster sweeteners. This suggests interindividual differences in the response of the NAS cluster sweeteners. It is noted that only one or two concentrations of each sweetener were tested in this study. The number of significantly altered protein groups with SAC2 was larger than that with SAC05. The same situation happened between THA2 and THA03, neotame at 2 mg/mL (NEO2) and NEO006, and acesulfame at 2 mg/mL (ACE2) and ACE005 (Figure 2.1F), which indicated that higher concentrations of sweeteners could likely have a larger impact on the human gut microbiome. The impacts of different concentrations of sweeteners need further investigation. In addition, most sugar substitute sweeteners had limited effects on the biomass of the microbiome. Only ISO and THA2 led to a biomass increase in all five microbiomes (see Supplementary Figure S2.3). Therefore, sugar substitute sweeteners that

2.5. DISCUSSION

affect the microbiome mostly do not affect the protein biomass but instead have systemic or individual effects on taxonomic balance and expressed proteins and enzymes.

A total of 1,075 protein groups were significantly altered by one or more of the sugar substitute sweeteners (see Supplementary Table S2.3). Although it is not possible to assess the possible functional impact of the changes in each protein group, some examples are noteworthy. For example, an average of 54 CAZymes, an important group of enzymes for microbiome function, were identified in each sample, a level consistent with a previous study (Huang et al., 2018). Most of the identified CAZymes were GHs, which can account for one-half of the classified CAZymes (Cantarel et al., 2009). A protein group from GH133 (top protein identification number V1.CD38-0_GL0163641) was reduced by 13 of the sugar substitute sweeteners. GH133 CAZymes are glycogen-debranching enzymes (α -1,6-glucosidases). The decrease in this enzyme class likely indicates that the sugar substitute sweeteners depressed the release of glucose from glycogen storage by some bacteria in the microbiome.

There were also specific groups of proteins that were affected by multiple sugar substitute sweeteners. For example, the protein group (top protein identification number MH0148_GL0062639) classified as outer membrane protein (OMP) OmpA from the genus *Alistipes* was decreased by four sugar substitute sweeteners. OmpA is one of the most abundant OMPs (Gribun et al., 2003) and is involved in many processes, particularly responses to stresses. It has been shown in some species to provide resistance to antibiotics (Nitzan et al., 2002). Its decrease by the four specific sugar substitute sweeteners could impact its roles in *Alistipes*.

The genus-level taxonomic changes of the gut microbiome with different sugar substitute sweeteners were evaluated. A sweetener cluster consisting of both glycoside-type NASs (REB, NDC, and MFE) and sugar alcohols (MAL, LAC, XYL, and ISO) significantly increased protein abundance from several genera in Firmicutes. A single-species-based study revealed that XYL was largely utilized by *Anaerostipes caccae* from Firmicutes to

produce butyrate and promoted the growth of the species (Sato et al., 2017). The effects of sweeteners on other genera from Firmicutes need to be studied further. In addition, controls FOS, KES, and glucose were clustered with sugar alcohols SOR, MAN, and HSH. It is noted that this clustering result was different from the functional profile clustering. While metaproteomics measure the abundances of different taxa by summarizing the overall protein intensities, functional profiles provide a deeper layer of information by comparing functional proteins in the community and are more relevant to the actual functionality and state of the microbiomes.

By analyzing the functional profiles of the cultured microbiomes, we segregated all of the sweeteners into two clusters, NAS and CHO. To further investigate the different effects of the two clusters of sweeteners on the human gut microbiome, discriminative proteins with PLS-DA VIP scores of > 2 were identified. FOS and KES clustered with the CHO cluster, which indicated that the CHO cluster might have functional effects similar to those of FOS and KES. Most of the discriminative proteins were from Clostridia. Clostridia accounts for at least 10 to 40% of the total bacteria in gut microbiota (Lopetuso et al., 2013), and members from this class have significant potential as probiotics (Guo et al., 2020).

Functional enrichment analysis of discriminative proteins from Clostridia revealed that proteins with greater abundances in the CHO cluster (CHO elevated group) and proteins with greater abundances in the NAS cluster (CHO depleted group) were both enriched in COG category C. Enriched proteins from the two groups corresponded to different functions (see Supplementary Table S2.4), with differences in specific pathways, as shown by the metabolic pathway analysis (Figure 2.4G). In particular, the enrichment of iron-dependent pathways in the TCA cycle and oxidative phosphorylation indicated increased cellular energy metabolism in Clostridia in response to the CHO cluster. Other proteins from the two groups also corresponded to different pathways, indicating that the two clusters of sweeteners had effects on different metabolic pathways of Clostridia. In addition, the CHO elevated group was enriched in COG function categories J (translation, ribo-

2.6. CONCLUSION

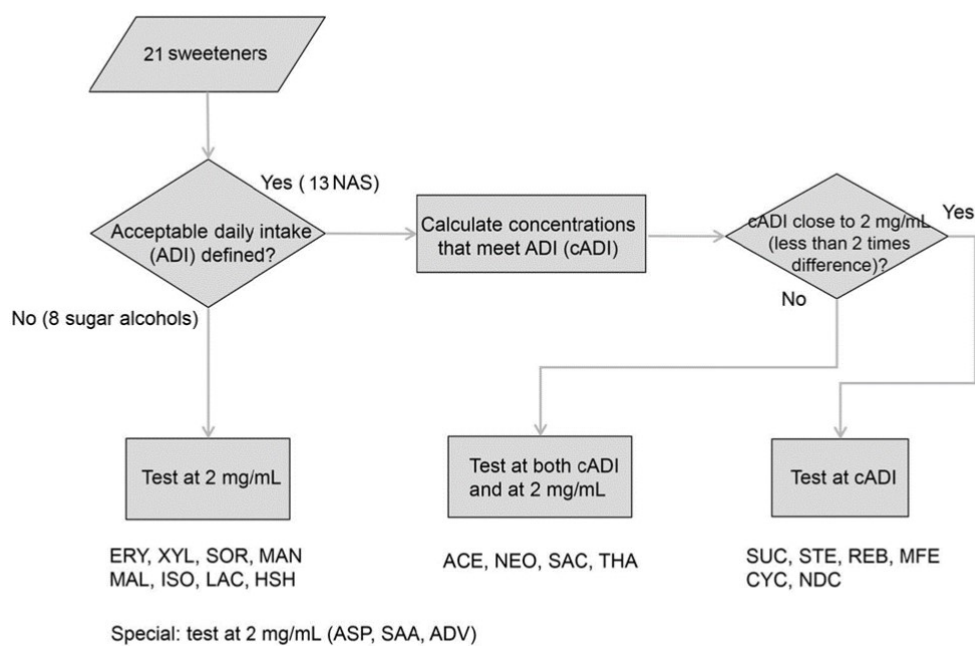
somal structure, and biogenesis) and O (posttranslational modification, protein turnover, and chaperone functions). The greater abundance of proteins from these two categories indicated that the CHO cluster sweeteners might promote the cellular growth of Clostridia. A previous study showed that the traditional prebiotic FOS and probiotic Lactobacillus can increase the proportion of Clostridia in gut microbiomes (Saulnier et al., 2008), and CHO cluster sweeteners might have similar potential as these prebiotics and probiotics. Discriminative proteins with greater abundance in the NAS cluster (CHO depleted group) were enriched in COG categories C, G, and E (amino acid transport and metabolism). The functional annotation of enriched proteins (see Supplementary Table S2.4) showed that certain numbers of these discriminative proteins in COG categories C (5 of 18 proteins) and G (4 of 12 proteins) were related to glycerol transport and alcohol metabolism. This suggested inhibition of the glycerol transport and alcohol metabolic pathways by the CHO cluster.

2.6 CONCLUSION

In conclusion, we examined the effects of 21 sweeteners on the *ex vivo* human gut microbiomes from five individuals. Seven sweeteners significantly altered the metaproteomes across the five microbiomes. Functional profiles of the microbiomes clustered all of the sweeteners into the NAS cluster and the CHO cluster. Prebiotics FOS and KES clustered with the CHO cluster. Most discriminative proteins for the two clusters were from Clostridia. Functional enrichment analysis and pathway analysis revealed that the two sweetener clusters had effects on different pathways of Clostridia. CHO cluster sweeteners had effects similar to those of the prebiotics FOS and KES. Our study revealed the functional effects of sweeteners on the human microbiome and suggested the prebiotic potential of the sugar alcohol sweeteners.

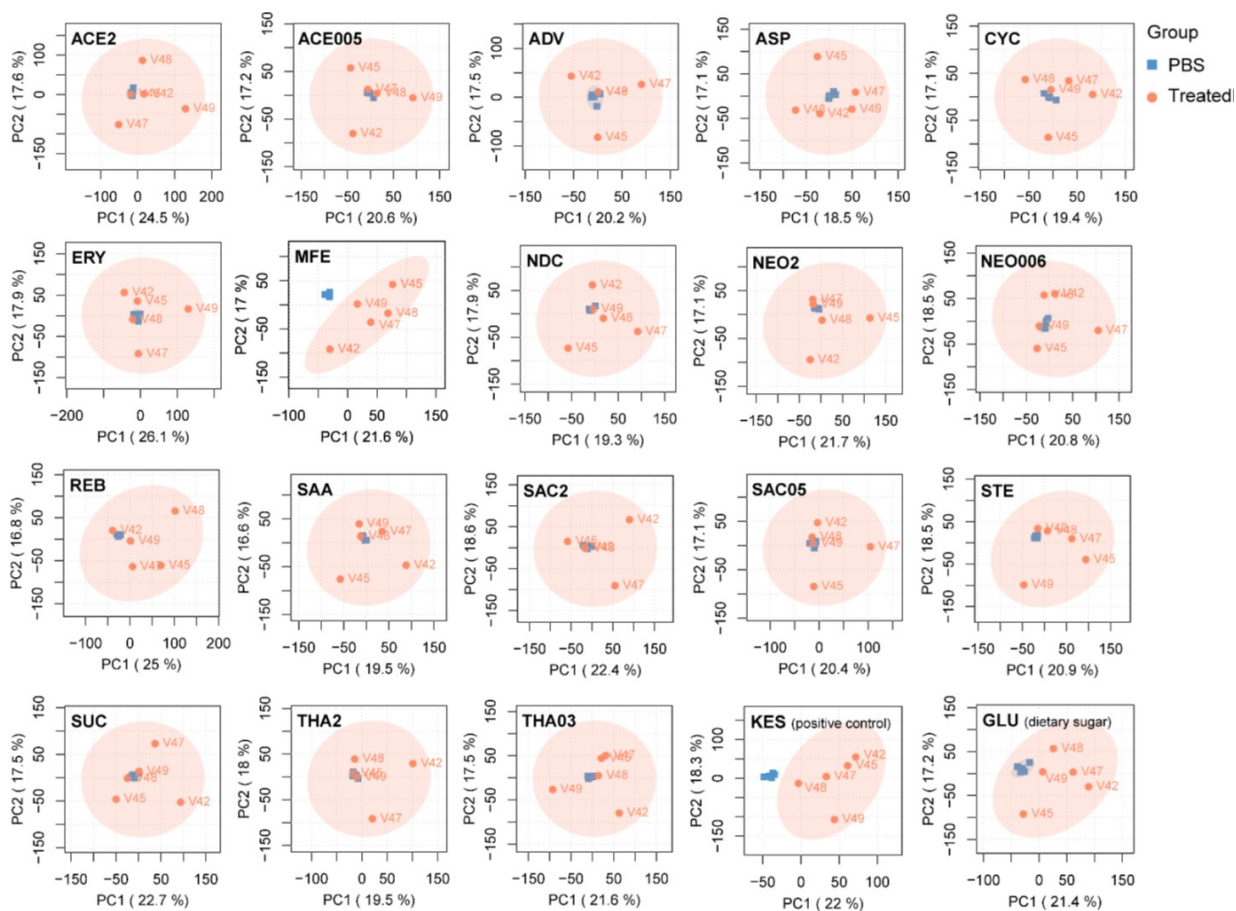
2.7 SUPPLEMENTARY MATERIALS

2.7.1 SUPPLEMENTARY FIGURES

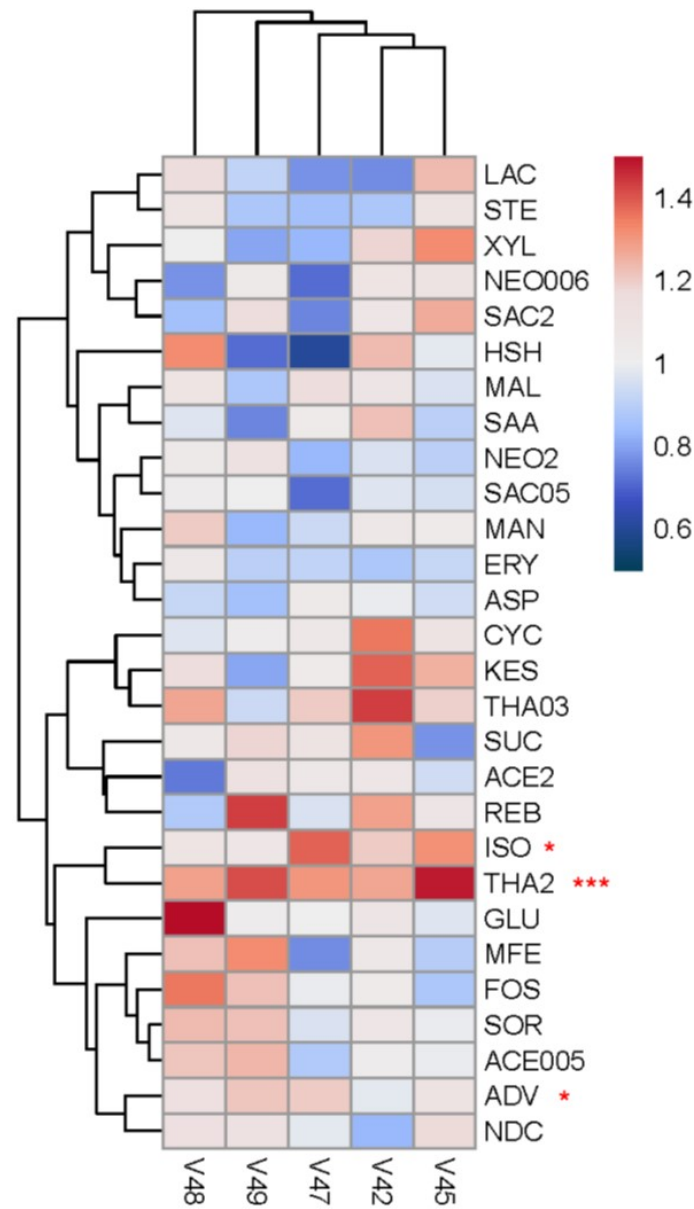


Supplementary Figure S2.1. Procedures used to determine sweetener concentrations.

2.7. SUPPLEMENTARY MATERIALS

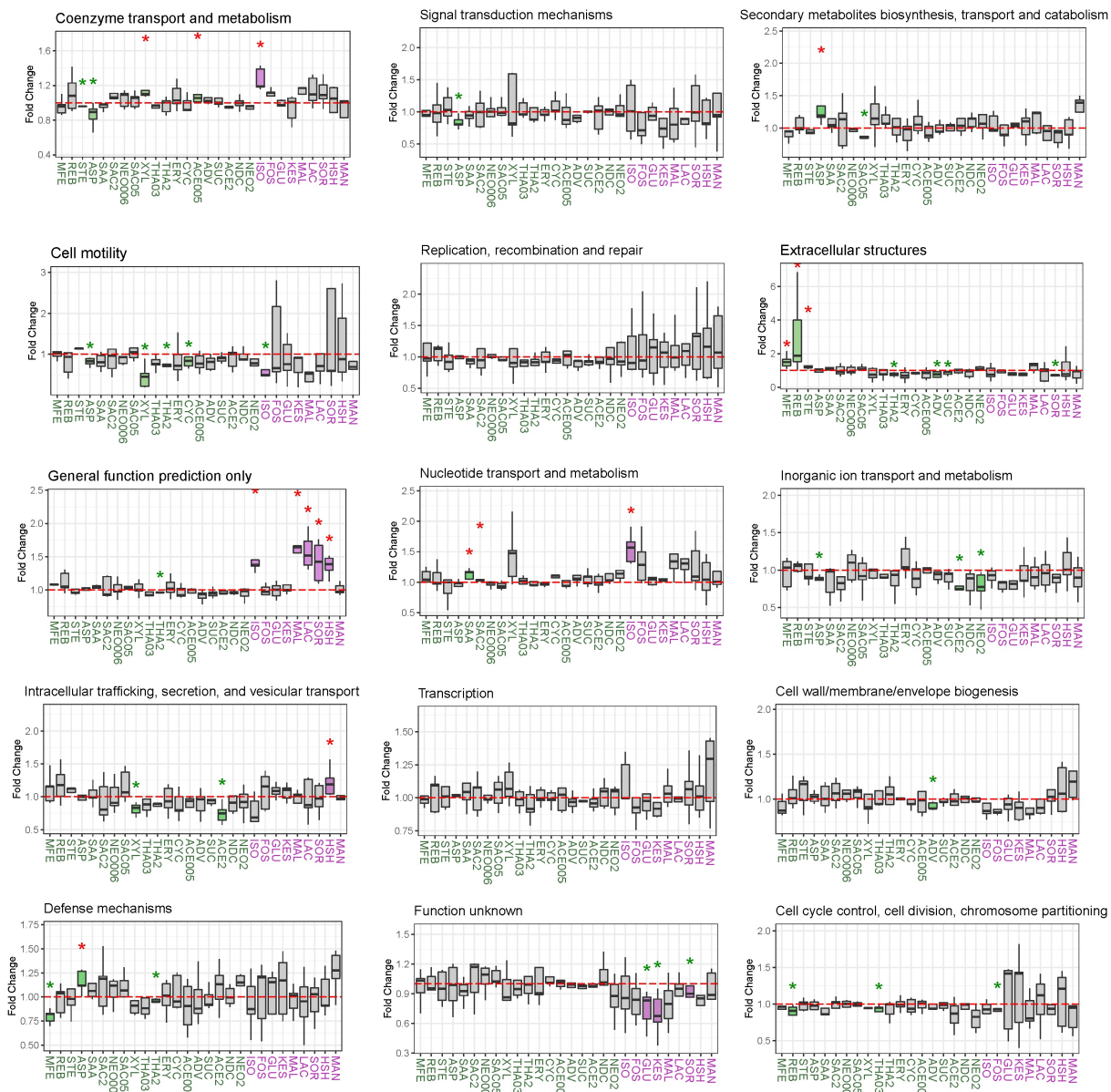


Supplementary Figure S2.2. Individual PCA plots of other sweeteners, positive control KES and dietary sugar control GLU.



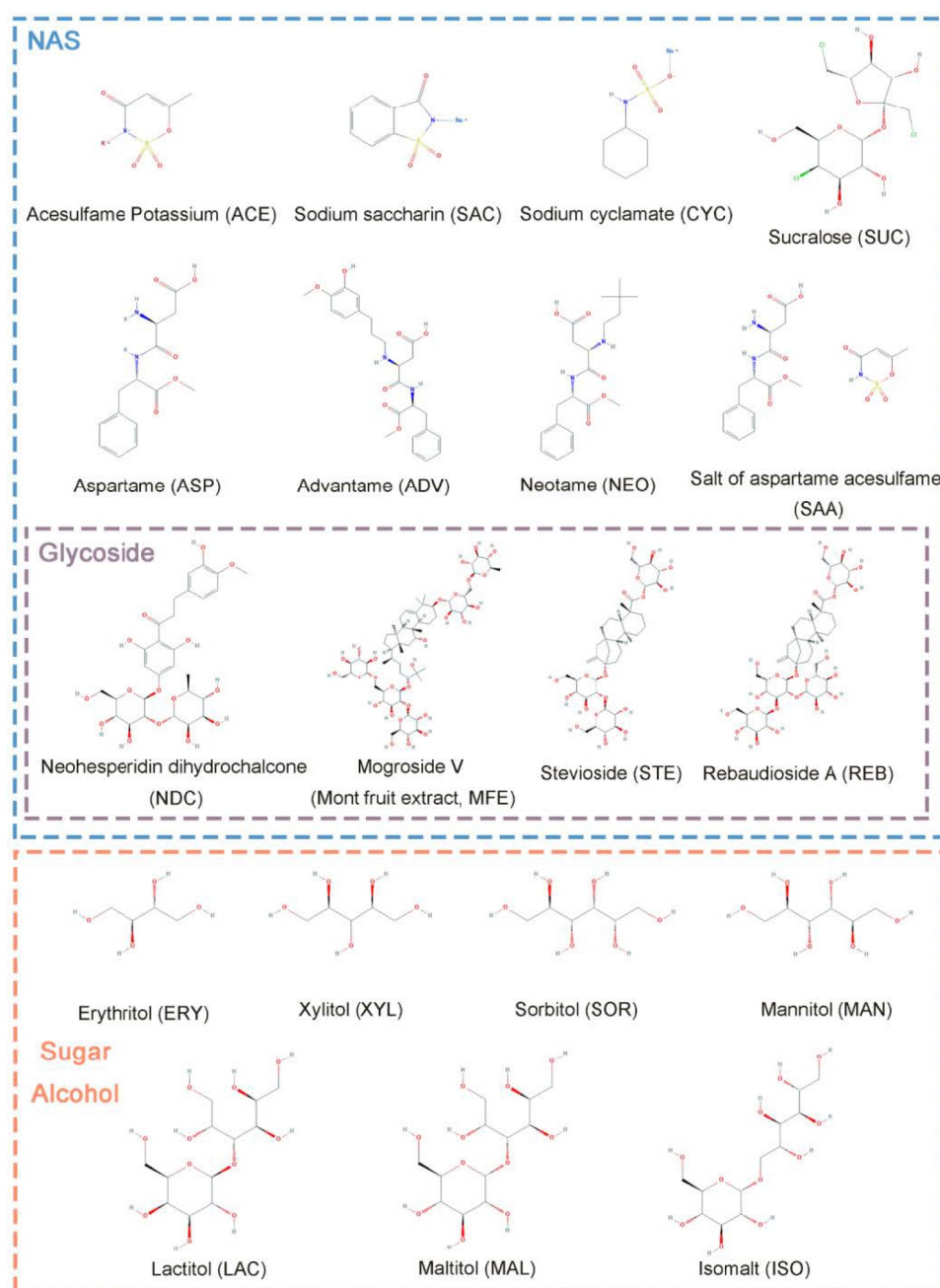
Supplementary Figure S2.3. Fold change of total protein amount obtained for each treatment group compared to PBS control.

2.7. SUPPLEMENTARY MATERIALS



Supplementary Figure S2.4. Responses of other COG categories in addition to Figure 2.3.

2.7. SUPPLEMENTARY MATERIALS



Supplementary Figure S2.5. Structure of sweeteners.

2.7.2 SUPPLEMENTARY TABLES

Supplementary Table S2.1. Detailed information of sweeteners used in the study, including classification, regulatory approval, and dosage.

Sweetener*	Classification	Abbreviation in this paper	PubChem CID	Molecular Weight (g/mol)	Approved by	Supplier and catalog number	Acceptable Daily Intake (ADI)	cADI (mg/mL)	Proportion that reaches the colon (%)	Concentration in medium (mg/mL)
Acesulfame K	NAS	ACE	11074431	201.24	HC, EFSA, FDA	TCI A1490	15 (FDA)	0.05	1 (Vogt et al., 2017)	0.05 (AE005) and 2 (ACE2)
Aspartame	NAS	ASP	134601	294.3	HC, EFSA, FDA	Alfa Aesar J61523	50 (FDA)	17.58	0 (Caballero et al., 2003)	2.00
Advantame	NAS	ADV	10389431	458.5	HC, EFSA, FDA	Sigma- aldrich 80054	32.8 (FDA)	10.31	89.5 (Aguilar et al., 2013)	2.00
Neotame	NAS	NEO	9810996	378.5	HC, EFSA, FDA	Sigma- aldrich 49777	0.3 (FDA)	0.07	63.7 (Aguilar et al., 2007)	0.067 (NEO006) and 2 (NEO2)
Saccharin (Sodium saccharin dihydrate)	NAS	SAC	517320	241.2	HC, EFSA, FDA	J&K 926097	15 (FDA)	0.53	10 (Xu et al., 2015)	0.5 (SAC05) and 2 (SAC2)

(Continued from previous page)

Sweetener*	Classification	Abbreviation in this paper	PubChem CID	Molecular Weight (g/mol)	Approved by	Supplier and catalog number	Acceptable Daily Intake (ADI)	cADI (mg/mL)	Proportion that reaches the colon (%)	Concentration in medium (mg/mL)
Sucralose	NAS	SUC	71485	397.6	HC, EFSA, FDA	Alfa Aesar J66736	5 (FDA)	1.49	85 (Roberts et al., 2000)	1.50
Stevia extract (Stevioside)	NAS	STE	442089	804.9	HC, EFSA, FDA	TCI S0594	4 (FDA)	1.41	100 (Xu et al., 2015)	1.40
Stevia extract (Rebaudioside A)	NAS	REB	6918840	967	HC, EFSA, FDA	TCI R0095	4 (FDA)	1.41	100 (Xu et al., 2015)	1.40
Monk fruit extract	NAS	MFE	N/A	N/A	HC, FDA	Sigma USP 1445492	6.8 (FDA)**	2.39	100 (Xu et al., 2015)	2.40
Thaumatococcus coccineus extract	NAS	THA	N/A	N/A	HC, EFSA, FDA	TCI T1144	1.1 (FDA)***	0.39	0 (Renwick, 1986)	0.39 (THA03) and 2 (THA2)
Cyclamate (Sodium cyclamate)	NAS	CYC	23665706	201.22	EFSA	Alfa Aesar A18666	7 (EFSA)	1.55	63 (Lyn, 2011)	1.60
Neohesperidin Dihydrochalcone	NAS	NDC	30231	612.6	EFSA	TCI N0675	5 (EFSA)	1.76	100 (Sundh et al., 2010)	1.80
Salt of Aspartame- Acesulfame	NAS	SAA	25130065	457.5	EFSA	Sigma USP 1043750	20.46 (EFSA)****	7.19	0 (Lyn, 2011)	2.00
Sorbitol (D-Sorbitol)	sugar alcohols	SOR	5780	182.17	HC, EFSA, FDA	TCI S0065	N/D	N/A	75 (Livesey, 2003)	2.00

(Continued from previous page)

Sweetener*	Classification	Abbreviation in this paper	PubChem CID	Molecular Weight (g/mol)	Approved by	Supplier and catalog number	Acceptable Daily Intake (ADI)	cADI (mg/mL)	Proportion that reaches the colon (%)	Concentration in medium (mg/mL)
Mannitol (D-Mannitol)	sugar alcohols	MAN	6251	182.17	HC, EFSA, FDA	TCI M0044	N/D	N/A	75 (Livesey, 2003)	2.00
Isomalt	sugar alcohols	ISO	88735	344.31	HC, EFSA, FDA	Sigma- aldrich PHR1769	N/D	N/A	90 (Livesey, 2003)	2.00
Maltitol	sugar alcohols	MAL	493591	344.31	HC, EFSA, FDA	TCI M0797	N/D	N/A	60 (Livesey, 2003)	2.00
Lactitol (Lactitol monohydrate)	sugar alcohols	LAC	3067270	362.33	HC, EFSA, FDA	J&K 126721	N/D	N/A	98 (Livesey, 2003)	2.00
Xylitol	sugar alcohols	XYL	6912	152.15	HC, EFSA, FDA	TCI X0018	N/D	N/A	50 (Livesey, 2003)	2.00
Erythritol (meso- Erythritol)	sugar alcohols	ERY	222285	122.12	HC, EFSA, FDA	TCI E0021	N/D	N/A	10 (Livesey, 2003)	2.00
Hydrogenated starch hydrolysates	sugar alcohols	HSH	N/A	N/A	HC, EFSA, FDA	CarboMer Inc. 68425-17-2	N/D	N/A	60 (Livesey, 2003)	2.00
Glucose (D-Glucose)	dietary sugar (positive control)	GLU	5793	180.16	N/A	N/A	N/A	N/A	N/A	2.00

(Continued from previous page)

Sweetener*	Classification	Abbreviation in this paper	PubChem CID	Molecular Weight (g/mol)	Approved by	Supplier and catalog number	Acceptable Daily Intake (ADI)	cADI (mg/mL)	Proportion that reaches the colon (%)	Concentration in medium (mg/mL)
1-Kestose	oligosaccharide (positive control)	KES	440080	504.4	N/A	N/A	N/A	N/A	N/A	2.00
Fructooligosaccharide	oligosaccharide (positive control)	FOS	N/A	N/A	N/A	N/A	N/A	N/A	N/A	2.00

*Names of the compounds that were used to represent the sweetener are shown in the brackets.

**The estimated 90th percentile intake of MFE in the general population by the FDA is used here, as its ADI is not specified (Heimbach, 2016).

*** The highest estimated exposure level of THA in the general population by the EFSA is used here, as its ADI is not specified (Canfora et al., 2015).

****The ADI of SAA is calculated based on ADI of both ASP and acesulfame, as specified by the EFSA (Hamer et al., 2007).

2.7. SUPPLEMENTARY MATERIALS

Table S2.2: Functional annotation of the top protein of the identified protein groups.

Table S2.3: Change of ComBat normalized protein group intensities under different treatment of sweeteners.

Table S2.4: Functional annotation of enriched discriminative proteins from Clostridia.

Supplementary Table S2.2, S2.3, and S2.4 could be accessed by:

<https://doi.org/10.1128/spectrum.00412-22>

A CORE PEPTIDE DATABASE FOR FASTER
HUMAN GUT METAPROTEOMICS
DATABASE SEARCHES

Preface

The following chapter consists of a research article of which I am the first author. The article was published in *Computational and Structural Biotechnology Journal*.

Zhongzhi Sun, Zhibin Ning, Kai Cheng, Haonan Duan, Qing Wu, Janice Mayne, and Daniel Figeys. MetaPep: A core peptide database for faster human gut metaproteomics database searches, *Computational and Structural Biotechnology Journal*, 2023.

Author contributions are described using the CRediT system as listed below.

Zhongzhi Sun: Data curation, Formal analysis, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Zhibin Ning:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Kai Cheng:** Methodology, Investigation. **Haonan Duan:** Methodology, Investigation. **Qing Wu:** Methodology, Investigation, Writing – review & editing. **Janice Mayne:** Resources, Project administration, Writing – review & editing. **Daniel Figeys:** Conceptualization, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

3.1 ABSTRACT

Metaproteomics has increasingly been applied to study functional changes in the human gut microbiome. Peptide identification is an important step in metaproteomics research, with sequence database search (SDS) and spectral library search (SLS) as the two main methods to identify peptides. However, the large search space in metaproteomics studies causes significant challenges for both identification methods. Moreover, with the development of mass spectrometry, it is now feasible to perform metaproteomic projects involving 100–1000 individual microbiomes. These large-scale projects create a conundrum for searching large databases. In this study, we constructed MetaPep, a core peptide database (including both collections of peptide sequences and tandem MS spectra) greatly accelerating the

peptide identifications. Raw files from fifteen metaproteomics projects were reanalyzed and the identified peptide-spectrum matches (PSMs) were used to construct the MetaPep database. The constructed MetaPep database achieved rapid and accurate identification of peptides for human gut metaproteomics. MetaPep has a large collection of peptides and spectra that have been identified in published human gut metaproteomics datasets. MetaPep database can be used as an important resource in the current stage of human gut metaproteomics research. This study showed the possibility of applying a core peptide database as a generic metaproteomics workflow. MetaPep could also be an important resource for future human gut metaproteomics research, such as DIA (data-independent acquisition) analysis.

3.2 INTRODUCTION

The human gut microbiome has been associated with human health and diseases. It is one of the most promising research areas for the development of new therapeutic approaches in the past two decades (de Vos et al., 2022; Proctor et al., 2019). Different omics technologies are important tools in the studies of the human gut microbiome (Franzosa et al., 2015). Among all these technologies, metaproteomics studies the proteins in microbial communities and directly shows the functionality of the gut microbiome. Metaproteomics is increasingly applied in the study of the human gut microbiome with the development of sample preparation processes, mass spectrometry instruments, and bioinformatics tools (Zhang & Figeys, 2019).

For metaproteomics studies, peptide identification is an important step, as all the downstream analyses are based on the peptide identification results. Sequence database search (SDS) and spectral library search (SLS) are two different methods for peptide identification. SDS searches the acquired spectra against the theoretical spectra generated with in-silico digested peptides without using the intensity information of each peak. In con-

trast, SLS searches the acquired spectra against the pre-identified experimental spectra using the intensity of each peak. There have been different search engines developed such as MaxQuant (Tyanova et al., 2016), pFind (Chi et al., 2018), and MSFragger (Kong et al., 2017) for SDS, and SpectraST (Lam et al., 2007) and COSS (Shiferaw et al., 2020) for SLS. As well, there are several SDS tools specific for the peptide identification of raw files from metaproteomics research, like MetaLab (Cheng et al., 2017), Galaxy-P (Jagtap et al., 2015), and MPA (Schiebenhoefer et al., 2020).

However, challenges remain for peptide identification from complex human gut microbiome samples. A major challenge is that the composition of the microbial community is unknown in most situations, and therefore a database that relies on published microbiome genomes is very large. In addition, SLS is absent from the current identification tools for the human gut microbiome. It has been shown that compared with SDS, SLS has higher sensitivity and lower computational complexity (Zhang et al., 2011). Therefore, it is worth exploring the performance of applying SLS to peptide identification in microbiomes. This exploration is very timely as DIA (data-independent acquisition) mode proteomics is becoming more popular. Traditional DDA (data-dependent acquisition) mode proteomics acquires MS/MS scans with narrow isolation windows centered on peptide precursors detected in an MS scan. In contrast, DIA acquires MS/MS scans with wide isolation windows that do not target any particular precursor. It has shown that DIA has advantages over DDA in terms of proteome coverage, reproducibility, and accuracy in quantification (Aakko et al., 2020). However, DIA mode proteomics relies on either pre-identified spectra, or predicted spectra based on a small peptide sequence database to resolve the complex spectra acquired (Ludwig et al., 2018). There is a greater need for an applicable peptide database for human gut metaproteomics to perform faster database searches.

In this study, we collected and re-analyzed previously published raw files from human gut metaproteomics studies, and combined the identified spectra and peptides to construct a core peptide database (MetaPep) of the human gut microbiome. MetaPep achieved rapid

and accurate identification of peptides from human gut metaproteomics data, demonstrating that searching a small peptide database can lead to rapid and accurate peptide identification. MetaPep (including a peptide sequence database and a spectral library) could also be a resource for future metaproteomics research.

3.3 MATERIALS AND METHODS

3.3.1 RAW FILE COLLECTION AND PEPTIDE IDENTIFICATION

To collect metaproteomics raw files of the human gut microbiome, a literature search was performed with *Web of Science* in January 2022. We searched literature with topics on the human gut and metaproteomics. For the papers returned from the search results, we manually verified their content to collect PRIDE (Perez-Riverol et al., 2022) (Proteomics IDentifications Database) accession IDs. As the HCD method coupled with FTMS analysis has higher resolution and accuracy and is able to improve peptide identification (McAlister et al., 2010), to increase the number of peptide sequence collections and keep identical product ions characteristics in the spectral library, only raw files from human gut microbiome samples with MS2 scan mode in HCD-FTMS (Higher energy Collisional Dissociation - Fourier Transform Mass Spectrometry) format were compiled. Raw files with the same PRIDE accession ID were considered as a project. Raw files from each project were submitted to the MetaLab MAG (Cheng et al., 2023) (version 1.0) with a pFind (Chi et al., 2018) search engine to perform a sequence database search to identify peptide sequences from the spectra in the raw files. During the MetaLab MAG search, human proteome fasta sequences were appended to the integrated UHGG (Almeida et al., 2021) (Unified Human Gastrointestinal Genome) database to generate a sample-specific database for the search of each project. Carbamidomethyl[C] was set as a fixed modification. Oxidation[M] and Acetyl[ProteinN-term] were set as variable modifications. All other parameters used the

3.3. MATERIALS AND METHODS

default settings in the MetaLab MAG with FDR (False discovery rate) on both peptide and protein levels set at 1%. MetaLab search results from each project were compiled to construct the spectral library.

3.3.2 PEPTIDE SEQUENCE DATABASE AND PEPTIDE SPECTRAL LIBRARY CONSTRUCTION

First, for each project, the *pFind-Filtered.spectra* file from the MetaLab MAG search results was converted to a *.tsv* format table that has six columns (raw file name, scan number, identified precursor, probability, score, and protein names). As we appended human protein sequences to the search space of the MetaLab MAG, the PSMs with peptides exclusively from human protein sequences were also removed. Simultaneously, all the downloaded *.raw* format raw files were converted to SpectraST (Lam et al., 2008) compatible *.mzXML* format using the msConvert from ProteoWizard (Chambers et al., 2012) toolkit (release 3.0.22005).

Next, the converted *.tsv* format table and *.mzXML* format raw files were used as inputs for SpectraST to generate a spectral library for each project (`spectrast.exe -cN`). Then libraries of all 15 projects were combined with a self-written script. If one peptide has been identified more than once, either in the same project or different projects, only the PSM with the highest pFind raw score was kept. The combined library was processed with SpectraST (`spectrast.exe -cJU -cAC -cN`) to reduce the size of the library by removing low-intensity peaks. Only the most intense 150 peaks of each spectrum were kept to produce a consensus library. Then the consensus library was further processed with SpectraST (`spectrast.exe -cAQ -cN`) for quality filtering based on the spectra quality, and as a result of this processing, the final spectral library was generated. Peptide sequences in the spectral library were extracted, with charge state and modification removed, to generate a non-redundant stripped peptide sequence database. In this manuscript, MetaPep refers to

3.3. MATERIALS AND METHODS

both the peptide sequence database and the peptide spectral library resulting from the process described above.

3.3.3 EVALUATION OF THE SEARCH SPEED AND PEPTIDE IDENTIFICATION

Two sets of metaproteomics raw files (PXD033624, PXD017467) were selected to evaluate the performance of MetaPep for human gut metaproteomics. Raw files were searched against the peptide sequence database with SDS methods and they were also searched against the spectral library with SLS methods. For SDS, two SDS tools, MaxQuant (Tyanova et al., 2016) (version 1.5.2.8) and pFind (Chi et al., 2018) (version 3.1.5) were selected. Each raw file dataset was searched against the peptide sequence database with the default parameters of these two search tools. The FDR was set at 1% at the peptide level and the FDR control of search results was conducted in the MaxQuant and pFind. For SLS, two SLS tools, SpectraST (Lam et al., 2007) (version 5.0) and COSS (Shiferaw et al., 2020) (version 1.1) were selected. Decoy libraries were generated with both SpectraST (-cAD -cc -cy 1) and COSS (using the reverse sequence strategy). The target-decoy spectral library generated with COSS was searched with SpectraST (SC), and the target-decoy spectral library generated with SpectraST was searched with SpectraST (SS). All the SLSs were performed with the default parameters of the SpectraST. COSS was not used to perform the search directly because of its limited capacity for handling large-size spectral libraries. Only the Top-1 PSM hit was collected for the peptide level FDR control at 1%. The MetaLab MAG search for these raw files was also performed with default parameters. The search speed was evaluated using our in-house Windows 10 Pro workstation with Intel(R) Xeon(R) Gold 5118 CPU @ 2.30 GHz 2.29 GHz (2 processors) and 192 GB RAM, and the search time of different methods was calculated from the records in the log file.

3.3. MATERIALS AND METHODS

After the FDR control, identified peptides by searching MetaPep with different search methods were compared with the identified peptides by MetaLab MAG. In addition, the pFind raw score of PSMs identified by searching MetaPep with pFind was compared to that of PSMs identified by MetaLab MAG. The Wilcoxon test of the raw score was conducted with *ggsignif* (version 0.6.4) package in R. To better evaluate the identification results, the MaxQuant scores of PSMs identified by searching MetaPep were collected as well. And PMD (precursor mass discrepancy) was also calculated from MaxQuant search results as described in the previous study (Hubler et al., 2020). The peptide quantification results of the two sets of raw files from MetaLab MAG and MetaPep (pFind search followed by FlashLFQ (Millikin et al., 2018)) were also compared. PCA (Principal component analysis) of the quantified peptides was performed with the *prcomp* function in the R package *ggfortify* (version 0.4.15).

To perform a more comprehensive evaluation of peptide identification by MetaPep on human gut metaproteomics raw files, another dataset (PXD032997) of six raw files was also searched against the MetaPep peptide sequence database and the MetaPep spectral library as described above. Raw files from this dataset were analyzed with MetaLab MAG with default parameters as well. The identified peptides from different methods were compared.

3.3.4 EVALUATION OF THE SEARCH SPECIFICITY AND SENSITIVITY

Raw files from *Pyrococcus furiosus* (Vaudel et al., 2012) (PXD001077) and human *Hela* cells were selected to evaluate the search specificity while searching MetaPep. Protein sequences from *Pyrococcus furiosus* (from Uniprot (Bateman et al., 2021)) and protein sequences of human *Hela* cells (Robin et al., 2018) were in-silico digested with ProteinDigest.py from msproteomicstools (Aebersold et al., 2016) (max miscleavage = 2 and minimum peptide length = 7). These raw files were first searched with the in-silico digested peptides of corresponding samples, with pFind and the same parameters for the previous metaproteomics

3.3. MATERIALS AND METHODS

files. Then raw files were searched against the MetaPep peptide sequence database with SDS tools MaxQuant and pFind, and also were searched against the MetaPep spectral library with SLS methods SC and SS, as described in the previous section. The identified peptides from each raw file were compared with the in-silico digested peptides of corresponding samples.

Raw files of four bacterial strains (*Escherichia coli* DSM 101114, *Bacteroides thetaio-taomicron* ATCC 29148^T, *Bacteroides ovatus* ATCC 8483^T, and *Blautia hydrogenotrophica* DSM 10507^T) were used to evaluate the search sensitivity of MetaPep. All these bacterial species are known to be present in the human gut microbiome. Single-strain raw files were acquired as described in previous studies (Cheng et al., 2023; Li et al., 2023). These raw files were searched against the peptide sequence database with SDS tools MaxQuant and pFind, as well as against the spectral library with SLS methods SC and SS, as described in the previous section. These raw files were also analyzed with MetaLab MAG using default parameters. Protein sequences of each strain were extracted from the UniProt and in-silico digested with ProteinDigest.py from *msproteomicstools* (Aebersold et al., 2016) (max miscleavage = 2 and minimum peptide length = 7). The identified stripped peptides from each raw file were compared with the in-silico digested peptides of the corresponding strain.

3.3.5 TAXONOMIC PROFILING

As all peptide sequences in MetaPep were extracted from MetaLab MAG results by searching the well-annotated genomes from UHGP database, these peptide sequences were able to be taxonomically annotated. Proteins from 4,744 representative genomes in the UHGP (Almeida et al., 2021) database were in-silico digested with ProteinDigest.py from *msproteomicstool* (Aebersold et al., 2016) (max miscleavage = 2 and minimum peptide length = 7). Peptides that could only be found in one bacterial genome were referred as to genome-

3.3. MATERIALS AND METHODS

distinct peptides. The identification of a genome-distinct peptide from MetaPep leads to a specific bacterial genome. After that, for each of the peptides that could be found in multiple genomes, the peptide was assigned to the LCA (last common ancestor) of the shared genomes and taxonomically annotated. And the identification of these peptides leads to a higher taxonomy level.

A SIHUMIx (Simplified Human Intestinal Microbiota) raw file dataset with eight known bacterial species from a previous publication (Van den Bossche et al., 2021) was searched against MetaPep with pFind, and the identified genome-distinct peptides were applied for the taxonomy profiling. The Sankey plots were generated with *sankeyNetwork* function from *networkD3* (version 0.4) package in R. After that, the peptides identified by MetaPep with taxonomy annotations at the family and genus levels were also applied for taxonomic profiling. The taxonomic profiling acquired by MetaPep was compared with the taxonomy profiling acquired by MetaLab MAG. In addition, a raw file from the previous additional dataset (PXD032997) was applied for the evaluation of the taxonomic profiling of MetaPep on complex real metaproteomics samples. And the taxonomic profiling of MetaPep was also compared with the taxonomic profiling acquired by MetaLab MAG.

3.3.6 PRELIMINARY EVALUATION OF APPLYING METAPEP FOR DIA RAW FILES

Three Q-Exactive type DIA raw files of the human gut microbiome were downloaded from the PRIDE (Perez-Riverol et al., 2022) database with accession ID PXD008738. Raw files were searched against both the MetaPep peptide sequence database and the MetaPep spectral library with DIA-NN (Demichev et al., 2020) using default parameters. The peptide identifications from searching MetaPep were then compared with the identification results from the original paper (Aakko et al., 2020).

Additionally, an in-house DIA-PASEF raw file was also searched against both the

3.4. RESULTS

MetaPep peptide sequence database and the spectral library with DIA-NN (Demichev et al., 2020) using default parameters to evaluate the performance of MetaPep on DIA-PASEF raw files. The methods for the acquisition of this in-house DIA-PASEF raw file were as described by Gomez-Varela et al. (Gómez-Varela et al., 2023a), with minor modifications as detailed in Supplementary Methods.

3.4 RESULTS

The overall workflow for the construction and evaluation of MetaPep is shown in Figure 3.1. We collected raw files from different human gut metaproteomics studies which were reanalyzed by MetaLab. The PSM with the highest pFind score across all samples was used to construct MetaPep, constituted with tandem MS spectra and peptide sequences. The search time, peptide identification, and sample profiling of searching MetaPep were then evaluated. The search specificity and sensitivity of searching MetaPep were evaluated with single species raw files. MetaPep was also assessed for the application of the taxonomic profiling of human gut microbiomes.

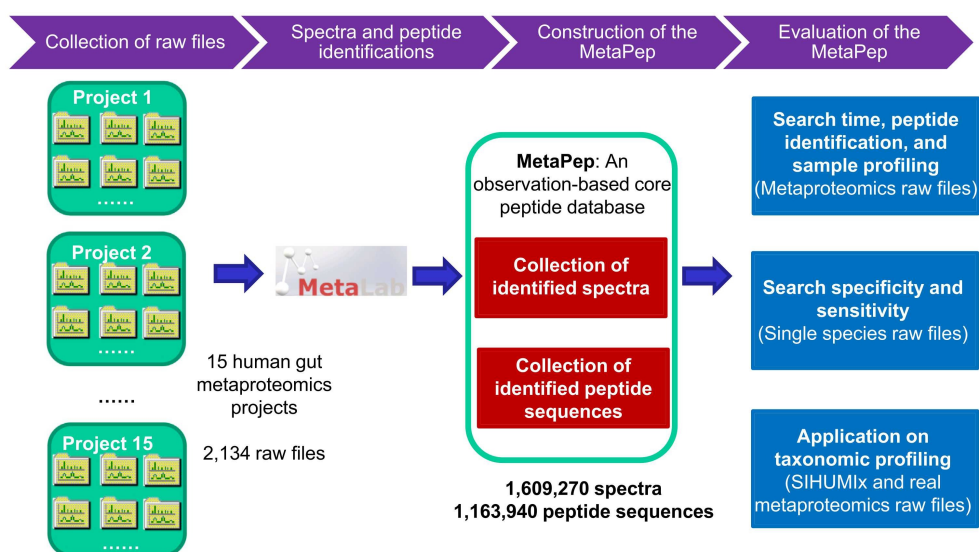


Figure 3.1. Workflow of the construction and evaluation of MetaPep.

3.4.1 METAPEP FROM IDENTIFIED SPECTRA AND PEPTIDE SEQUENCES

The literature search on the *Web of Science* returned 139 articles with topics on the human gut and metaproteomics. Thirty of these papers had at least one PRIDE accession ID for metaproteomics raw files of the human gut microbiome. Finally, 15 projects (each project has a unique PRIDE accession ID) from 13 articles with MS2 scan mode in HCD + FTMS were selected to perform the peptide identification using MetaLab MAG. These projects encompassed 2,134 raw files and 415 individuals (Detailed in Supplementary Table S3.1).

Raw files from each project were reanalyzed by MetaLab MAG separately. The number of identified peptides ranged from 26,333 to 888,139 for each project (Supplementary Table S3.2 and Supplementary Figure S3.1). After removing duplicated identification results from the different projects, in total, 2,438,058 spectra and 1,524,162 stripped peptides were identified from all 15 projects (Supplementary Figure S3.2). With PSMs from human protein sequences removed, a peptide sequence database with 1,163,940 peptide sequences Supplementary Data and a 7.77 GB spectral library with 1,609,270 spectra were constructed. Both the MetaPep peptide sequence database and the MetaPep spectral library have been uploaded to the open dataset repository, Zenodo (<https://zenodo.org/>), with DOI: 10.5281/zenodo.8101702. Distribution of the pFind raw score of collected PSMs, as well as the charge state and MW of collected precursors, and the length of collected peptides were shown in Supplementary Figure S3.3.

3.4.2 SEARCHING METAPEP ACHIEVED FASTER SEARCH SPEEDS WITH COMPARABLE IDENTIFICATIONS FOR HUMAN GUT METAPROTEOMICS

Searching MetaPep with different search methods was faster than canonical MetaLab MAG search (Figure 3.2A). Searching the peptide database with the SDS tool pFind had the fastest speed, at an average of 16.2% time of the MetaLab MAG search. Not only faster,

3.4. RESULTS

searching MetaPep with SDS tools MaxQuant and pFind also achieved a large number of high-quality identifications. Searching MetaPep with SDS tools MaxQuant and pFind identified similar, or larger numbers of peptides than the traditional MetaLab MAG search for raw files from both datasets (Figure 3.2B and C). All raw files from the same dataset showed similar trends. (Supplementary Figure S3.2 and Supplementary Figure S3.5). In addition, for both datasets, a large fraction of the identified peptides from searching MetaPep with SDS methods overlapped with the identified peptides from MetaLab MAG. Specifically, for the SDS tool MaxQuant, an average of 66.6% and 45.5% of the identified peptides overlapped with the MetaLab MAG in the two datasets. And for the pFind, an average of 74.5% and 42.9% of the identifications overlapped with the MetaLab MAG. The SLS methods did not perform as well as the SDS methods, searching spectral libraries with SLS methods identified fewer peptides than MetaLab MAG, especially in the PXD017467 dataset (Figure 3.2). The peptide identification result from the additional dataset (PXD032997) showed a similar but slightly different pattern (Supplementary Figure S3.6) compared to the PXD017467 dataset. For PXD032997 dataset, searching MetaPep peptide sequence database also identified a comparable number of peptides with MetaLab MAG. While searching MetaPep spectral library identified a limited number of peptides, it was higher than the number identified in PXD017467 dataset. The Venn Plot of the peptide identification results from all three datasets (Supplementary Figure S3.7, Supplementary Figure S3.8, and Supplementary Figure S3.9) showed that most of the identified peptides from different search tools were verified by at least one other search tool. In addition, many peptides were only identified by either SDS methods or SLS methods, indicating different search methods have their own preference.

3.4. RESULTS

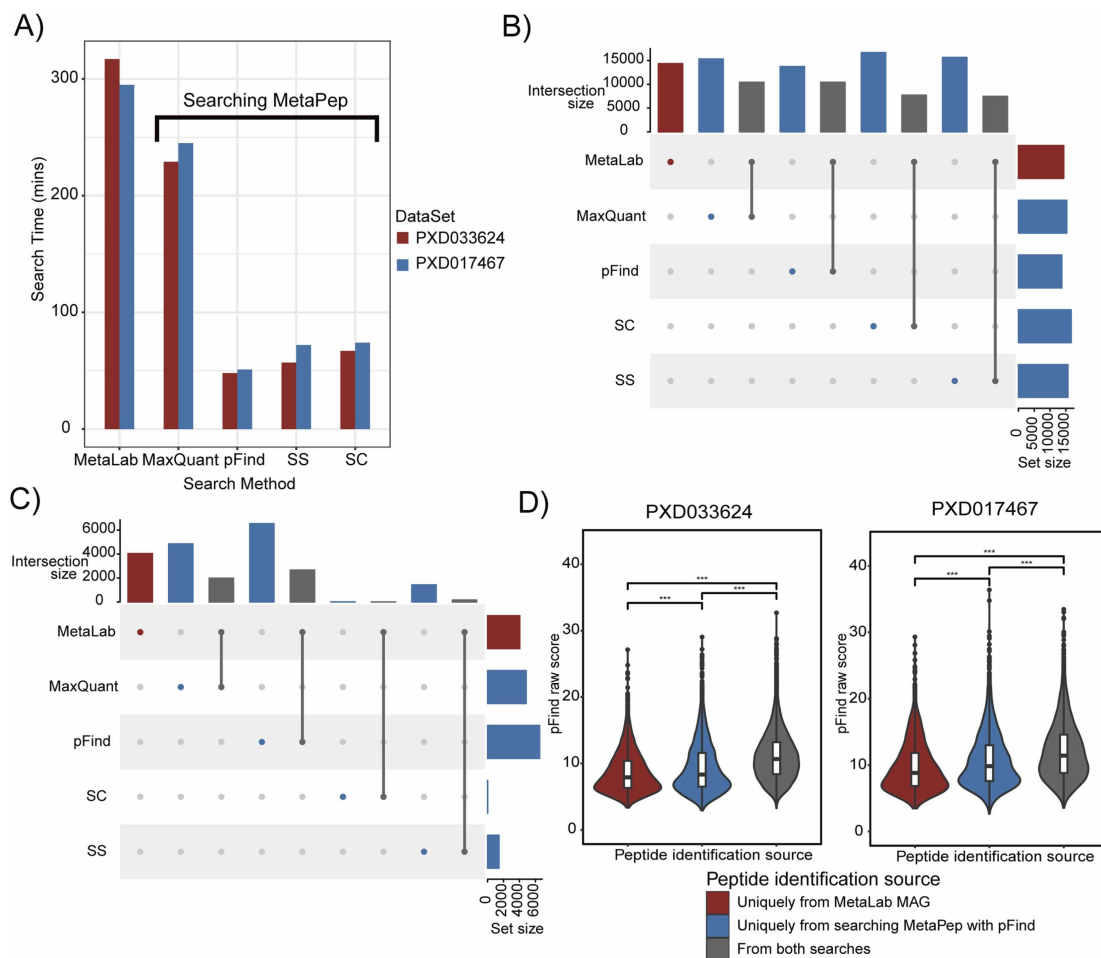


Figure 3.2. The peptide identification by searching human gut metaproteomics raw files against MetaPep. A) The search time of searching raw files with MetaLab MAG and searching raw files against MetaPep with different search methods. Raw files were from two different datasets and each dataset had six raw files. B-C) The number of identified peptides and the overlap of identified peptides between the MetaLab MAG search and searching MetaPep with different search methods for a representative raw file from B) the PXD033624 dataset and C) the PXD017467 dataset. Red bar: Identified peptides from MetaLab MAG search. Blue bar: Identified peptides from searching MetaPep with different search methods. Grey bar: The overlap of identified peptides between MetaLab MAG search and searching MetaPep. D) The distribution of pFind raw scores for PSMs that were exclusively identified by MetaLab MAG, exclusively identified by searching MetaPep with pFind, and identified by both methods from the representative raw files from the two datasets.

To investigate the reliability of identifications from searching MetaPep, the pFind raw score of PSMs from MetaLab MAG and PSMs from searching MetaPep with pFind were compared. In both test datasets, PSMs identified from both MetaLab MAG and searching MetaPep had higher raw scores than PSMs uniquely identified by one method (Fig-

ure 3.2D). PSMs exclusively identified by searching MetaPep had a higher raw score than PSMs exclusively identified by the MetaLab MAG (Figure 3.2D), verifying the reliability of identifications from searching MetaPep. The MaxQuant scores and the PMD of PSMs identified by searching MetaPep of the two datasets (PXD033624 and PXD017467) were shown in Supplementary Figure S3.10.

Searching MetaPep also showed the ability to discriminate raw files under different treatments and raw files of different individuals. Similar to the PCA plot of the identification and quantification results from MetaLab MAG, the PCA plot of the identification and quantification results from searching MetaPep separated raw files under two different treatments in PXD033624 dataset. (Supplementary Figure S3.11). Although there were differences in the quantification results of peptides that were uniquely quantified by either MetaLab MAG or MetaPep (Supplementary Figure S3.12 and Supplementary Figure S3.13). The quantification correlations of commonly quantified peptides from MetaLab MAG and MetaPep were highly correlated (Supplementary Figure S3.14 and Supplementary Figure S3.15).

3.4.3 SEARCHING METAPEP ACHIEVED PEPTIDE IDENTIFICATION WITH HIGH SPECIFICITY AND SENSITIVITY

The specificity of searching MetaPep was evaluated using proteomic raw files from samples with low overlaps with MetaPep including *Pyrococcus furiosus* (a hyperthermophile with a genome dramatically different from other species) and human *Hela* cells. These samples were selected as their peptide identification results theoretically have limited overlap with the peptides from the human gut microbiome, as only 345 of 295,495 *Pyrococcus furiosus* in-silico digested peptides and 2,781 of 2,593,869 human *Hela* in-silico digested peptides overlapped with MetaPep. As positive controls, searching the *Pyrococcus furiosus* and the human *Hela* raw file against their in-silico digested peptides with pFind achieved ID rates

3.4. RESULTS

of 78.8% and 80.0%, respectively, indicating that these raw files were of good quality. In contrast, searching these raw files against MetaPep identified a limited number of PSMs. For the *Pyrococcus furiosus* raw file, most of the identified PSMs were peptides that do not overlap with the in-silico digested peptides of this archaea species, suggesting that these identifications were false positives (Figure 3.3A). Indeed, the highest identification rate was 0.48% (pFind), indicating a well-controlled false positive identification. For the human *Hela* raw file, searching MetaPep identified a higher number of PSMs. The estimated false positive identifications (PSMs with peptides that have no overlap with human *Hela* in-silico digested peptides) account for up to 1.19% (MaxQuant) of the total spectra in the raw file, and the estimated false positive identifications were lower than 1% for other search methods. The evaluation with the *Pyrococcus furiosus* and the human *Hela* raw files showed that searching MetaPep using the search methods MaxQuant, pFind, SS, and SC achieved high specificity in most situations.

3.4. RESULTS

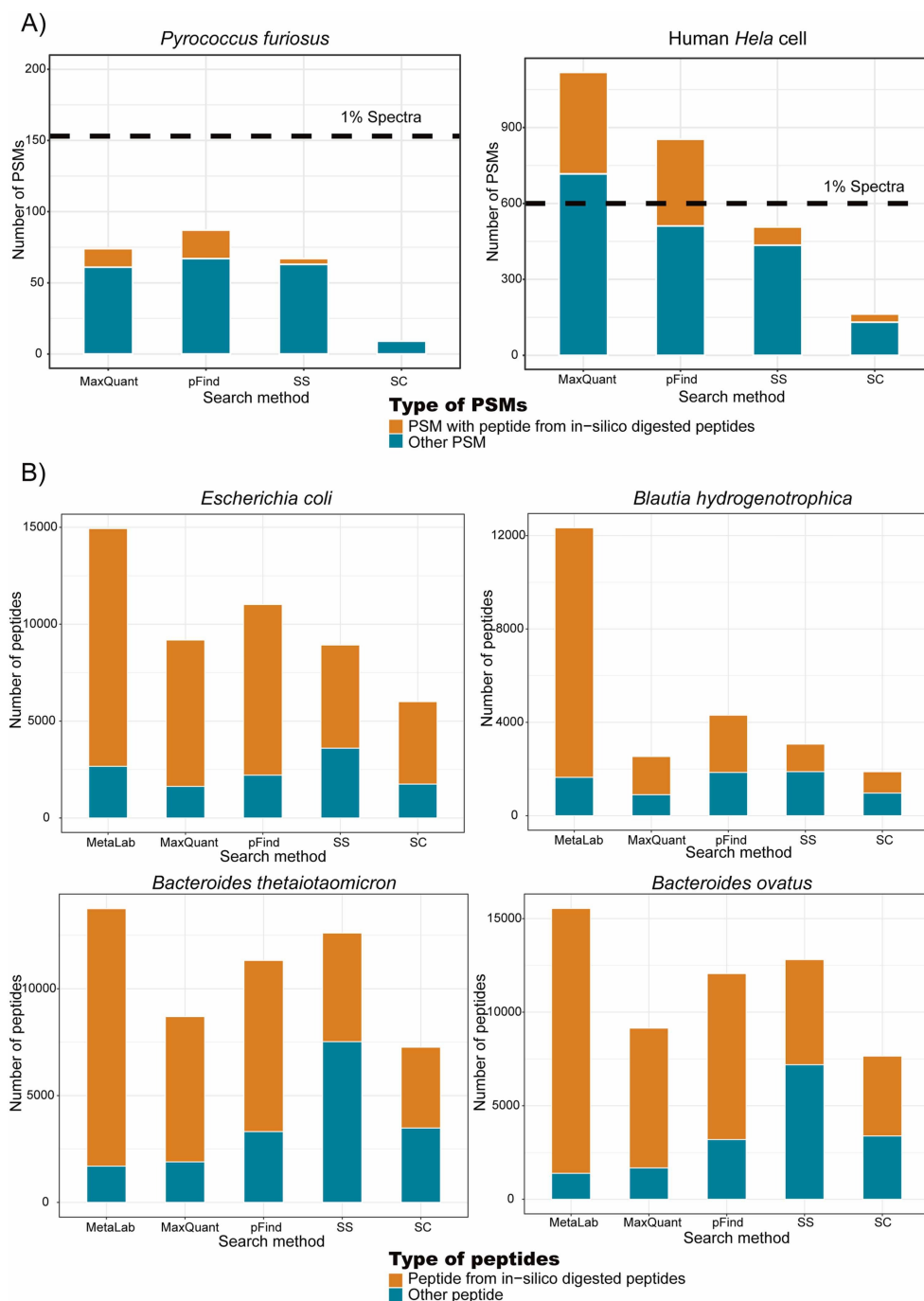


Figure 3.3. Evaluation of the specificity and the search sensitivity of searching MetaPep. A) The number of identified peptide-spectrum matches (PSMs) searching *Pyrococcus furiosus* and human *Hela* raw files against MetaPep at 1% FDR on peptide level. PSMs with peptides overlapped with in-silico digested *Pyrococcus furiosus* or human *Hela* peptides were in orange, and the PSMs with other peptides were in blue-green. B) The number of identified peptides in searching single strain raw files against MetaPep at 1% FDR on peptide level. Peptides overlapped with in-silico digested peptides of the corresponding bacterial strain were shown in orange, and the other peptides were shown in blue-green.

3.4. RESULTS

The sensitivity of searching MetaPep was evaluated by searching individual raw files from four bacteria known to be present in human gut microbiomes (*Escherichia coli* DSM 101114, *Bacteroides thetaiotaomicron* ATCC 29148, *Blautia hydrogenotrophica* DSM 10507^T, and *Bacteroides ovatus* ATCC 8483^T). The MetaLab MAG identified most peptides from single-strain raw files through its two-step search strategy against the comprehensive UHGP database (Figure 3.3B). In comparison, searching MetaPep identified fewer peptides than the MetaLab MAG. For more common strains in the human gut, (*Escherichia coli* DSM 101114, *Bacteroides thetaiotaomicron* ATCC 29148, and *Bacteroides ovatus* ATCC 8483^T) searching MetaPep with pFind identified 11,015, 12,060, and 11,321 peptides, respectively. For the less common strain in the human gut, *Blautia hydrogenotrophica* DSM 10507^T, searching MetaPep identified fewer peptides, ranging from 1,855 (SC) to 4,309 (pFind). In addition, for all the species, a large fraction of the identified peptides by SDS methods (64.6%–82.2% for MaxQuant and 56.9%–80.0% for pFind) overlapped with the in-silico digested peptides of the corresponding species (Figure 3.3B). The identification of a large number of single-strain digested peptides from single-strain raw files showed the high sensitivity of searching MetaPep with SDS methods.

3.4.4 METAPEP PRESERVED THE ABILITY OF TAXONOMIC PROFILING

After collecting genome-distinct peptides from the UHGP database and selecting those peptides from MetaPep, it was found that 25.2% of peptides in MetaPep were genome-distinct peptides (Figure 3.4A and B). In addition, in MetaPep, 25.3% and 13.8% of peptides were assigned to specific genera and families, respectively. The left peptides were assigned to higher taxonomic levels (Supplementary Figure S3.16). In contrast, 84.5% of the in-silico digested peptides from UHGP were only found in a single specific genome (Supplementary Figure S3.17). In total, 4,149 of 4,744 representative bacterial genomes have genome-distinct peptides in MetaPep, with 1,928 of them having ≥ 10 genome-distinct peptides (Figure 3.4C). However, when it came to higher taxonomy levels, by summing the num-

3.4. RESULTS

ber of genome-distinct from the same genus or same family, 421 of 480 genera have ≥ 10 genome-distinct peptides, and 127 of 137 families have ≥ 10 genome-distinct peptides. The high-level coverage of genome-distinct peptides at the genus and family level enabled rough taxonomic profiling by searching MetaPep.

3.4. RESULTS

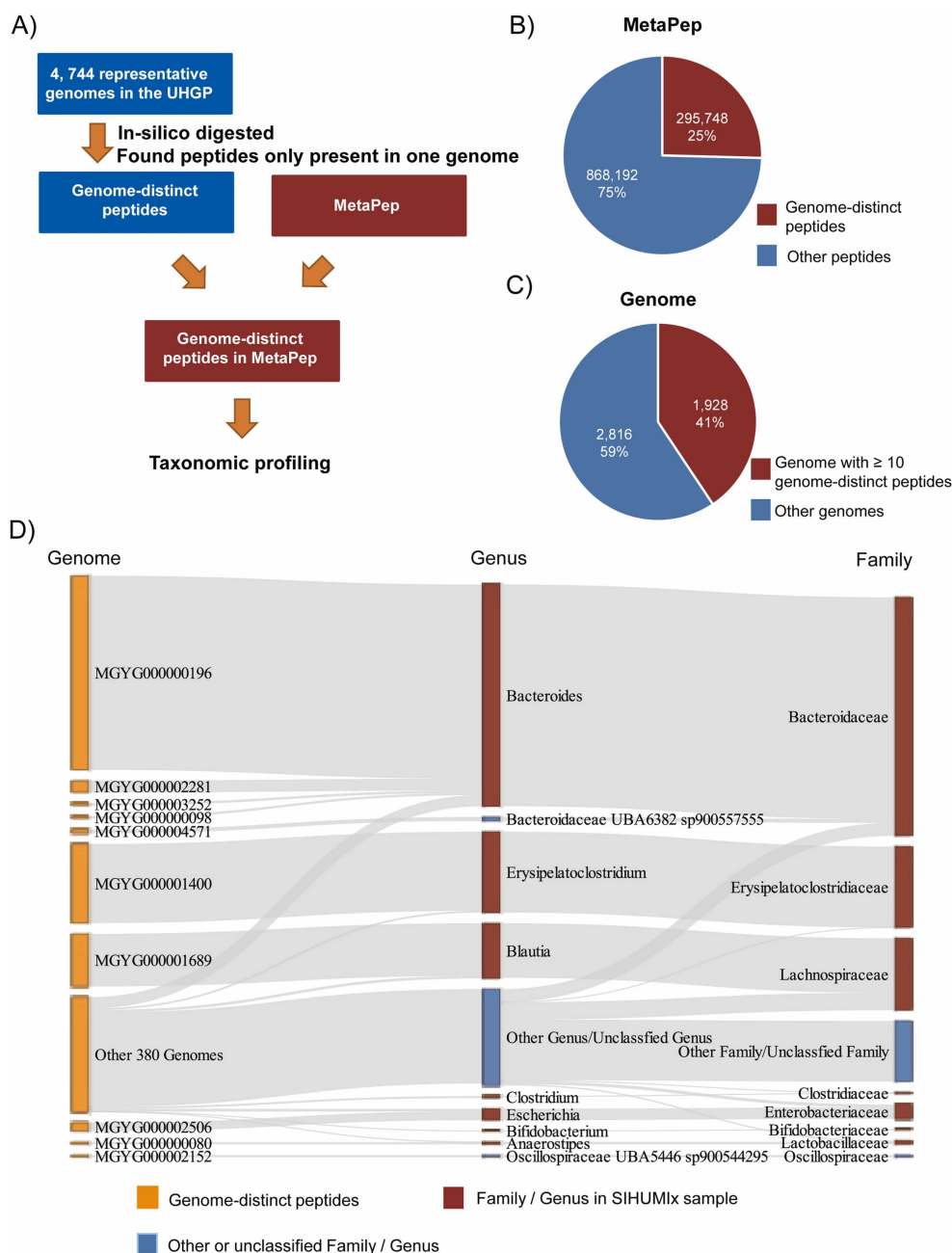


Figure 3.4. Application of genome-distinct peptides in MetaPep for taxonomic profiling. A) The workflow for collecting genome-distinct peptides from the UHGP database and selecting genome-distinct peptides in MetaPep. B) The number and the proportion of genome-distinct peptides in MetaPep. C) Proportion of genomes with ≥ 10 genome-distinct peptides in the UHGP representative bacterial genomes. D) Taxonomic profile of a SIHUMIx sample based on the genome-distinct peptides identified by searching MetaPep.

To evaluate the taxonomic profiling ability of MetaPep, a SIHUMIx raw file was searched against MetaPep, and the identified genome-distinct peptides were applied to

3.4. RESULTS

perform taxonomic profiling. The SIHUMIx was composed of eight species, and the previous metagenomics results revealed that the five most abundant species (*Bacteroides thetaiotaomicron*, *Blautia producta*, *Escherichia coli*, *Erysipelatoclostridium ramosum*, and *Anaerostipes caccae*) accounted for the majority of the sample’s abundance (Van den Bossche et al., 2021). Searching MetaPep with pFind identified 34,195 stripped peptides including 3,230 genome-distinct peptides. A large number of the identified genome-distinct peptides (1,323) were assigned to the genome of *Bacteroides thetaiotaomicron* (Figure 3.4D). The top-8 genomes with the most identified genome-distinct peptides covered all the five most abundant species revealed by metagenomics. In addition, most of the identified genome-distinct peptides were assigned to the genera and families from the SIHUMIx sample (Figure 3.4D). However, for the left three bacterial species in the SIHUMIx sample (*Clostridium butyricum*, *Lactobacillus plantarum*, and *Bifidobacterium longum*), only one genome-distinct peptide was assigned to *Clostridium butyricum* and no genome-distinct peptide was assigned to the other two species due to their low abundance (Van den Bossche et al., 2021). After that, the peptides identified by MetaPep with taxonomy annotations at the family and genus levels were also applied for taxonomic profiling. The four families with the highest number of identified peptides by MetaPep (Bacteroidaceae, Lachnospiraceae, Enterobacteriaceae, Erysipelatoclostridiaceae) were all present in the SIHUMIx sample, and these four families accounted for 96.2% of all peptides with family-level taxonomy annotations (Supplementary Figure S3.18). The five genera with the highest number of peptides identified by MetaPep (*Bacteroides*, *Blautia*, *Escherichia*, *Erysipelatoclostridium*, and *Anaerostipes*) were all present in the SIHUMIx sample, and these five genera accounted for 92.1% of all peptides annotated with genus-level taxonomy (see Supplementary Figure S3.18). We compared the family-level and genus-level taxonomic profiling of the SIHUMIx files obtained by MetaPep with the taxonomic profiling obtained by MetaLab MAG search. The four families and five genera mentioned above were also the most abundant in the taxonomic profiling of the MetaLab MAG. The four families accounted for 99.2% of

3.4. RESULTS

the total sample abundance, and the five genera accounted for 97.6% of the total sample abundance (Supplementary Figure S3.18).

To further test the effectiveness of MetaPep, we used it for the taxonomic profiling of a real human gut sample with more complex taxonomy compositions. Again, the taxonomic profiling obtained by MetaPep was compared with that obtained by MetaLab MAG. Searching MetaPep identified genome-distinct peptides from 240 UHGP representative species. Meanwhile, MetaLab MAG identified 133 representative species in the sample, with 33 representative species identified by both methods. However, the taxonomic profilings acquired by two different methods were more similar at higher taxonomy levels. At the family level, the four most abundant families identified by MetaPep (Bacteroidaceae, Lachnospiraceae, Ruminococcaceae, Tannerellaceae) accounted for 87.2% of all identified peptides with family-level taxonomic annotations (see Supplementary Figure S3.19). These four families were also the most abundant families identified by MetaLab MAG, accounting for 96.0% of the total abundance of the samples (Supplementary Figure S3.19). At the genus level, the six most abundant genera identified by MetaPep (*Faecalibacterium*, *Phocaeicola*, *Blautia*, *Agathobacter*, *Parabacteroides*, and *Bacteroides*) accounted for 58.9% of all peptides with taxonomic annotations at the genus level (see Supplementary Figure S3.19). These six genera were also the most abundant genera identified by MetaLab MAG, accounting for 80.5% of the total abundance of the samples (Supplementary Figure S3.19).

All these results suggest that searching MetaPep provides a reasonable taxonomical profile of human gut metaproteomics. And to facilitate the application of MetaPep for taxonomic profiling, the detailed taxonomic annotations of qualified MetaPep peptides (peptide lengths between 7 and 30) were provided in Supplementary Table S3.3.

3.4.5 METAPEP WAS APPLICABLE FOR DIA METAPROTEOMICS

For the three Q-Exactive type DIA raw files (PXD008738), searching MetaPep peptide sequence database and MetaPep spectral library with DIA-NN identified an average of 38,698 and 20,997 peptides, respectively (Supplementary Figure S3.20). These numbers were both much higher than the average of 7,294 identified peptides in the original study by searching a spectral library built from DDA raw files (Aakko et al., 2020). In addition, an average of 63.4% of peptides identified in the original study were identified either by searching MetaPep sequence database or searching MetaPep spectral library (Supplementary Figure S3.20).

For the in-house DIA-PASEF raw file of the human gut microbiome, searching MetaPep peptide sequence database and MetaPep spectral library with DIA-NN identified 25,302 peptides and 16,565 peptides, respectively. These identification results indicated that both the MetaPep peptide sequence database and the MetaPep spectral library could be valuable resources for employing DIA-PASEF technology in human gut metaproteomics studies.

3.5 DISCUSSION

A healthy individual harbors a complex gut microbial community composed of around 200 bacterial species with a large dynamic range (Duan et al., 2022). When applying metaproteomics to study this community, mass spectrometry-based approaches tend to identify peptides with preferential properties by the workflows (including sample preparation, instrument, etc.), and only proteotypic peptides were repeatedly and consistently identified from the protein mixture (Duan et al., 2022; Mallick et al., 2007). It is estimated that only 10% of genome-coding protein sequences could be identified in metaproteomics research (Zhang & Figeys, 2019), and the identified peptides account for an even lower coverage of genome-coding protein sequences. A popular view in proteomics suggested we should only

search for peptides we are interested in (Noble, 2015). This view is especially applicable in metaproteomics, as a large number of protein-coding sequences from the complex microbial community would result in a huge search space and long search time. However, the current database search strategy in metaproteomics research is searching against a search space filled with peptides that are unlikely to be identified. Here, we used the pre-identified spectra and peptide sequences to construct a core peptide database, MetaPep. We then verified the application of MetaPep to peptide identification in metaproteomics studies.

MetaPep compiles peptides and spectra identified from 2,134 raw files from fifteen published projects. These projects had broad representations, with samples collected from seven countries with various diet habits and involving people of different ages (Supplementary Table S3.1). It is worth noting that although the human proteome is an important component of the human gut metaproteomics samples (Armengaud, 2023), MetaPep mainly focuses on reducing the search space for bacterial species. PSMs with peptide sequences exclusively for humans were not included in MetaPep. When necessary, MetaPep is able to append peptide sequences and spectra in the available human spectral libraries from PeptideAtlas (Deutsch et al., 2015) or NIST (NIST, 2022). The publicly available MetaPep sequence database and the MetaPep spectral library could be applied as the input of different search engines, such as MaxQuant (Tyanova et al., 2016) and pFind (Chi et al., 2018) for SDS, SpectraST (Lam et al., 2007) for SLS, and DIA-NN (Demichev et al., 2020) for analyzing DIA raw files.

Searching MetaPep with SDS methods identified a large number of peptides with a high proportion of overlap with the search results of the MetaLab MAG (Cheng et al., 2023) while greatly speeding up the search (Figure 3.2A-C). Despite MetaLab MAG being significantly faster than traditional search methods by applying a two-step search strategy (Cheng et al., 2023, 2017), the smaller size of MetaPep further speeds up search times by about 6 times compared to MetaLab MAG. However, searching MetaPep with SLS methods did not perform as well as the SDS methods, especially for the PXD017467 dataset

(Figure 3.2C). It is worth noting that the reduced performance of SLS was improved in the additional dataset, PXD032997 (Supplementary Figure S3.6). This suggested that differences in the qualities of the raw files and the preference of the spectral library for the instrument with specific parameter settings, also impacted the SLS results, in addition to the effect of the MetaPep spectral library. Six out of the 15 projects collected in MetaPep were from our lab that had similar settings, and the PXD033624 dataset with more peptide identifications with SLS methods was also an in-house dataset. The application of the constructed spectral library to datasets with different parameters should be done with caution. In addition, more advanced spectral library searching tools might increase the identification by introducing more variable modifications as well as the adaption to the complex metaproteomics raw files. The constructed spectral library in MetaPep could not only be used for the general SLS for DDA raw files but could also be used for DIA proteomics. Although some spectra-prediction methods have shown to be able to achieve peptide identification from the DIA raw files without pre-identified experiment-acquired spectra, searching against the experimental spectra would give more confidence in the identification (Yang et al., 2021b, 2020). Additionally, the constructed spectral library in MetaPep could also be used for training machine learning models, MRM (multiple reaction monitoring) method developments, and peptide identification validation.

When evaluating the search sensitivity of MetaPep, the number of identified peptides from searching MetaPep was fewer than the MetaLab MAG search, especially for a specific strain, *Blautia hydrogenotrophica* DSM 10507^T (Figure 3.3). MetaLab MAG generates a sample-specific database on a genome/proteome base, which means that the whole proteome is incorporated for the search. In contrast, we constructed MetaPep using identified spectra and peptides from human gut metaproteomics, and peptides from low abundance species in complex microbiome samples be inevitably reduced in MetaPep. Our study has shown that peptides that could be identified from single-strain raw files were difficult to be identified from complex microbiome samples (Duan et al., 2022). Therefore it makes

sense that the MetaLab MAG search outperformed searching MetaPep generated by the peptide-centered strategy for single-strain samples. This disadvantage should be alleviated for metaproteomic samples with similar composition because of the dramatic complexity of the sample. During the evaluation of human metaproteomics raw files, searching MetaPep identified a similar or higher number of peptides than the MetaLab search, and the disparity of the number was different from searching single-strain raw files. In addition, the cross-validated identification results verified the reliability of searching MetaPep, and the greatly faster search time showed the advantage of searching MetaPep.

The application of the MetaPep for the taxonomic profiling of the SIHUMIx raw file provided a reasonable profile. However, the profiling was slightly different from metagenomics data and metaproteomics data analyzed using other methods (Van den Bossche et al., 2021). MetaPep identified the *Bacteroides thetaiotaomicron* as the most abundant species, but its relative abundance was lower than the abundance from other methods. There were also 413 (12.8%) identified genome-distinct peptides assigned to unrelated bacterial families. However, most of these peptides (219 of 413) have short peptide lengths (peptide length <10). For the left longer peptides (194), 76 of them had $\geq 80\%$ sequence identity with in-silico digested peptide sequences from the SIHUMIx species. The bias in the taxonomic profiling of the MetaPep could result from different reasons. The first is that there were limitations of the peptide identification algorithms (Colaert et al., 2011). It could be difficult to discriminate peptides with high similarity and resulting in the identification of highly similar peptides that were assigned to other species. Another reason is that MetaPep covered different microbial species at different depths. The number of genome-distinct peptides is different across different bacterial taxa and might result in bias in the taxonomic profiling. The coverage of MetaPep might still need to be improved by incorporating more datasets.

The number of identified spectra and peptides kept increasing when more projects were added (Supplementary Figure S3.2). However, the growth in the number of identi-

fied microbial species slowed down as more research projects were added (Supplementary Figure S3.21). At present, all the sequences included in MetaPep have covered 4,110 of the 4,744 UHGG representative species. However, MetaPep has differences in the coverage of different species. Taking the strict identification of 10 genome-distinct peptides in one project as a criterion for reliably identifying the corresponding genome (species), 1,183 human gut representative species have been reliably identified in all included projects (Supplementary Figure S3.21). With more metaproteomics studies included, the balance between the coverage and the size of MetaPep will be optimized in future versions. Our aim will be keeping MetaPep to a manageable size, without losing the performance in terms of sensitivity and speed.

3.6 CONCLUSION

In conclusion, we collected high-quality PSMs from 15 different human gut metaproteomics studies and constructed MetaPep consisting of 1,609,270 spectra and 1,163,940 stripped peptide sequences. Evaluation of the search time, peptide identification, search specificity, and search sensitivity verified the feasibility of applying MetaPep to peptide identification for the human gut microbiome. Searching MetaPep also showed the ability to perform taxonomic profiling. MetaPep can be used as a general database for human gut metaproteomics, in types of modes, like DDA and DIA, with or without the spectral library. Database searches were extremely fast, while still maintaining a decent identification rate and sequence coverage compared to the canonical protein sequence searches with a comprehensive fasta database. MetaPep could be used as a first-pass search strategy to create a sample-specific search space. The workflow to construct the core peptide database (MetaPep) could be generalized to other metaproteomics studies. More raw files from different sources and instruments will be incorporated into MetaPep in the future. We anticipate that as more metaproteomic studies and newer techniques to go deeper in

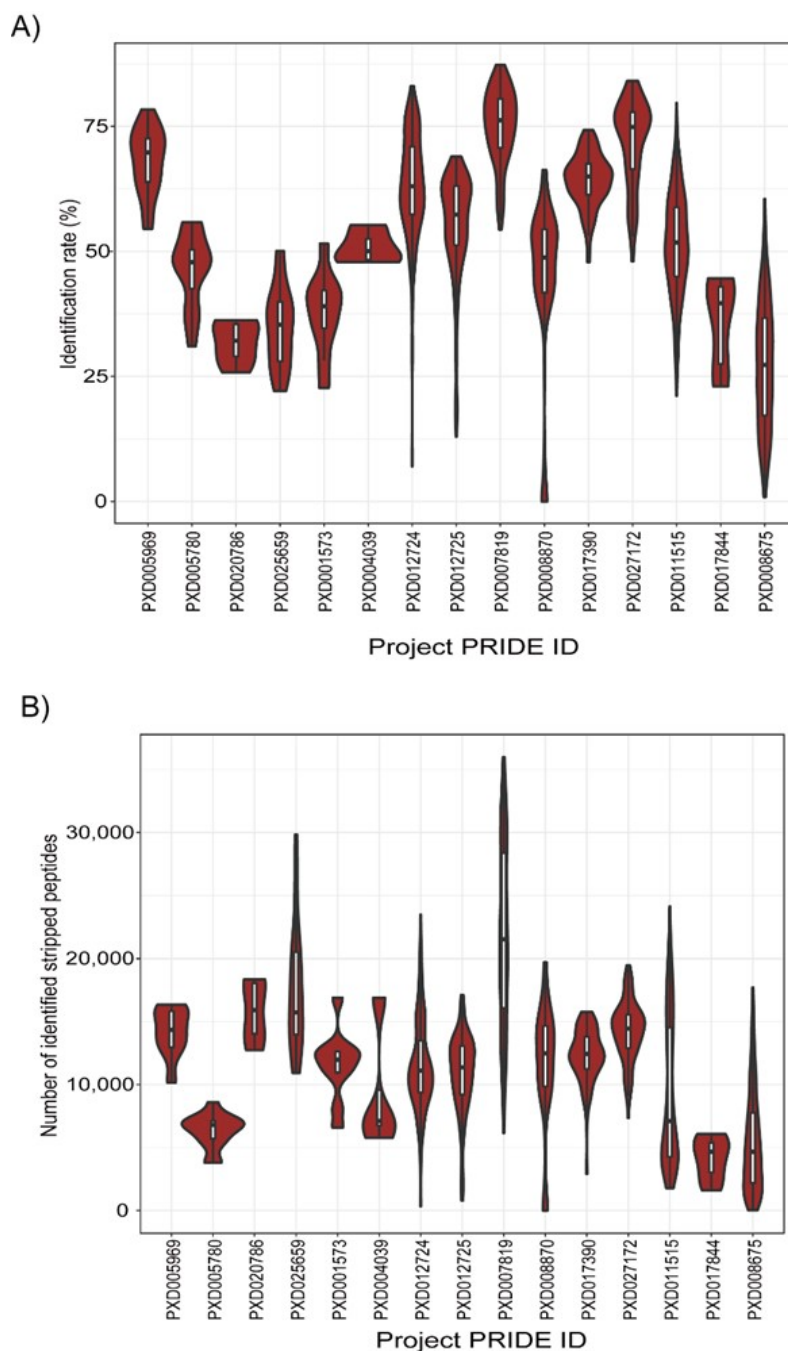
3.6. *CONCLUSION*

the metaproteome are reported, the number of peptides and spectra in MetaPep will increase. This will provide better peptide identification and taxonomic profiling of for lower abundance microbes.

3.6. CONCLUSION

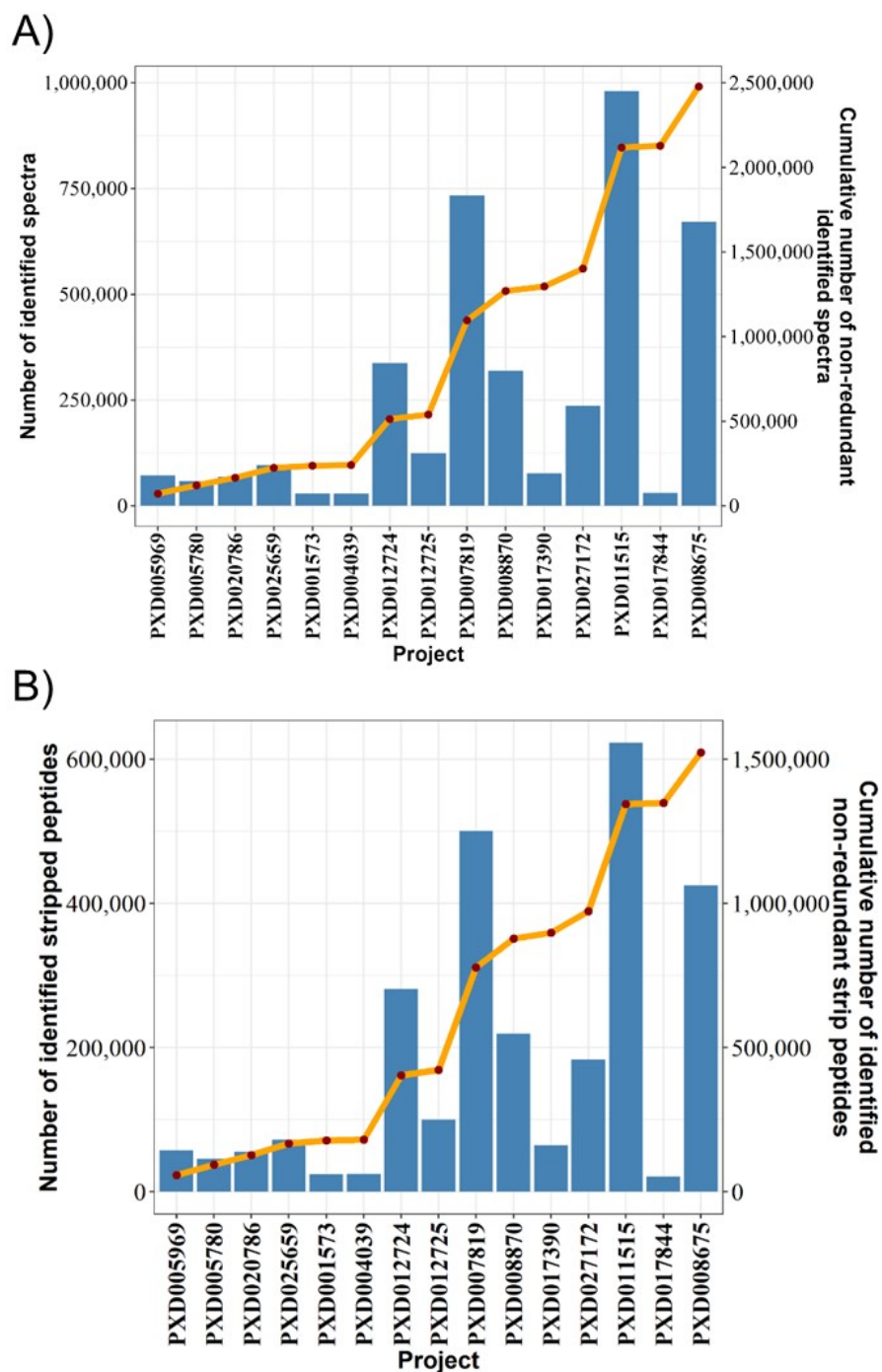
3.7 SUPPLEMENTARY MATERIALS

3.7.1 SUPPLEMENTARY FIGURES



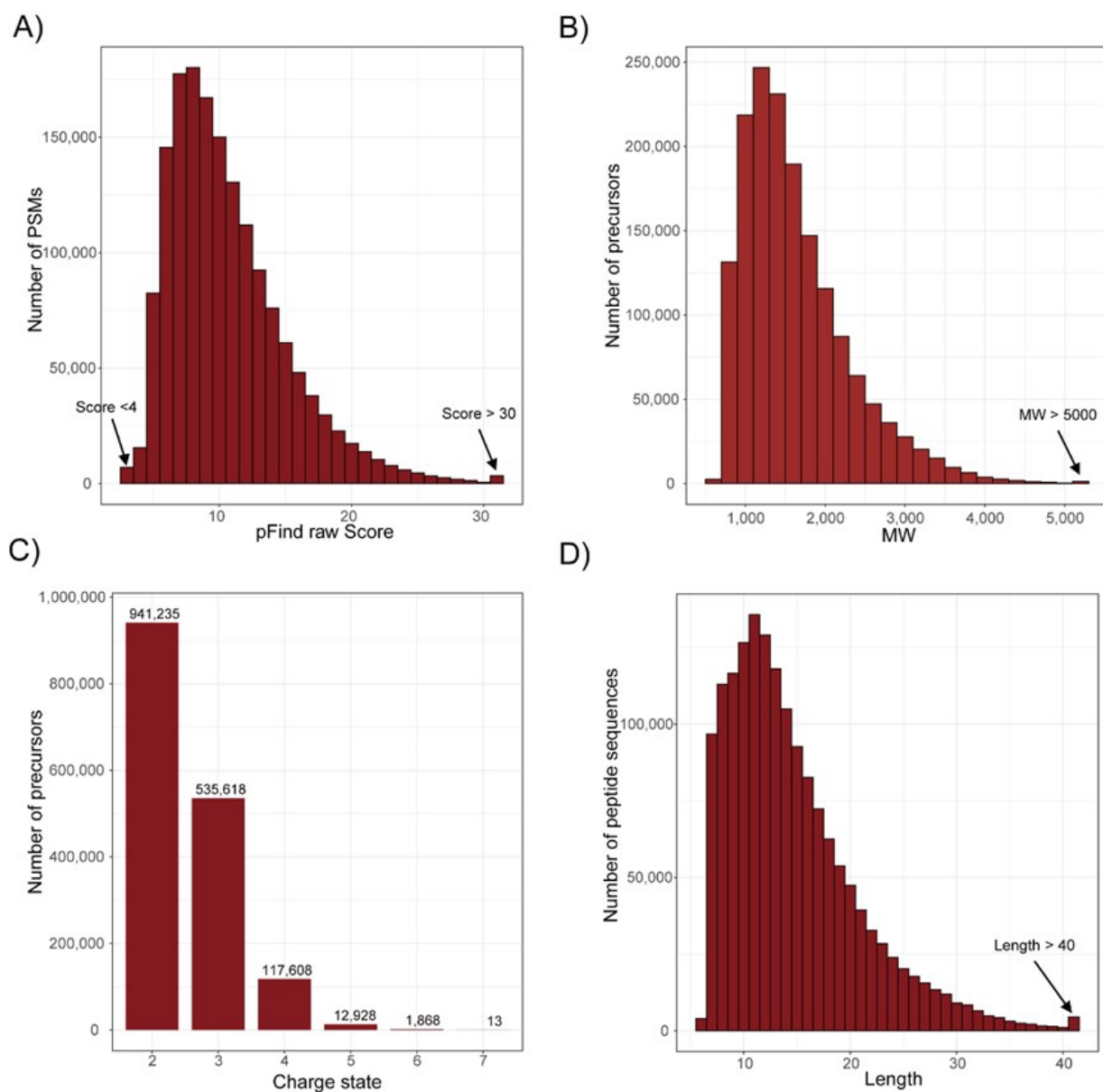
Supplementary Figure S3.1. The distribution of A) the ID rate and B) the number of identified stripped peptide sequences among the raw files of each project from the MetaLab MAG search.

3.7. SUPPLEMENTARY MATERIALS

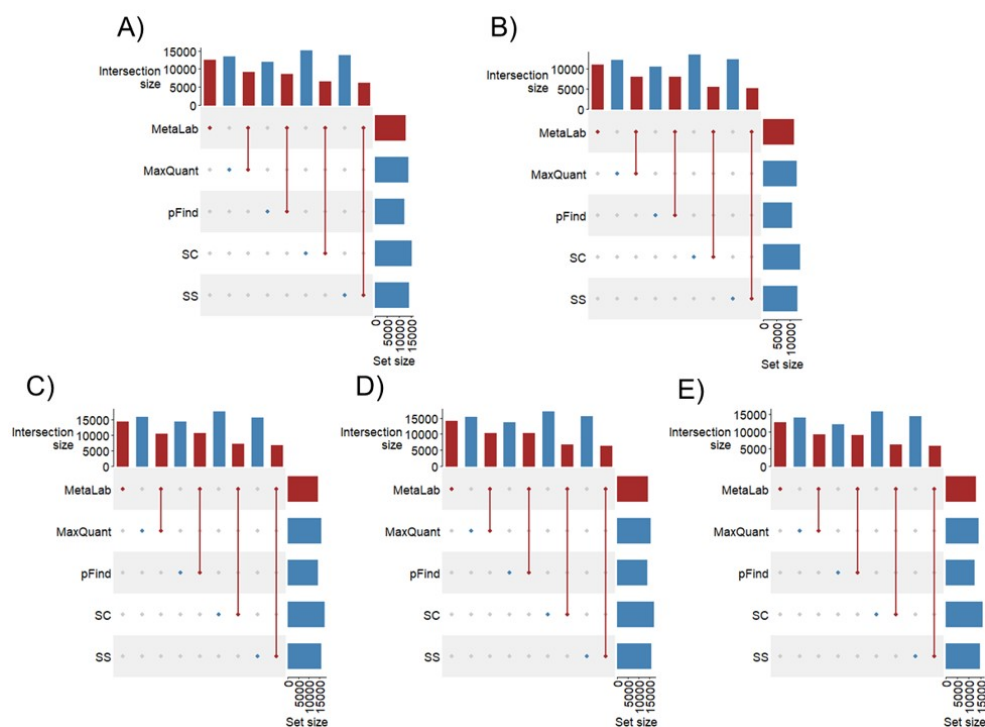


Supplementary Figure S3.2. The number of A) identified spectra and B) stripped peptide sequences from the MetaLab MAG search of each project. The cumulative number of identified non-redundant spectra and stripped peptides from all projects was shown with the orange lines.

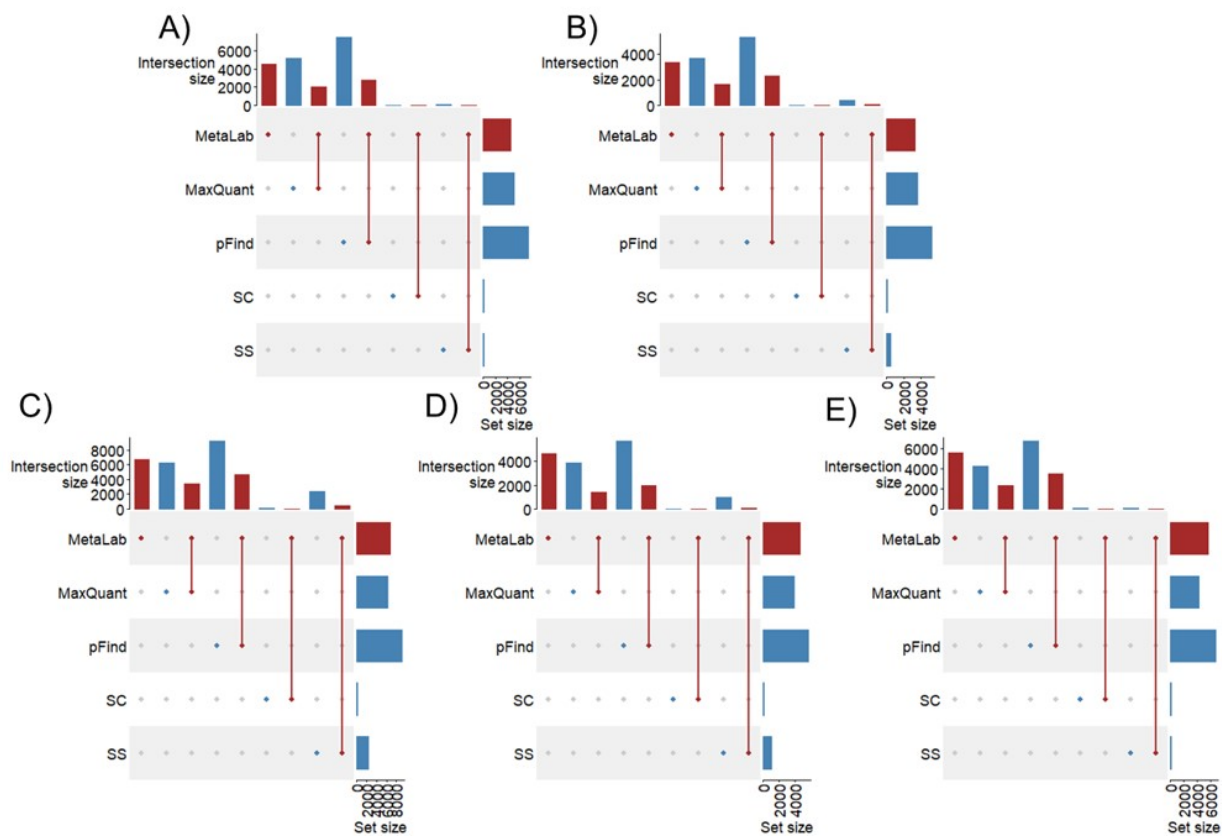
3.7. SUPPLEMENTARY MATERIALS



Supplementary Figure S3.3. Characteristics of the collected PSMs in the peptide library. A) Distribution of the pFind raw score of collected PSMs. B) Distribution of the MW of the collected precursors. C) Distribution of the charge state of the collected precursors. D) Distribution of the peptide length of the collected peptide sequences.

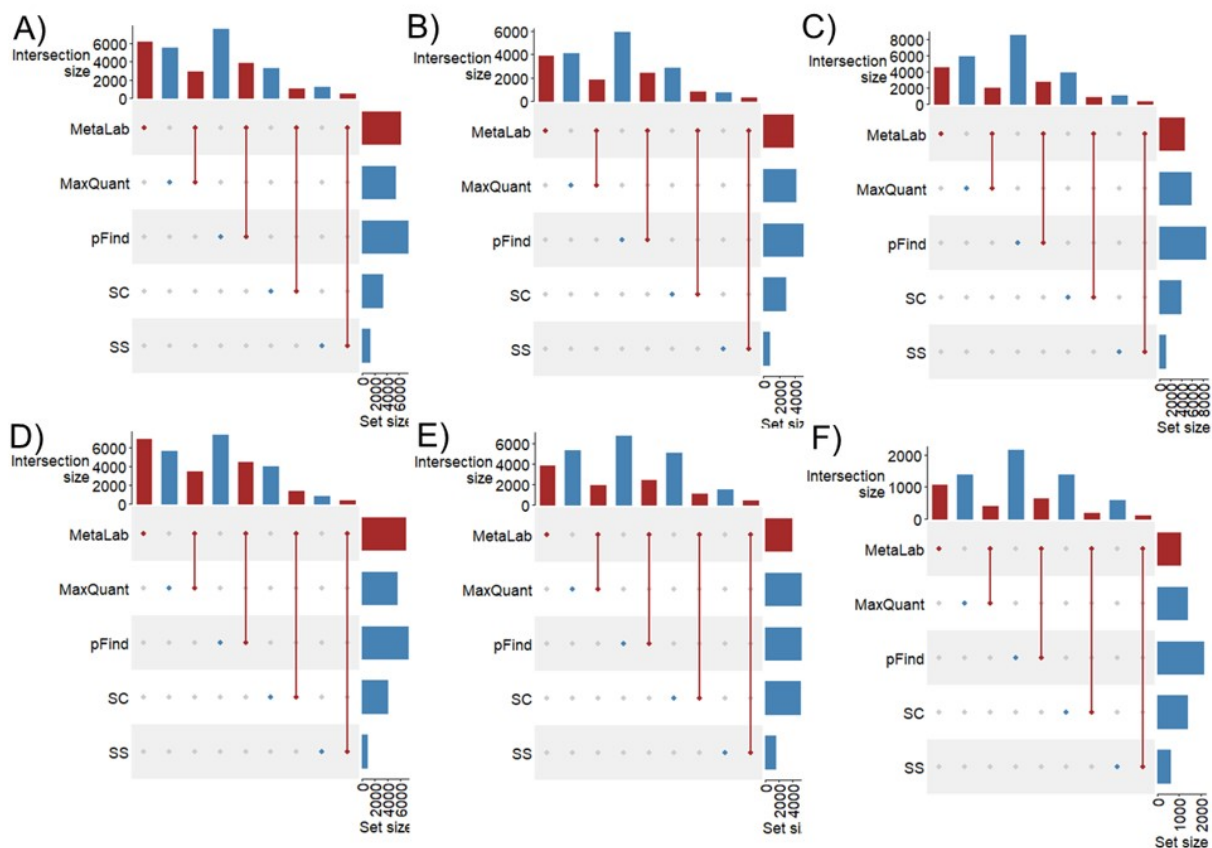


Supplementary Figure S3.4. The number of identified peptides and the overlap of identified peptides between the MetaLab MAG search and searching MetaPep with different search methods for raw files from PXD033624 dataset. Raw file names: A) Diclofenac2, B) Diclofenac3, C) DMSO1, D) DMSO2, and E) DMSO3. Red bar: Identified peptides from MetaLab MAG search. Blue bar: Identified peptides from searching MetaPep with different search methods. Grey bar: The overlap of identified peptides between MetaLab MAG search and searching MetaPep.



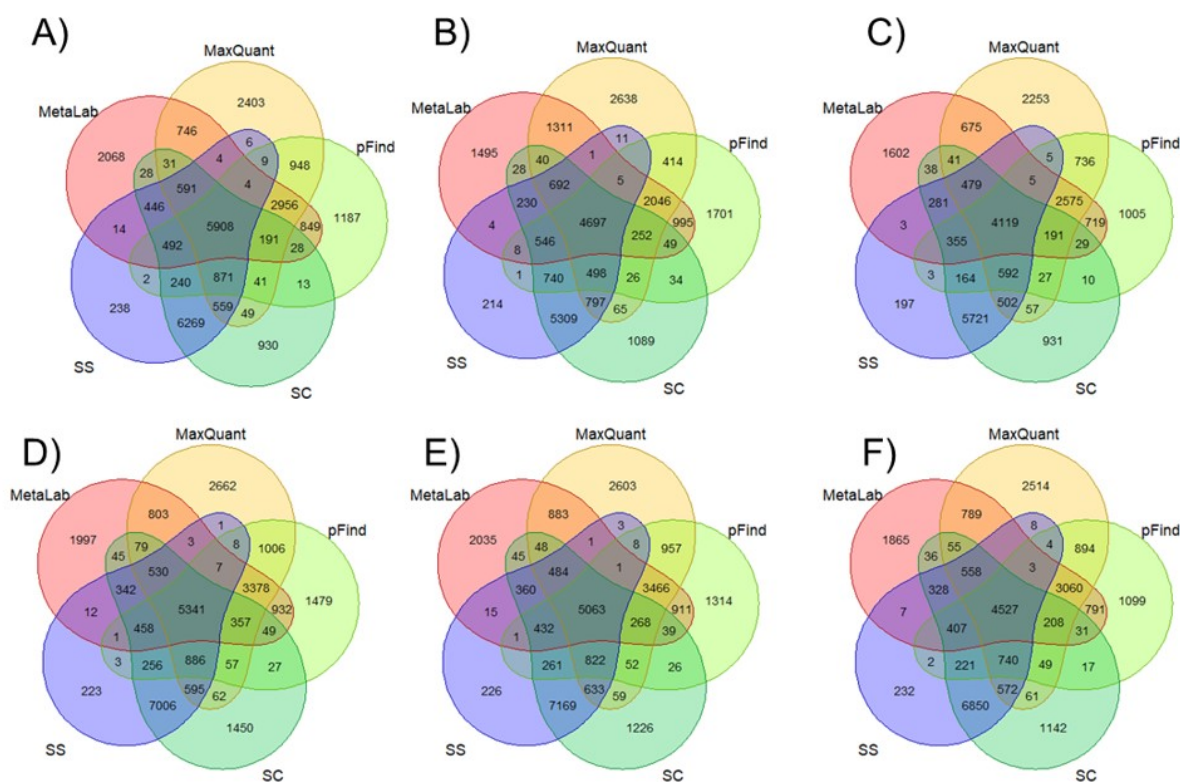
Supplementary Figure S3.5. The number of identified peptides and the overlap of identified peptides between the MetaLab MAG search and searching MetaPep with different search methods for raw files from PXD017467 dataset. Raw file names: A) S109, B) S121, C) S157, D) S161, and E) S175. Red bar: Identified peptides from MetaLab MAG search. Blue bar: Identified peptides from searching MetaPep with different search methods. Grey bar: The overlap of identified peptides between MetaLab MAG search and searching MetaPep.

3.7. SUPPLEMENTARY MATERIALS



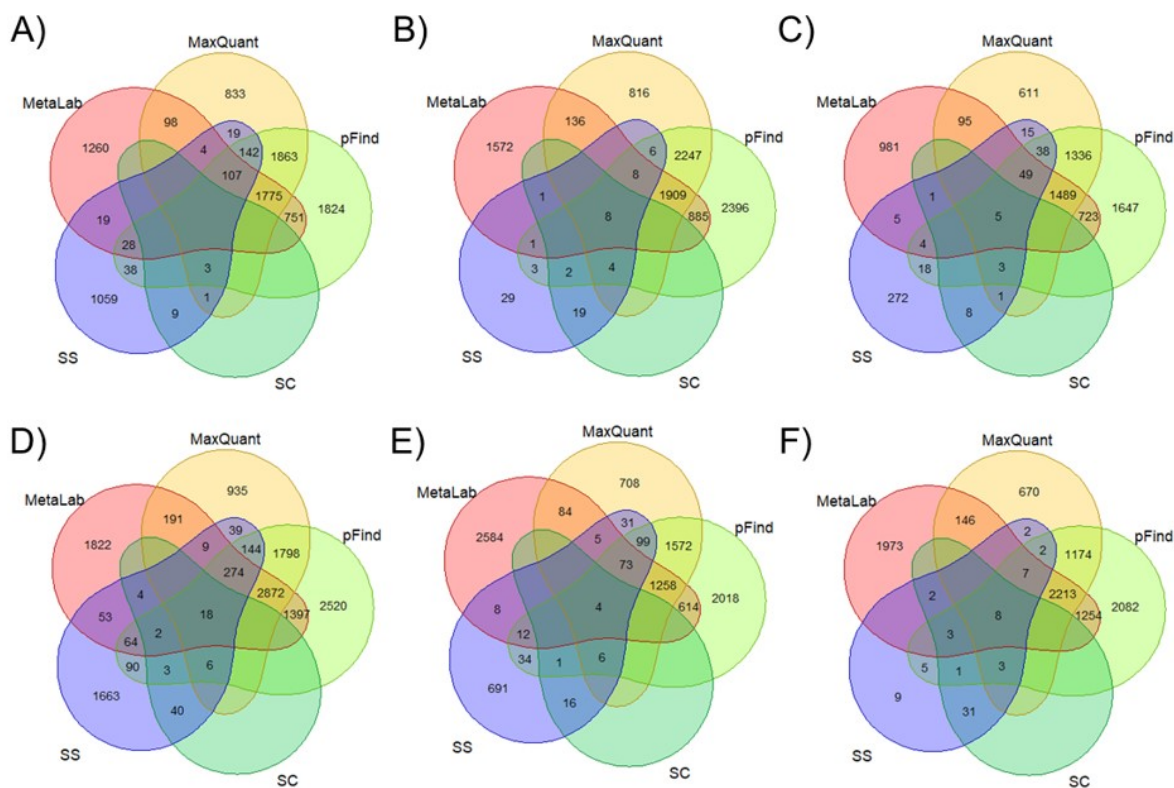
Supplementary Figure S3.6. The number of identified peptides and the overlap of identified peptides between the MetaLab MAG search and searching MetaPep with different search methods for raw files from PXD032997 dataset. Raw file names: A) 201905_Pat_VIGR_VIGR-118, B) 201905_Pat_VIGR_VIGR-088, C) 201905_Pat_VIGR_VIGR-089, D) 201905_Pat_VIGR_VIGR-120, E) 201905_Pat_VIGR_VIGR-385, and F) 201905_Pat_VIGR_VIGR-386. Red bar: Identified peptides from MetaLab MAG search. Blue bar: Identified peptides from searching MetaPep with different search methods. Grey bar: The overlap of identified peptides between MetaLab MAG search and searching MetaPep.

3.7. SUPPLEMENTARY MATERIALS



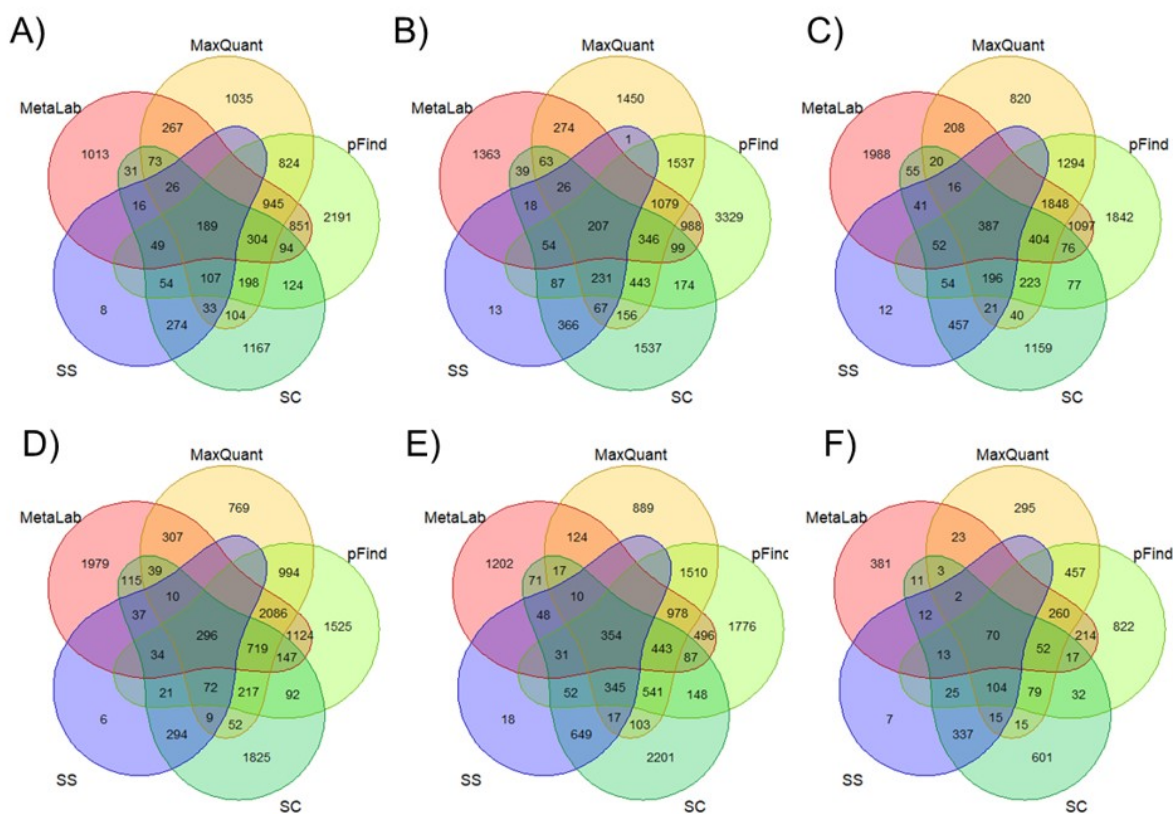
Supplementary Figure S3.7. Venn Plot of the peptides identified from the raw files of the PXD033624 dataset using MetaLab MAG and searching MetaPep with different search methods. Raw file names: A) Diclofenac1, B) Diclofenac2, C) Diclofenac3, D) DMSO1, E) DMSO2, and F) DMSO3.

3.7. SUPPLEMENTARY MATERIALS

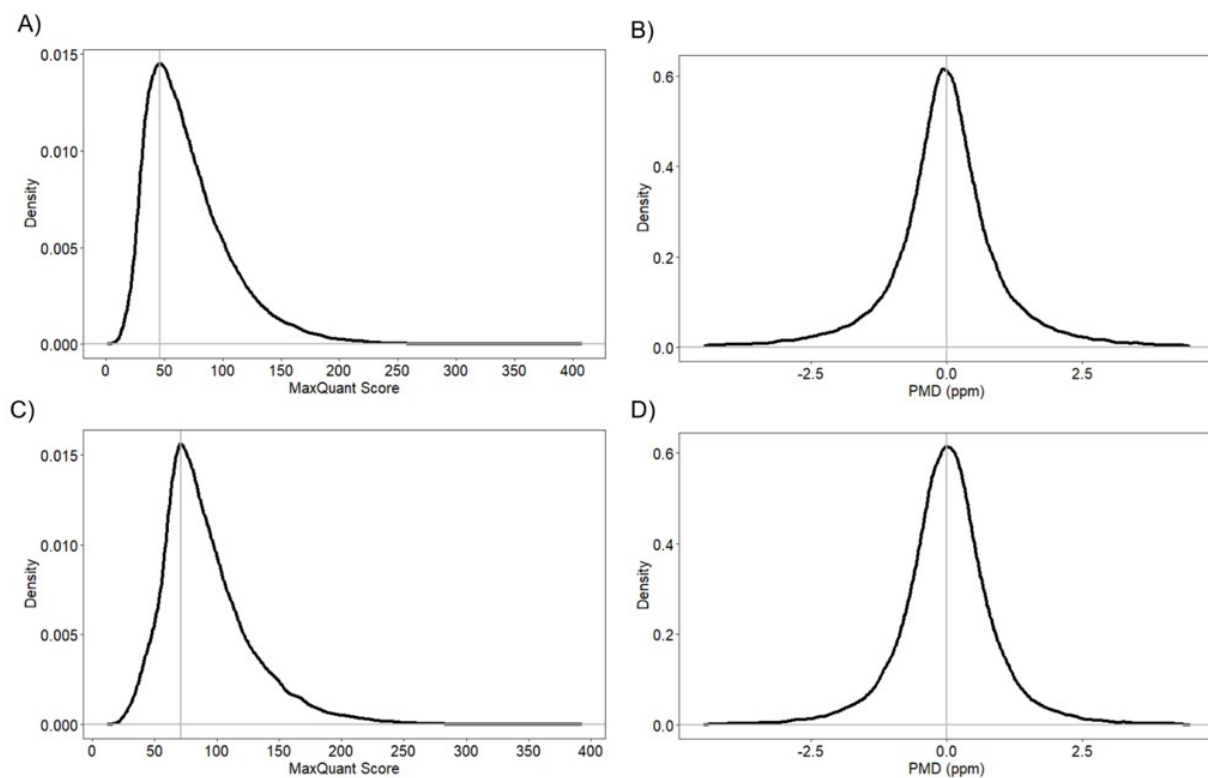


Supplementary Figure S3.8. Venn Plot of the peptides identified from raw files of PXD017467 dataset using MetaLab MAG and searching MetaPep with different search methods. Raw file names: A) S183, B) S109, C) S121, D) S157, E) S161, and F) S175.

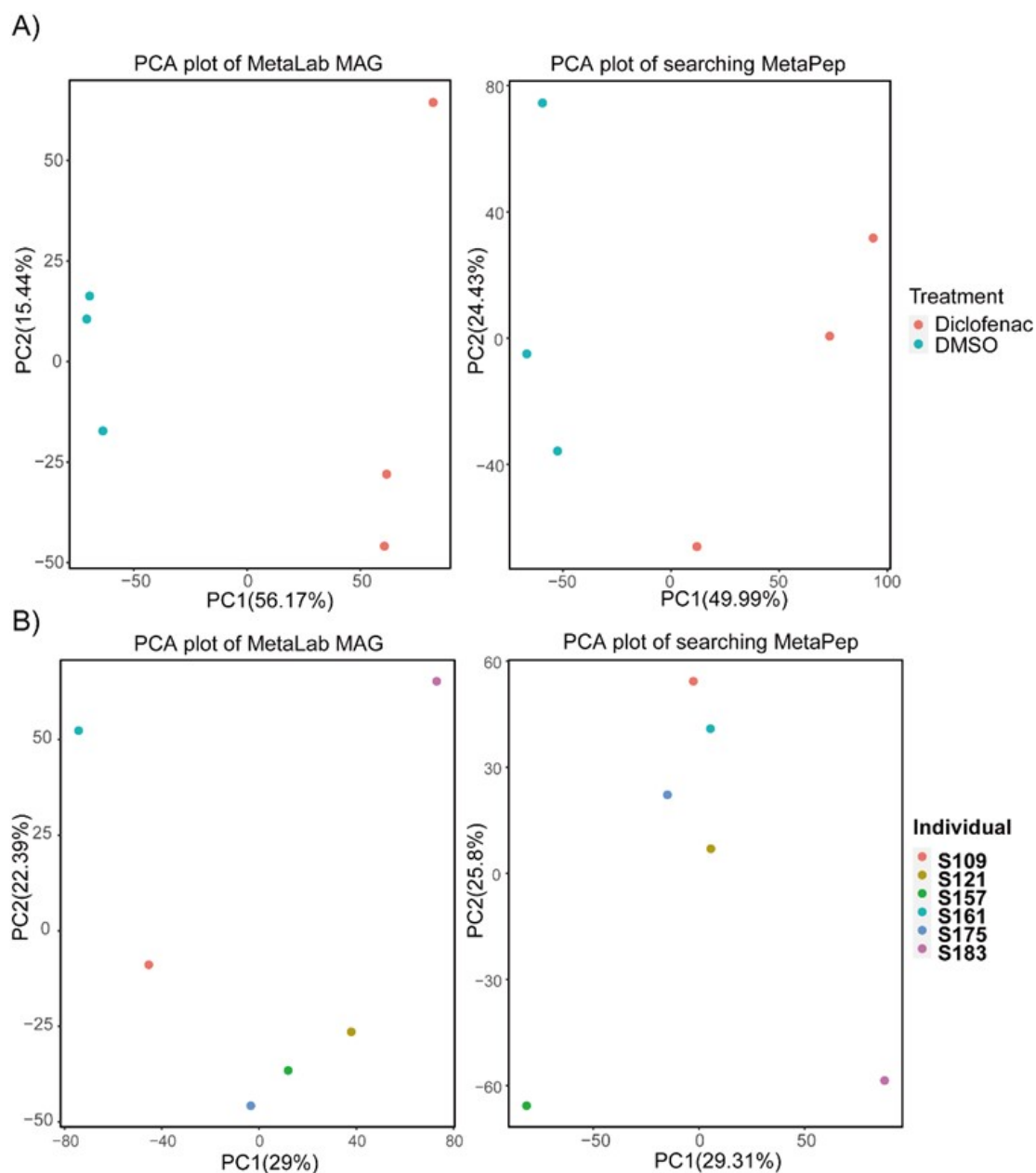
3.7. SUPPLEMENTARY MATERIALS



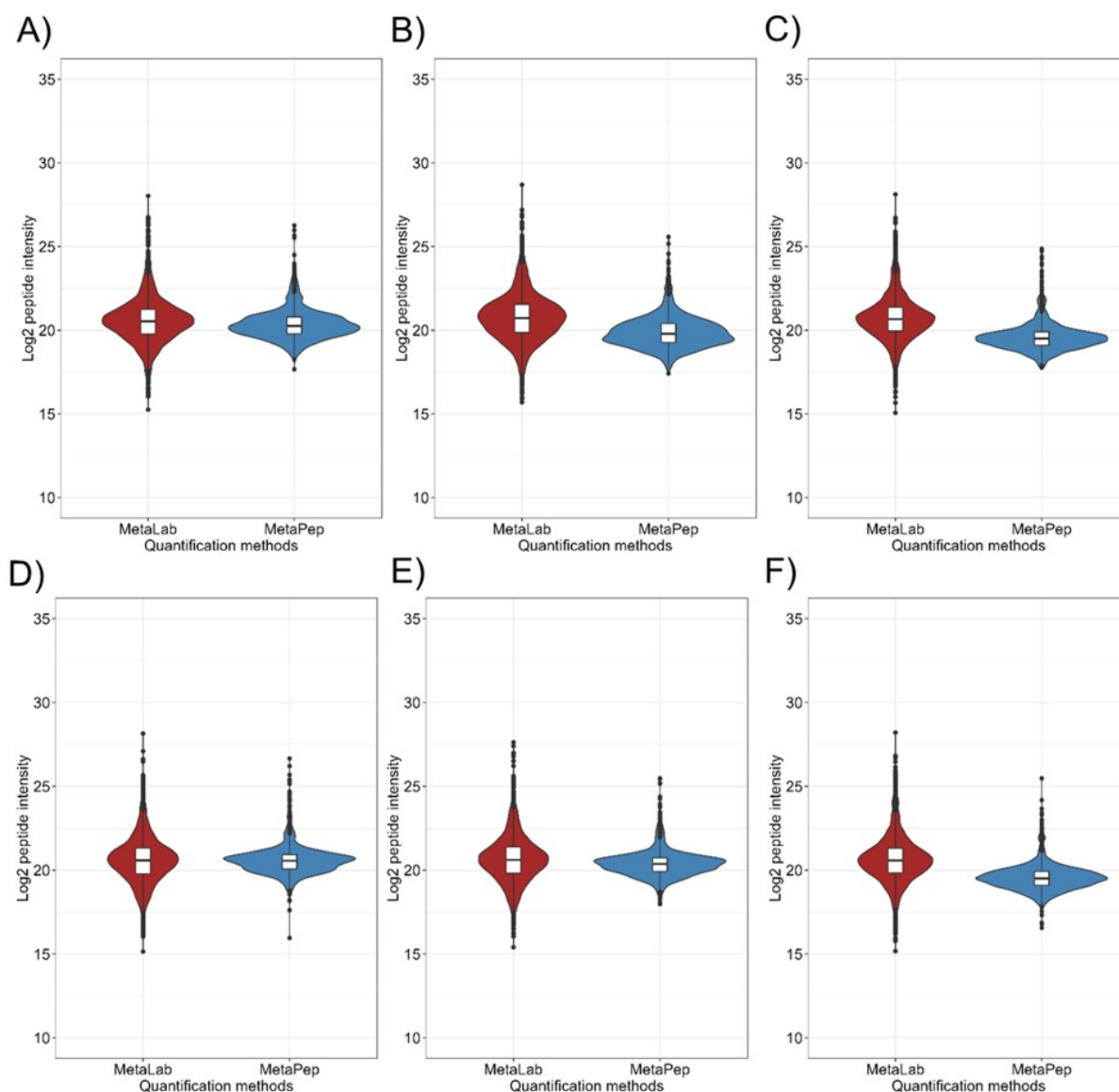
Supplementary Figure S3.9. Venn Plot of the peptides identified from raw files of PXD032997 dataset using MetaLab MAG and searching MetaPep with different search methods. Raw file names: A) 201905_Pat_VIGR_VIGR-118, B) 201905_Pat_VIGR_VIGR-088, C) 201905_Pat_VIGR_VIGR-089, D) 201905_Pat_VIGR_VIGR-120, E) 201905_Pat_VIGR_VIGR-385, and F) 201905_Pat_VIGR_VIGR-386.



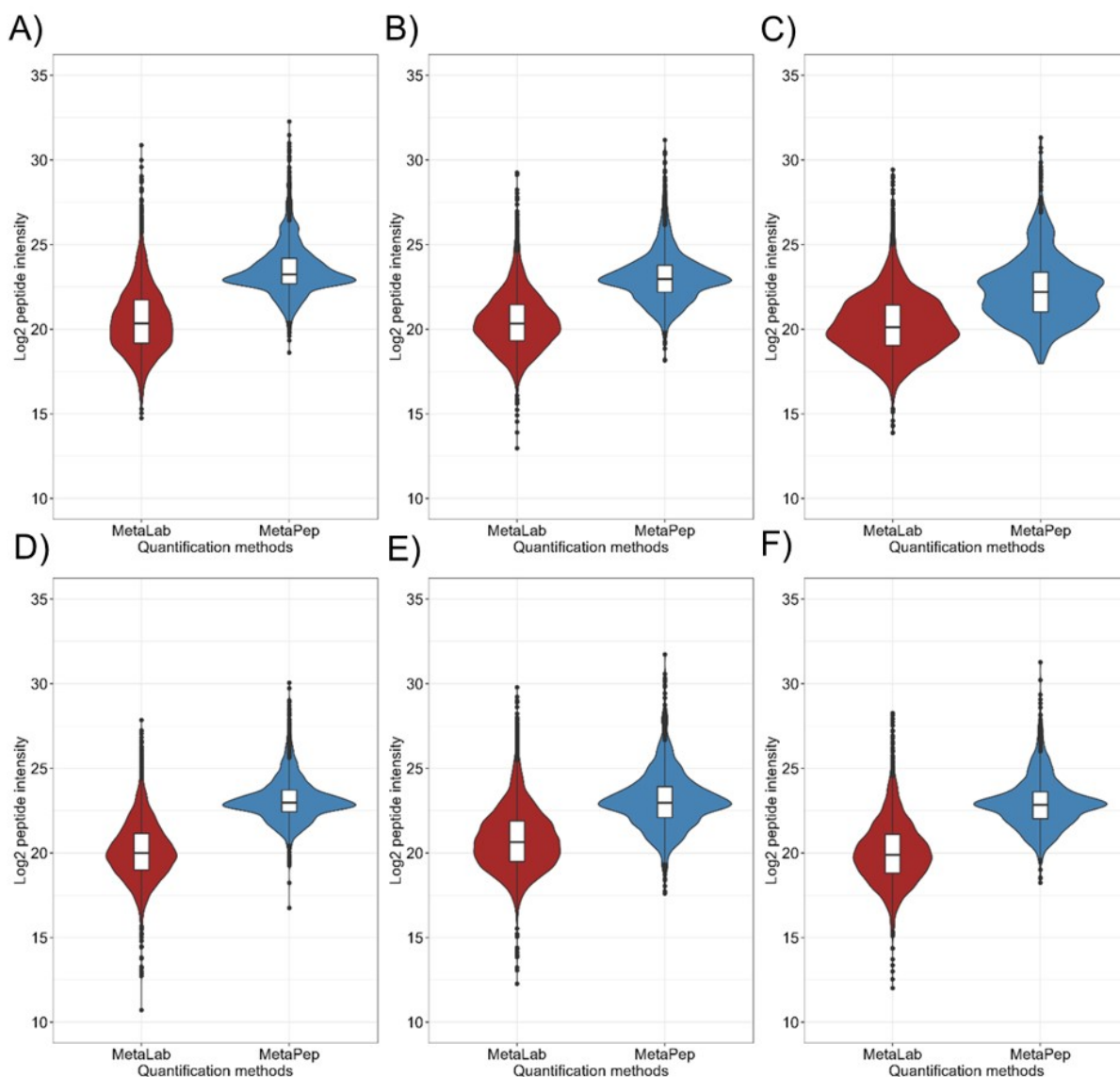
Supplementary Figure S3.10. The distribution of MaxQuant score and PMD (precursor mass discrepancy) of identified PSMs by searching MetaPep. A)-B) MaxQuant score and PMD from raw files of PXD033624 dataset. C)-D) MaxQuant score and PMD from raw files of PXD017467 dataset.



Supplementary Figure S3.11. The PCA plot of the identification and quantification results from MetaLab MAG analysis and searching MetaPep in two different datasets. A) PXD033624 dataset. The PCA plot of the quantification results of identified peptides in six raw files under two different treatments by MetaLab MAG analysis and searching MetaPep with pFind. B) PXD017467 dataset. The PCA plot of the quantification results of identified peptides in six raw files from six different individuals by MetaLab MAG analysis and searching MetaPep with pFind.

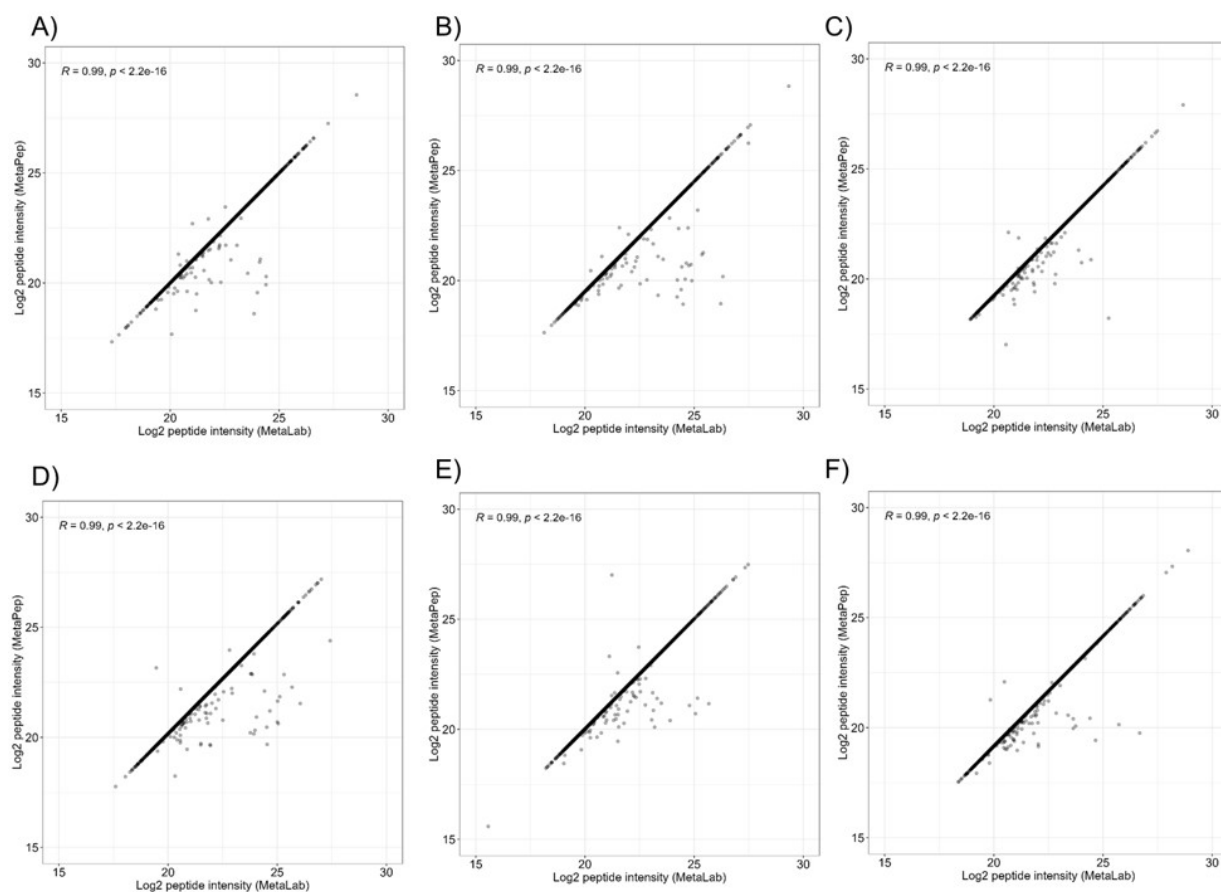


Supplementary Figure S3.12. Log₂ intensity of peptides that were uniquely quantified by either MetaLab MAG or MetaPep for raw files in PXD033624 dataset. Raw file names: A) Diclofenac1, B) Diclofenac2, C) Diclofenac3, D) DMSO1, E) DMSO2, and F) DMSO3.



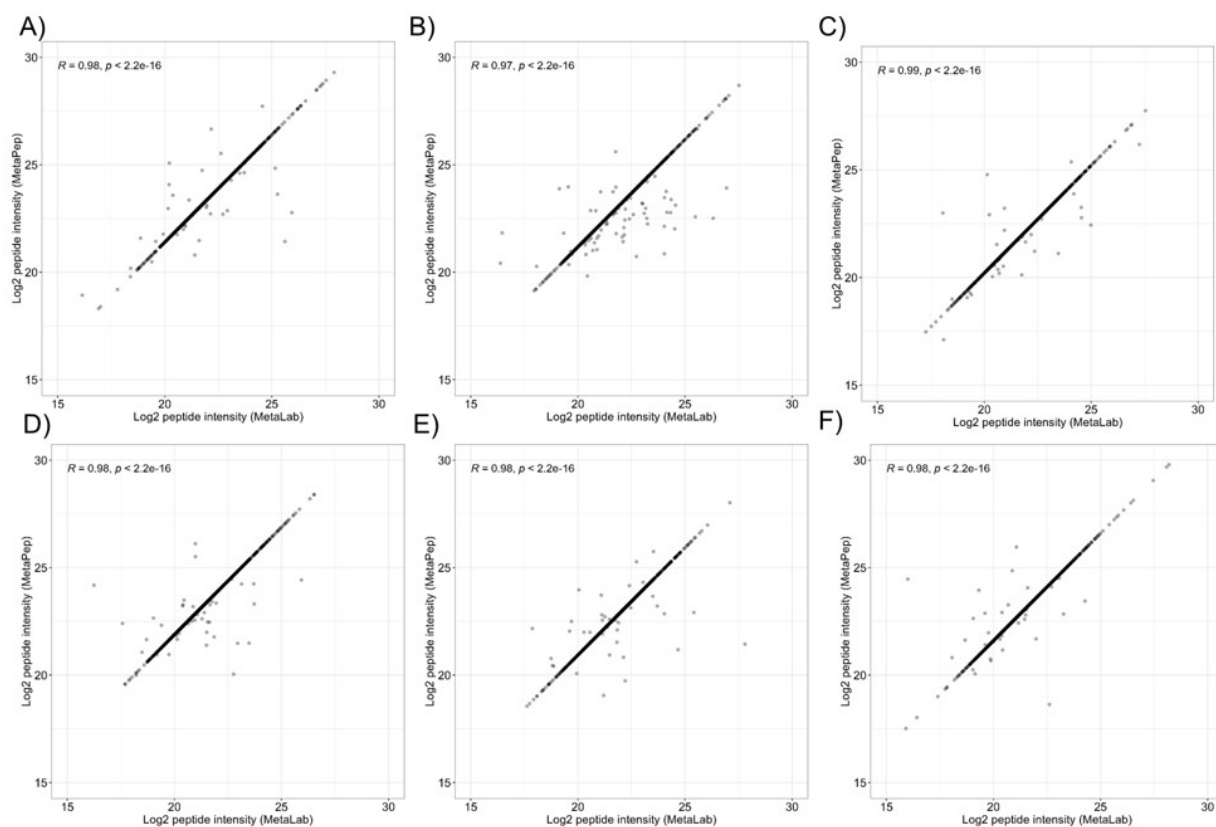
Supplementary Figure S3.13. Log₂ intensity of peptides that were uniquely quantified by either MetaLab MAG or MetaPep for raw files in PXD017467 dataset. Raw file names: A) S183, B) S109, C) S121, D) S157, E) S161, and F) S175.

3.7. SUPPLEMENTARY MATERIALS

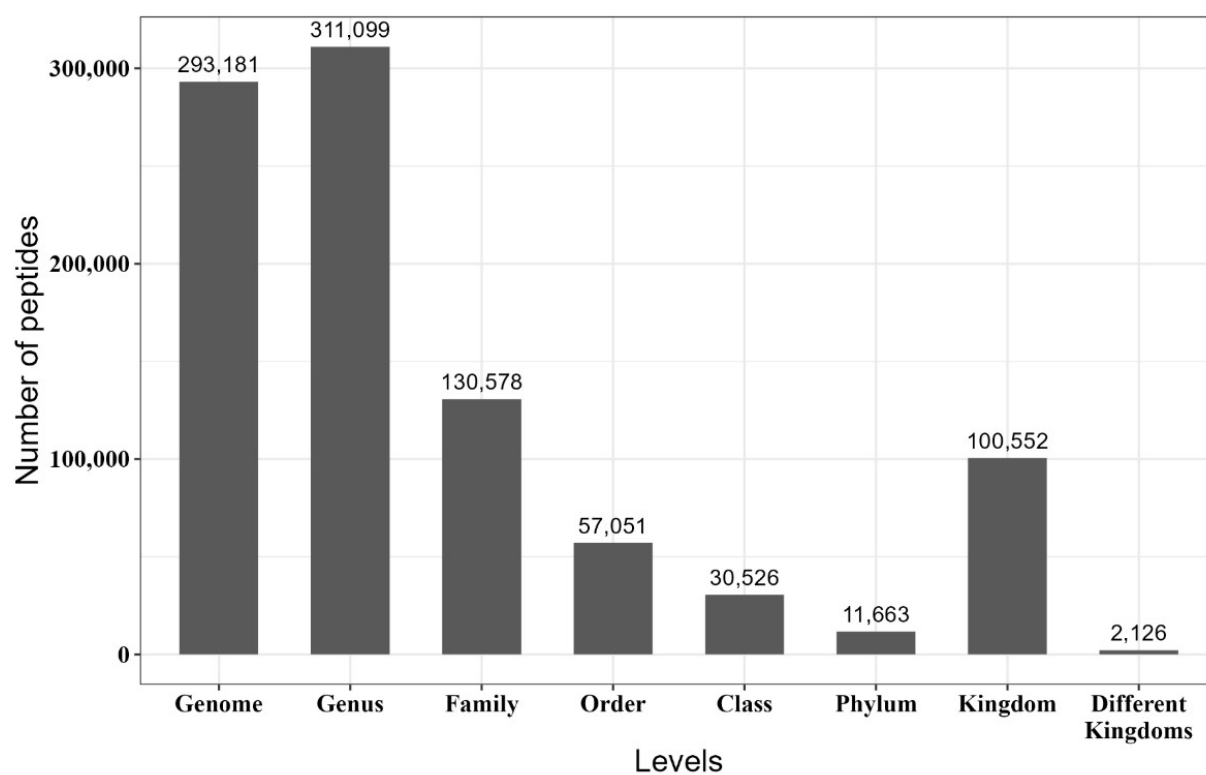


Supplementary Figure S3.14. The quantification correlation of commonly quantified peptides from both MetaLab MAG and searching MetaPep for raw files in PXD033624 dataset. Raw file names: A) Diclofenac1, B) Diclofenac2, C) Diclofenac3, D) DMSO1, E) DMSO2, and F) DMSO3.

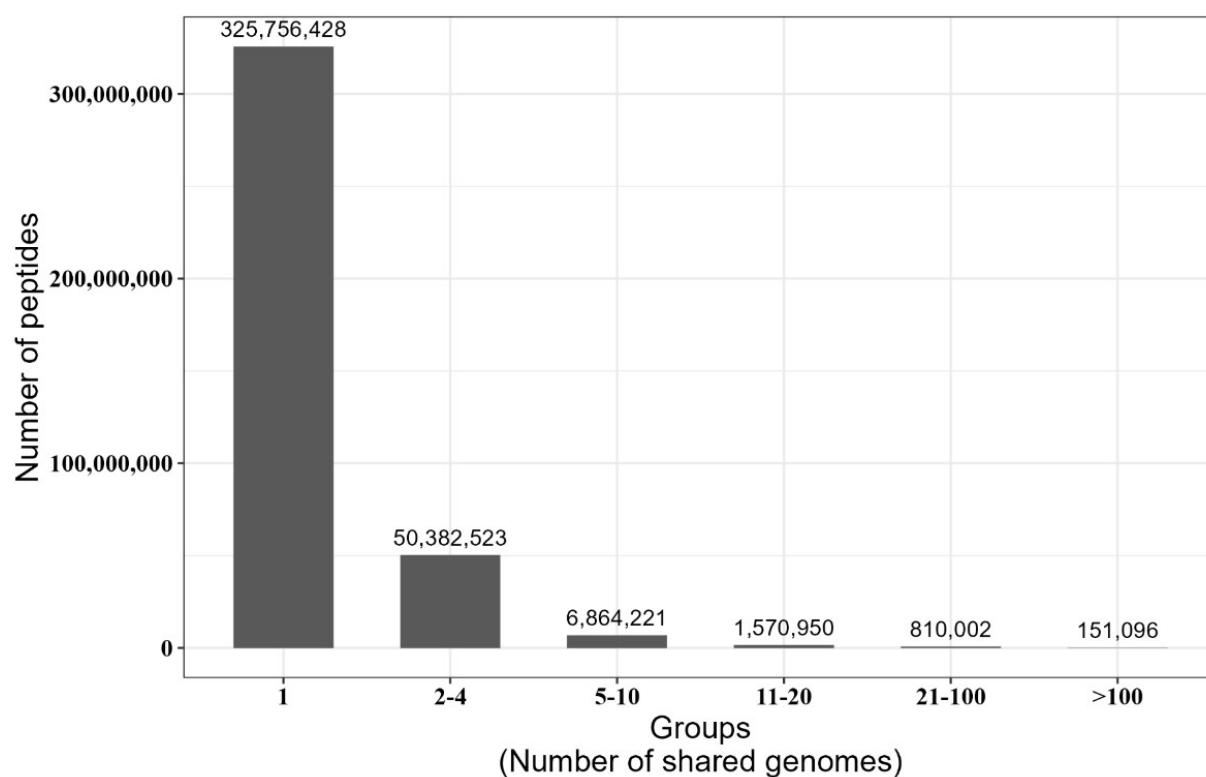
3.7. SUPPLEMENTARY MATERIALS



Supplementary Figure S3.15. The quantification correlation of commonly quantified peptides from both MetaLab MAG and searching MetaPep for raw files in PXD017467 dataset. Raw file names: A) S183, B) S109, C) S121, D) S157, E) S161, and F) S175.

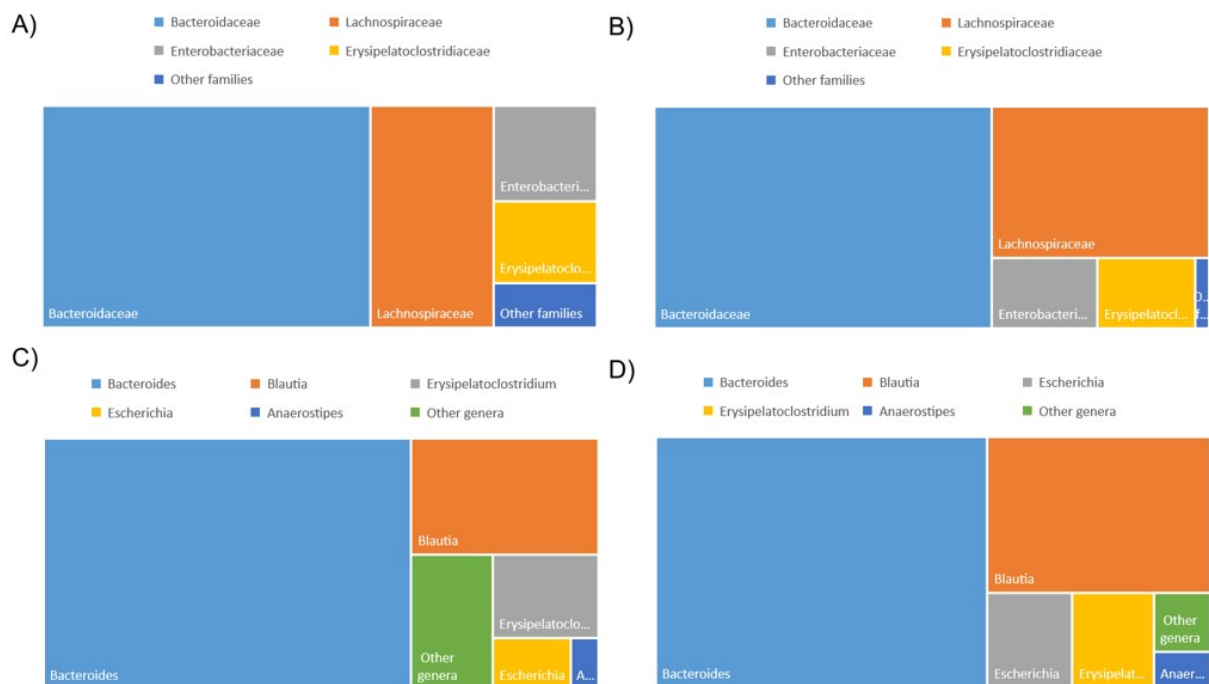


Supplementary Figure S3.16. The number of MetaPep peptides assigned to different taxonomy levels.



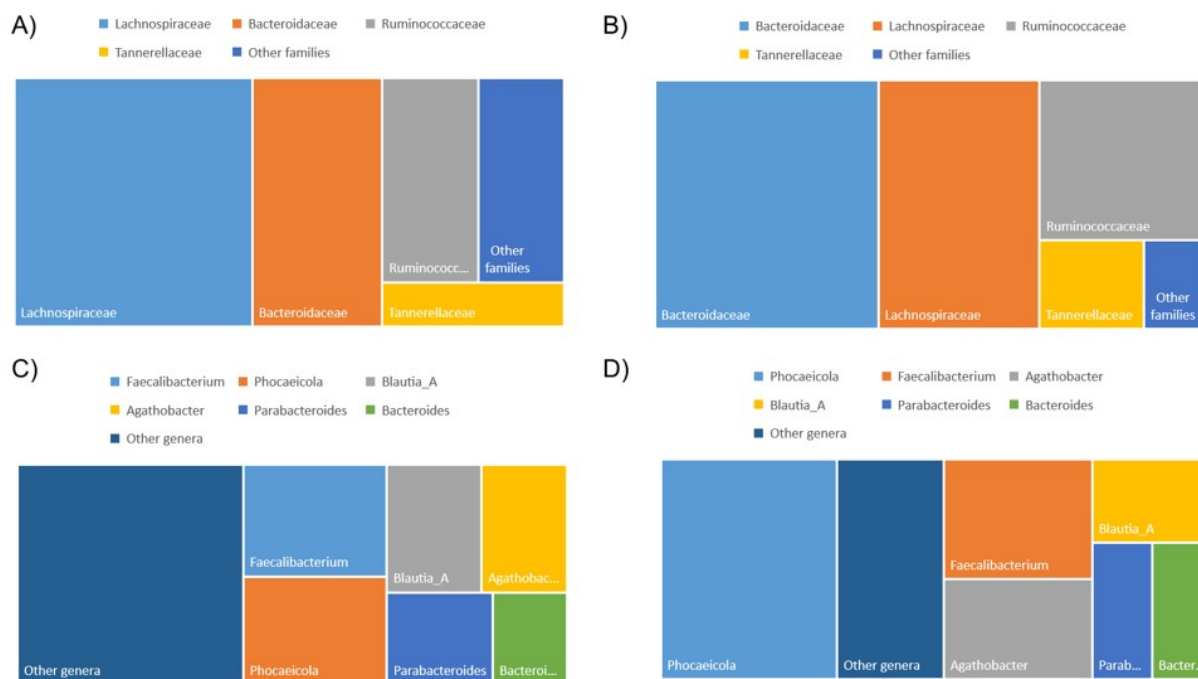
Supplementary Figure S3.17. The UHGP in-silico digested peptides shared by different numbers of UHGP representative genomes.

3.7. SUPPLEMENTARY MATERIALS



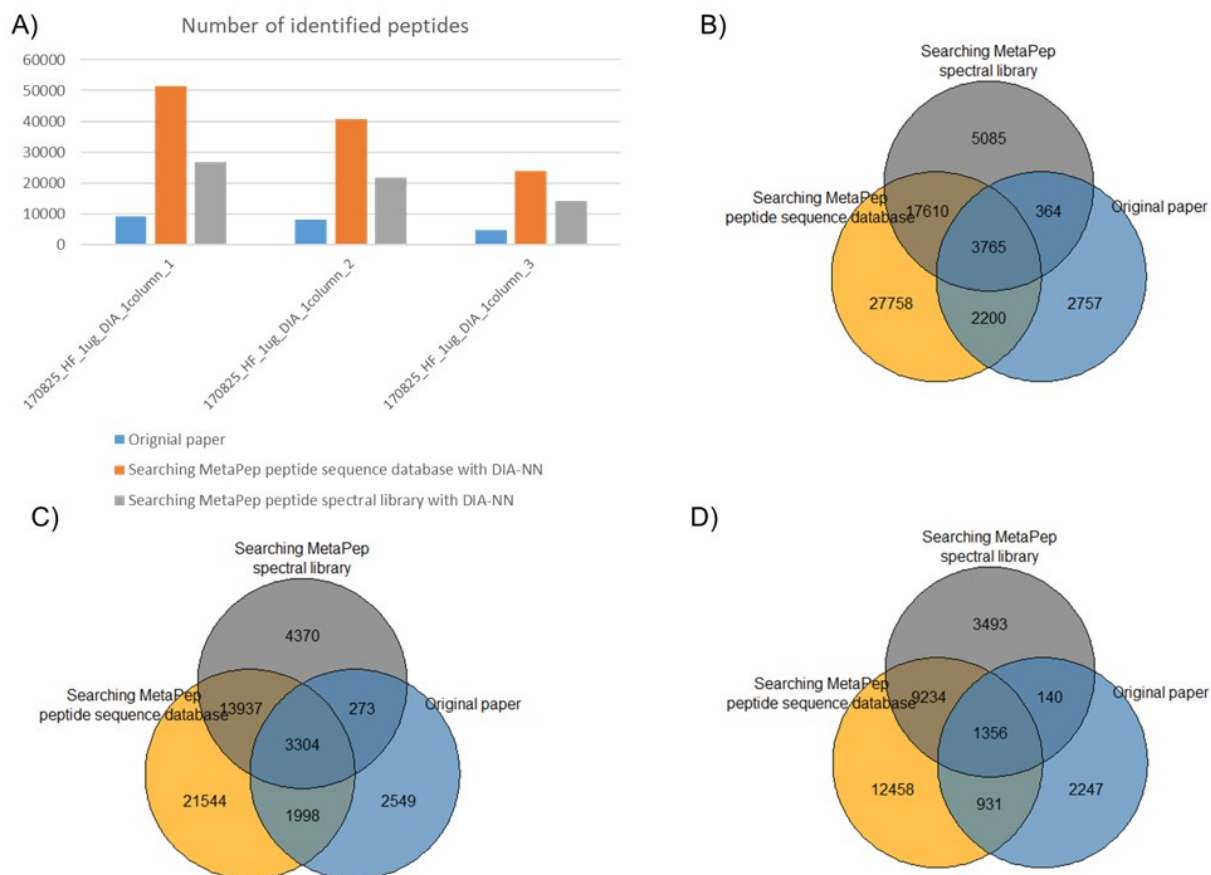
Supplementary Figure S3.18. Genus- and family-level taxonomic profiling of a SIHUMIx raw file from searching MetaPep and MetaLab MAG. The taxonomic profiling from searching MetaPep was based on the number of identified taxonomically annotated peptides from different taxonomic units. The taxonomic profiling from MetaLab MAG was acquired from taxa quantification results from the MetaLab MAG pipeline. A) Family level taxonomic profiling from searching MetaPep, B) Family level taxonomic profiling from MetaLab MAG, C) Genus level taxonomic profiling from searching MetaPep, and D) Genus level taxonomic profiling from MetaLab MAG.

3.7. SUPPLEMENTARY MATERIALS

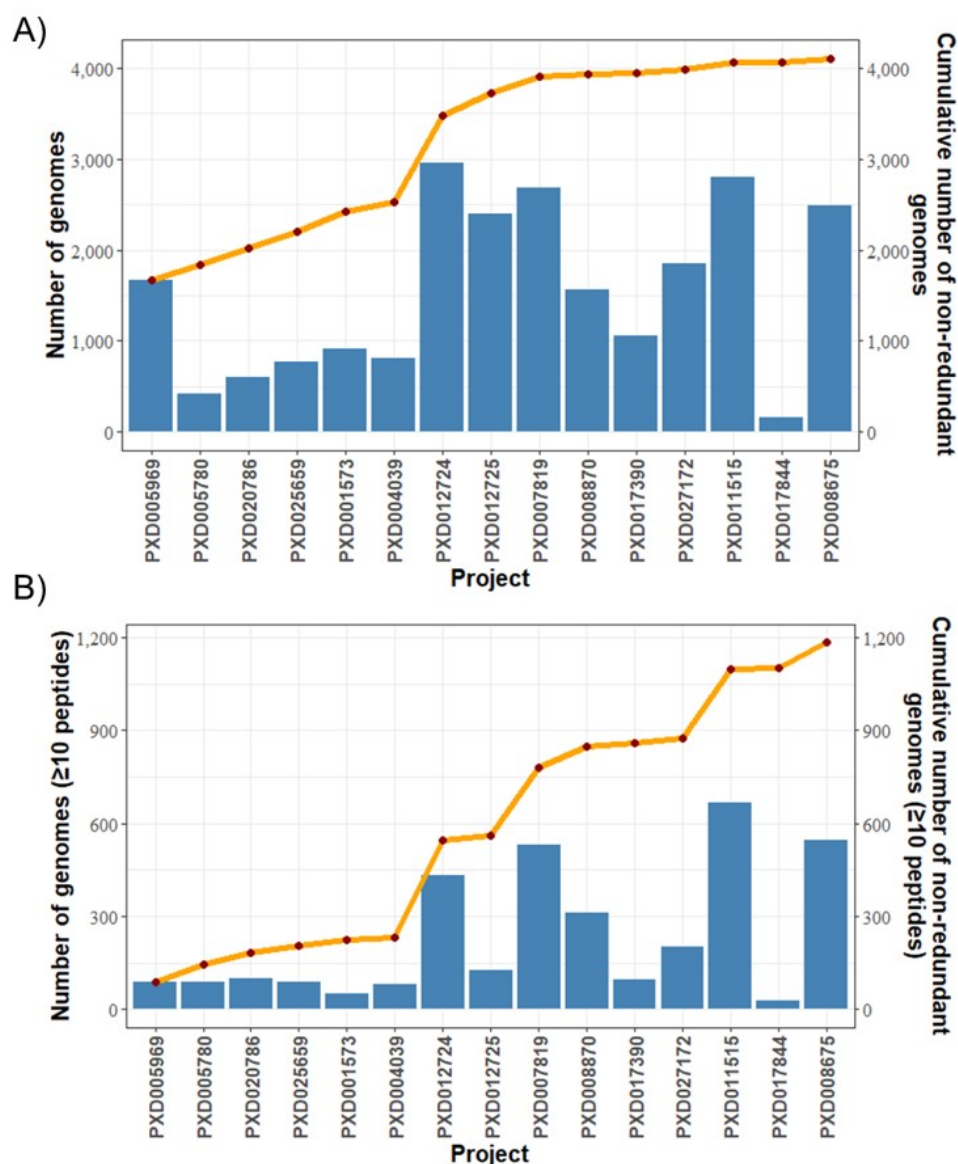


Supplementary Figure S3.19. Genus- and family-level taxonomic profiling of a real human gut metaproteomics raw file from searching MetaPep and MetaLab MAG. The taxonomic profiling from searching MetaPep was based on the number of identified taxonomically annotated peptides from different taxonomic units. The taxonomic profiling from MetaLab MAG was acquired from taxa quantification results from the MetaLab MAG pipeline. A) Family level taxonomic profiling from searching MetaPep, B) Family level taxonomic profiling from MetaLab MAG, C) Genus level taxonomic profiling from searching MetaPep, and D) Genus level taxonomic profiling from MetaLab MAG.

3.7. SUPPLEMENTARY MATERIALS



Supplementary Figure S3.20. The comparison of identified peptides from raw files of the added Q-Exactive type DIA dataset (PXD008738) by searching MetaPep to the identified peptides from the original study. A) The number of identified peptides from three raw files by searching MetaPep sequence database with DIA-NN, searching MetaPep spectral library with DIA-NN and the number of identified peptides from the original study. B)-D) The overlap of identified peptides for raw file B) 170825_HF_1ug_DIA_1column_1, C) 170825_HF_1ug_DIA_1column_2, and D) 170825_HF_1ug_DIA_1column_3.



Supplementary Figure S3.21. The number of UHGP representative bacterial genomes identified from each project. A representative genome was identified based on the correspondence between genome-distinct peptides and genomes under two different standards: A) identifying one genome-distinct peptide in a project as the standard for identifying the corresponding genome, and B) identifying ten genome-distinct peptides in a project as the standard for identifying the corresponding genome. The cumulative number of genomes identified across all projects was shown with the orange lines.

3.7.2 SUPPLEMENTARY TABLES

Supplementary Table S3.1. Detailed information on the collected human gut metaproteomics studies.

Research Article	PRIDE ID	Country	Instrument	Involved Individual(s)	Number of Raw Files
Assessing the impact of protein extraction methods for human gut metaproteomics	PXD005969	Canada	Q Exactive	1	30
Potential and active functions in the gut microbiota of a healthy human cohort	PXD005780	Italy	LTQ-Orbitrap Velos	15	15
Distinct Human Gut Microbial Taxonomic Signatures Uncovered With Different Sample Processing and Microbial Cell Disruption Methods for Metaproteomic Analysis	PXD020786	Spain	Q Exactive HF	6	6
Distinct Human Gut Microbial Taxonomic Signatures Uncovered With Different Sample Processing and Microbial Cell Disruption Methods for Metaproteomic Analysis	PXD025659	Spain	Q Exactive HF	6	18
A human gut metaproteomic dataset from stool samples pretreated or not by differential centrifugation	PXD001573	Italy	LTQ-Orbitrap Velos	1	10
The impact of sequence database choice on metaproteomic results in gut microbiota studies	PXD004039	Italy	LTQ-Orbitrap Velos	4	4
RapidAIM: a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs	PXD012724	Canada	Q Exactive	5	273
RapidAIM: a culture- and metaproteomics-based Rapid Assay of Individual Microbiome responses to drugs	PXD012725	Canada	Q Exactive	5	184
Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease	PXD007819	Canada	Q Exactive	71	211

(Continued from previous page)

Research Article	PRIDE ID	Country	Instrument	Involved Individual(s)	Number of Raw Files
Intestinal Metaproteomics Reveals Host–Microbiota Interactions in Subjects at Risk for Type	PXD008870	USA	Q Exactive	101	101
Metaproteomics Reveals Growth Phase–Dependent Responses of an In Vitro Gut Microbiota to Metformin	PXD017390	Canada	Q Exactive HF-X	1	84
Examining the Effects of an Anti-Salmonella Bacteriophage Preparation, BAFASAL(R), on Ex-Vivo Human Gut Microbiome Composition and Function Using a Multi-Omics Approach	PXD027172	Canada	Q Exactive	5	122
Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae	PXD011515	Germany	Q Exactive HF	56	424
Butyrate Levels in the Transition from an Infant- to an Adult-Like Gut Microbiota Correlate with Bacterial Networks Associated with Eubacterium Rectale and Ruminococcus Gnavus	PXD017844	Norway and Sweden	Q Exactive	6	11
Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases	PXD008675	USA	Q Exactive Orbitrap	132	641
Total	–	–	–	415	2,134

3.7. SUPPLEMENTARY MATERIALS

Table S3.2: The number of identified peptides and ID rates of the MetaLab search for collected studies.

PRIDE ID	Number of Raw Files	Average ID Rate	Number of Identified Peptides
PXD005969	30	68.0%	62,956
PXD005780	15	46.5%	53,533
PXD020786	6	32.3%	62,112
PXD025659	18	34.8%	83,429
PXD001573	10	37.7%	26,636
PXD004039	4	51.0%	26,546
PXD012724	273	65.2%	310,285
PXD012725	184	56.6%	111,346
PXD007819	211	75.0%	612,061
PXD008870	101	45.1%	280,255
PXD017390	84	64.2%	71,045
PXD027172	122	72.1%	203,260
PXD011515	424	52.2%	888,139
PXD017844	11	35.8%	26,333
PXD008675	641	29.5%	596,102

Table S3.3: Taxonomical annotation of MetaPep peptides. The table could be accessed by: <https://doi.org/10.1016/j.csbj.2023.08.025>

3.7.3 SUPPLEMENTARY METHODS

DIA-PASEF raw file acquisition

For the in-house DIA-PASEF human gut metaproteomics raw file, we followed the protein extraction, protein digestion, and peptide clean-up protocol of a human stool sample as described in our previous study (Zhang et al., 2016). The acquisition of this DIA-PASEF raw file was as described by Gomez-Varela et al. (Gómez-Varela et al., 2023a), with minor

3.7. SUPPLEMENTARY MATERIALS

modifications as follows.

Nanoflow reversed-phase liquid chromatography (Nano-RPLC) was performed on a NanoElute 2 system (Bruker Daltonik, Bremen, Germany) coupled with a timsTOF Pro 2 mass spectrometer (Bruker Daltonik), using a CaptiveSpray ion source (Bruker Daltonics). Peptides were separated with a 120 minutes gradient on a PepSep Fifteen Column (15 cm x 75 μ m column packed with 1.9 μ m C18 particles) (Bruker Daltonik). Mobile solvent A consisted of 0.1% FA and mobile phase B of 100% acetonitrile, 0.1% FA. The flow rate was set to 500 nL/min for the whole gradient. In the first 115 minutes separation, the mobile phase B was linearly increased from 3 to 35%, flowed by a steep increase to 95% in 3.5 minutes. Then a flow rate of 500 nL/min at 95% was maintained for 6.5 minutes to elute all hydrophobic peptides. The timsTOF Pro 2, Bruker Daltonik, Bremen, Germany) was performed in data-independent acquisition (DIA) coupled with parallel accumulation serial fragmentation (PASEF) mode. MS scan was performed between 100 and 1700 m/z. The TIMS analyzer was operated in a 100% duty cycle with equal accumulation and ramp times of 100 ms each, and the ion mobility range was set to 0.60–1.60 V·s/cm². For DIA-PASEF MS/MS scan, precursors with m/z between 400 and 1200 were defined in 16 scans containing 32 ion mobility steps with an isolation window of 26 Da in each step with 1 Da overlap with neighbouring windows. The acquisition time of each DIA-PASEF scan was set to 100 ms, which led to a total cycle time estimate 1.8 sec. The collision energy for DIA-PASEF scan was ramped linearly from 59 eV at $1/k_0 = 1.6$ V·s/cm² to 20 eV at $1/k_0 = 0.6$ V·s/cm².

PEPTIDE ABUNDANCE CORRELATIONS IN
METAPROTEOMICS ENHANCE TAXONOMIC
AND FUNCTIONAL ANALYSIS OF THE
HUMAN GUT MICROBIOME

Preface

The following chapter consists of a research article of which I am the first author. The article has been deposited in *bioRxiv* and is currently under review for publication in a peer-reviewed journal.

Zhongzhi Sun, Zhibin Ning, Qing Wu, Leyuan Li, Andrew C. Doxey, and Daniel Figeys. Peptide abundance correlations in metaproteomics enhance taxonomic and functional analysis of the human gut microbiome. *bioRxiv*, 2025

Author contributions are described using the CRediT system as listed below.

Zhongzhi Sun: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft. **Zhibin Ning:** Conceptualization, Methodology, Investigation, Writing – original draft. **Qing Wu:** Methodology, Investigation, Writing – review & editing. **Leyuan Li:** Resources, Writing – review & editing. **Andrew C. Doxey:** Methodology, Supervision, Writing – review & editing. **Daniel Figeys:** Conceptualization, Funding acquisition, Supervision, Writing – original draft.

4.1 ABSTRACT

Mass spectrometry (MS)-based proteomics is widely used for quantitative protein profiling and has become a powerful tool for studying protein interactions. However, most current research focuses on single-species proteomics to study protein interactions. Protein interactions within more complex microbiomes, composed of 100's of bacterial species, remain largely unexplored. The human gut microbiome, closely linked to human health and disease, has become a key area of study using metaproteomics. Yet, due to the complexity of the microbiome, the interactions between gut microbes and their host remain largely unknown. In this study, we analyzed peptide abundance correlations within a metaproteomics dataset derived from *in vitro* cultured human gut microbiomes subjected to various drug

treatments. Our analysis revealed that peptides from the same protein or taxon exhibited correlated abundance changes. By using t-SNE for visualization, we generated a peptide correlation map in which peptides from the same taxon formed distinct clusters. Furthermore, peptide abundance correlations enabled genome-level taxonomic assignments for a greater number of peptides. In single-species subsets of the dataset, peptide correlation networks constructed using taxon-based normalized peptide abundance (TNPA) linked peptides from functionally related proteins. These networks also provided insights into the potential functions of previously uncharacterized proteins. Altogether, our study demonstrates that analyzing peptide abundance correlations enhances both taxonomic and functional analyses in human gut metaproteomics research.

4.2 INTRODUCTION

Despite many studies linking the human gut microbiome with human health and disease (de Vos et al., 2022; Lloyd-Price et al., 2019; Lynch & Pedersen, 2016; Sanna et al., 2019; Wu et al., 2024), elucidating the mechanisms underlying the impact of the microbiome on host biology remains challenging, largely due to the complexity and the dark matter in gut microbial communities (Heintz-Buschart & Wilmes, 2018; Lobb et al., 2015). Different omics technologies, including metagenomics, metatranscriptomics, metaproteomics, and metabolomics, have become important tools for studying the human gut microbiome (Franzosa et al., 2015). One of these techniques, metaproteomics, measures the presence and abundance of proteins that drive biological functions within microbial communities, providing a direct glimpse of gut microbiome pathways (Kleiner, 2019).

Proteins involved in the same biological pathways tend to be co-regulated (Wang et al., 2017). These functionally related proteins typically have similar expression profiles, leading to positive abundance correlations. Conversely, proteins involved in mutual inhibition, feedback regulation, or competition for binding sites tend to display negative abundance

correlations (Brandman & Meyer, 2008; Garrido-Rodriguez et al., 2022). Given accurate quantification using mass spectrometry, abundance correlations between proteins could reveal functional linkages, even for proteins that do not physically interact or colocalize, thus enabling the prediction of unknown protein functions (Kustatscher et al., 2019). Protein abundance correlations have been applied to predict the functions of uncharacterized proteins in various model organisms, such as *Escherichia coli* (Mateus et al., 2020), *Saccharomyces cerevisiae* (Messner et al., 2023), and *Homo sapiens* (Kustatscher et al., 2019). However, these studies have been largely limited to single-species proteomics. In complex microbial communities, like the gut microbiome, protein abundance correlations remain unexplored. A previous study applied gene abundance correlation in metagenomics to predict microbial functional organization (Ma et al., 2018), but since proteins directly execute biological processes, studying protein abundance correlations is expected to more faithfully reflect functional dynamics in microbial communities. Given that over 40% of proteins from the human gut microbiome remain functionally uncharacterized (Almeida et al., 2021), and that some of these unknown proteins have been implicated in disease development (Haynes et al., 2018; Kustatscher et al., 2022), understanding their functions is critical for advancing our knowledge of host-microbiome interactions.

Studying protein abundance correlations is more challenging in metaproteomics compared to single-species proteomics due to the difficulties in protein inference (Schiebenhoefer et al., 2019). In complex microbial communities, peptides can be shared among homologous proteins from different species, making it hard to assign them to their actual proteins of origin. Our previous studies showed that a large number of identified peptides were shared by multiple species, complicating taxonomic and protein source assignments (Sun et al., 2023). Moreover, even after protein inference, it remains debatable whether it is reasonable to aggregate peptide quantities into a single protein quantity, as peptides from the same protein-coding sequence could express different quantitative responses due to the additional complexity introduced by post-translational modifications (Plubell et al.,

2022). Recently, peptide-centric metaproteomics analysis, which directly links peptides to their taxonomic and functional annotations, provides an alternative approach (Gurdeep Singh et al., 2019). Since mass spectrometry directly measures peptides instead of proteins, peptide-centric analysis is inherently reasonable and has been shown to be more sensitive and uncover features that are masked at the protein level (Simopoulos et al., 2020). Given the challenges in protein inference and the inherent advantages of peptide-centric analysis, we focus on peptide abundance correlations in this study. By leveraging peptide-centric analysis, we aim to overcome the biases associated with protein inference and provide a more accurate representation of functional and taxonomic relationships in microbial communities.

In this study, we analyzed peptide abundance correlations in a metaproteomics dataset derived from *in vitro* cultured human gut microbiomes subjected to different drug treatments, also referred to as perturbations. We explored the biological foundations of these correlations and visualized them using a peptide correlation map. Additionally, we applied peptide abundance correlations to improve the taxonomic assignment of peptides. Focusing on single-species subsets, we calculated taxon-based normalized peptide abundance (TNPA) and constructed peptide abundance correlation networks. These networks revealed functional linkages, providing new insights into the roles of previously uncharacterized microbial proteins.

4.3 MATERIALS AND METHODS

4.3.1 METAPROTEOMICS DATASET

In this study, we analyzed a metaproteomics dataset comprising 672 raw files from human gut microbiomes of six individuals subjected to various drug treatments (112 samples per individual). Specifically, stool samples were treated with 109 different compounds—107

4.3. MATERIALS AND METHODS

drugs, two DMSO samples as negative controls, and three kestose samples as positive controls (detailed in Supplementary Table S4.1)—and then *in vitro* cultured using the RapidAIM assay (Li et al., 2020), a culture- and metaproteomics-based rapid method for studying individual microbiome responses to drugs. Detailed information on sample collection, preparation, and LC-MS/MS analysis has been documented in a separate study (Li et al., 2025). Metaproteomics raw files obtained from LC-MS/MS were searched against the IGC (Integrated Gene Catalog) database (MetaHIT Consortium et al., 2014) with MetaLab2.3 (Cheng et al., 2020) with the MaxQuant (Tyanova et al., 2016) workflow, using default settings. Peptide quantification results were extracted from the “peptides.txt” file generated by MetaLab for subsequent analysis.

4.3.2 PEPTIDE ABUNDANCE CORRELATION CALCULATION

Pre-processing and peptide filtering. Peptide abundance correlations across different perturbations were calculated separately for each individual. First, peptide identification and quantification results were extracted individually. Peptides were filtered by retaining only those with non-zero intensity in at least 20% of the samples for each individual (≥ 23 samples). The quantification results of peptides that passed this filtering step were used for further analysis.

Peptide abundance log2-fold change (log2-FC) calculation. To avoid zero values, 1 was added to peptide raw intensity values. The log2-fold change in peptide abundance for each sample was then calculated by dividing the peptide intensity under each treatment by the average intensity of the two DMSO-treated control samples, followed by log2 transformation.

Spearman correlation coefficients (SCC) calculation. Peptide abundance correlations across different treatments were calculated for all peptide pairs using Spearman

4.3. MATERIALS AND METHODS

correlation coefficients (SCCs) of the log2-FC values across the treatments. The SCCs were computed using the *cor* function in R.

4.3.3 PEPTIDE ANNOTATIONS

To enable meaningful biological analysis, peptides were assigned to their protein sources and annotated with taxonomic and functional information using the following procedures.

Protein source refinement. We first generated a genome-level taxonomic profile for each individual to refine peptide protein source assignments and improve peptide annotation resolution. All identified peptides from each individual were mapped to their taxonomic sources by aligning them against MetaPep (Sun et al., 2023) records. Genome-distinct peptides, defined as peptides exclusively present in a single bacterial genome, were utilized to enhance the accuracy of assignments. In the subsequent step, only genomes with at least three genome-distinct peptides identified across all samples were considered for peptide protein source assignments.

Peptide protein sources annotation. Next, for peptide protein source assignments, protein sequences from the 4,744 UHGG representative genomes (Almeida et al., 2021) were in-silico digested using DeepDigest (Yang et al., 2021a) with the following parameters: miscleavage = 2, minimum peptide length = 7, maximum peptide length = 47 (default settings), and trypsin as the protease. Subsequently, all identified peptides were mapped to their unique protein source or multiple protein sources.

Peptide taxonomic and functional sources annotation. Each peptide was taxonomically annotated by assigning it to a specific genome source or the lowest common ancestor (LCA) of the genomes corresponding to all its protein sources. For functional annotations, protein functional annotations were extracted from the UHGG (Almeida et

4.3. MATERIALS AND METHODS

al., 2021) database, and each peptide was functionally annotated based on the functional annotation of its source protein(s).

4.3.4 VISUALIZE PEPTIDE ABUNDANCE CORRELATIONS WITH GLOBAL PEPTIDE CORRELATION MAPS

t-SNE (t-distributed stochastic neighbor embedding) was applied to visualize the peptide abundance correlations. The SCC matrix of peptide pairs from each individual was used as the input for the *Rtsne* package (Jesse & Laurens van der, 2023) with default parameters (perplexity = 30). In the peptide correlation map, each point represents a peptide, and the peptides were colored based on their taxonomic or functional annotations acquired as described in the previous section.

4.3.5 APPLYING A MACHINE LEARNING MODEL TO PREDICT PEPTIDE GENOME SOURCE

Genome-distinct peptides from the two genomes with the highest number of such peptides within the same bacteria family were extracted. Seventy percent of these peptides were randomly selected as the training dataset, while the remaining 30% constituted the test dataset. Additionally, genome-distinct peptides from other genomes within the same family and 500 randomly selected peptides from other families were included for further evaluation of the model’s performance.

For the training dataset, the Spearman correlation coefficients (SCCs) of each peptide’s abundance changes relative to all peptides in the dataset were calculated and used as input features. Specifically, this produced an SCC matrix in which each value represented the calculated SCC between a pair of peptides in the training set. A Random Forest classifier was then trained using the *randomForest* package (Leo et al., 2024) in R with `ntree = 500`,

importance = True.

During model evaluation, the SCC matrix between the peptides in the test dataset and those in the training dataset was computed and used as the input for the trained classifier. The model’s performance was assessed using confusion matrices calculated on two different combined test datasets: (1) genome-distinct peptides from the test dataset combined with genome-distinct peptides from other genomes of the same family, and (2) genome-distinct peptides from the test dataset combined with 500 randomly selected peptides from other families. Following these assessments, the trained model was used to predict genome sources of peptides from the same bacterial family that lacked genome-level taxonomic annotations.

4.3.6 SINGLE-SPECIES PEPTIDE ABUNDANCE CORRELATIONS

Calculation of taxon-based normalized peptide abundance (TNPA). To mitigate the impact of changes in taxon abundance on peptide abundance and to better study the effect of different functions on peptide abundance, taxon-based normalized peptide abundance (TNPA) was calculated based on the original peptide abundance (OPA), total peptide abundance of all peptides from the same taxon in the sample (TPATS, Total Peptide Abundance per Taxon in a Sample), and an average level of peptide abundance (10^8) using the following formula:

$$\text{TNPA} = \left(\frac{\text{OPA}}{\text{TPATS}} \right) \times 10^8$$

TNPA was calculated only for the top 10 genomes from each individual microbiome with the largest number of identified genome-distinct peptides. The subsequent single-species peptide abundance correlation analysis also focused on these genomes.

Single-species peptide abundance correlation network. Two types of single-species peptide abundance correlation networks were constructed using the SCCs of the log2-FC of OPA and the SCCs of log2-FC of TNPA. Peptide pairs with the top 5% SCCs were retained for network construction, which was carried out using the *igraph* package (Gábor et al., 2025) in R. Nodes and edges from each network were extracted with the same package, and all peptides in networks were mapped to their protein source annotations.

Single-species peptide abundance correlation networks were visualized using *Gephi* 0.10.1 with the Yifan Hu layout (Bastian et al., 2009). Network modularity and clustering coefficients were calculated using the modularity and transitivity functions from the *igraph* package, respectively. Networks constructed with TNPA were used to study the correlations between peptides with related functions and to predict the functions of previously uncharacterized proteins.

Applying network connections to predict protein functions. Based on the functional annotations assigned to peptides, the percentage of connections between peptides derived from PUFs in each of the single-species peptide correlation networks of the selected ten genomes, as well as the percentage among all peptide pairs from each species, was calculated. After that, the proportion of connections between peptides from proteins of unknown functions (PUFs) and those from annotated proteins, relative to all connections in each single-species peptide correlation network, was further calculated. Connections between peptides in the networks were analyzed based on protein source and functional annotations of the peptides. InterProScan (<https://www.ebi.ac.uk/interpro/search/sequence/>) and the AlphaFold Protein Structure Database (Varadi et al., 2024) were queried online using default settings, with protein sequences extracted from the UHGP catalog (Almeida et al., 2021).

4.3.7 DATA AVAILABILITY

Metaproteomics raw files used to compile analysis were deposited at the ProteomeXchange Consortium (Deutsch et al., 2019) via the PRIDE (Perez-Riverol et al., 2022) partner repository as described in our previous study (Li et al., 2025). These files will be publicly available upon publication. Additionally, another dataset generated using a Q Exactive mass spectrometer is available with the data set identifiers PXD012724. All codes to perform the analysis in this study are available on GitHub at https://github.com/northomics/Peptide_Abundance_Correlations.

4.4 RESULTS

4.4.1 PEPTIDE ABUNDANCE CORRELATIONS IN THE METAPROTEOMICS DATASET

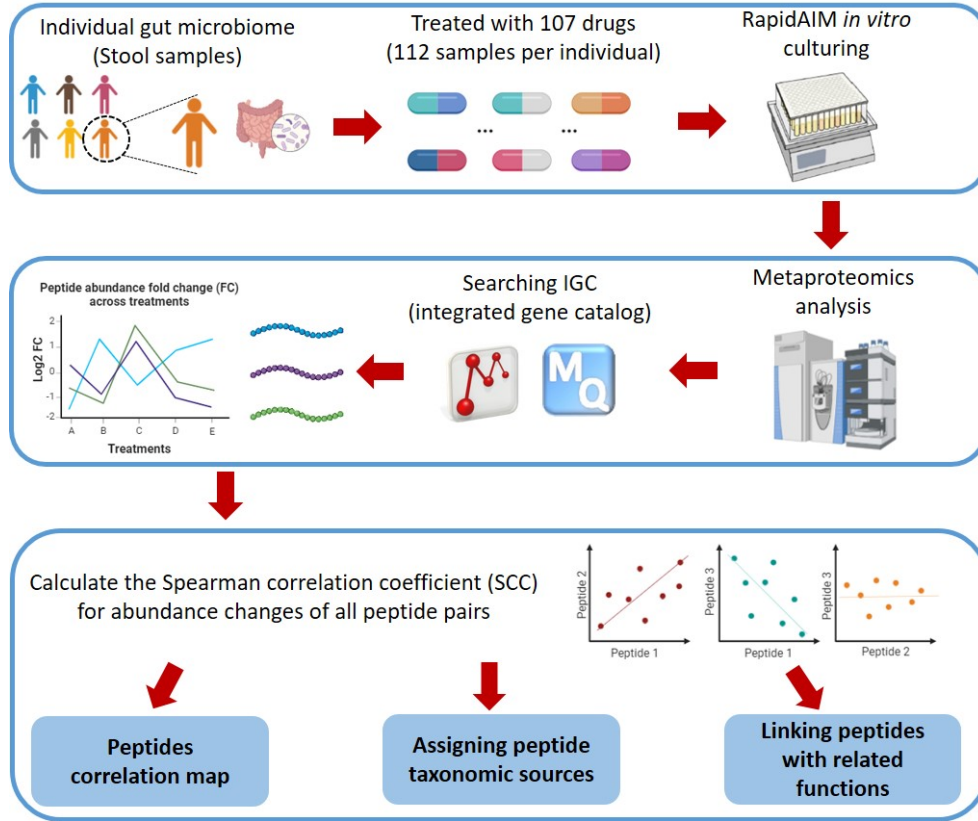


Figure 4.1. Experimental design and research framework. Gut microbiomes from six individuals were extracted from human stool samples. The microbiomes were then treated with 107 various drugs, resulting in 112 samples per individual, and *in vitro* cultured using RapidAIM. After culturing, samples were loaded to LC-MS/MS for metaproteomics analysis. Raw files were searched against the IGC database using the MaxQuant workflow from MetaLab V2.3, yielding a peptide abundance table. To measure peptide abundance correlations, we calculated the Spearman correlation coefficient (SCC) of peptide abundance fold changes relative to the control sample for all peptide pairs. We then created a peptide correlation map to visualize these correlations. Furthermore, we utilized these peptide abundance correlations to assign peptide taxonomic sources, connect peptides from functionally related proteins, and predict the functions of previously uncharacterized proteins.

4.4. RESULTS

Figure 4.1 summarizes the overall experimental design and research framework. Briefly, 6 individual microbiomes were individually exposed to 107 drugs and five controls using the RapidAIM assay (Li et al., 2020), which is a 96-well plate assay that maintains the composition and function of the microbiome. The metaproteomes were extracted from each well and analyzed by metaproteomics. Peptide identification and quantification results were acquired as previously described (Li et al., 2025). After acquiring peptide quantification results, peptides with non-zero values in $\geq 20\%$ of samples from each individual were selected. Peptide abundance correlations for all peptide pairs from each individual were calculated using Spearman correlation coefficients (SCCs) for log2-transformed abundance fold changes from all 112 samples against the control group. Global peptide abundance correlation maps created with t-SNE of all six individuals colored with family-level taxonomic annotation of peptides showed a clear clustering of peptides from the same family (Figure 4.2A). In addition, these maps revealed substantial inter-individual variation in the taxonomic composition of microbiomes, which was further supported by peptide-based taxonomic profiling for each individual (Supplementary Figure S4.1). Among all six individuals, V52 has the largest number of quantified peptides (Figure 4.2B) and was selected as an example in the main text. For 21,363 peptides with non-zero values in $\geq 20\%$ of samples from V52, SCC calculation yielded a total of 228,178,203 peptide pairs with an average SCC of 0.13 ± 0.35 (Figure 4.3A).

4.4. RESULTS

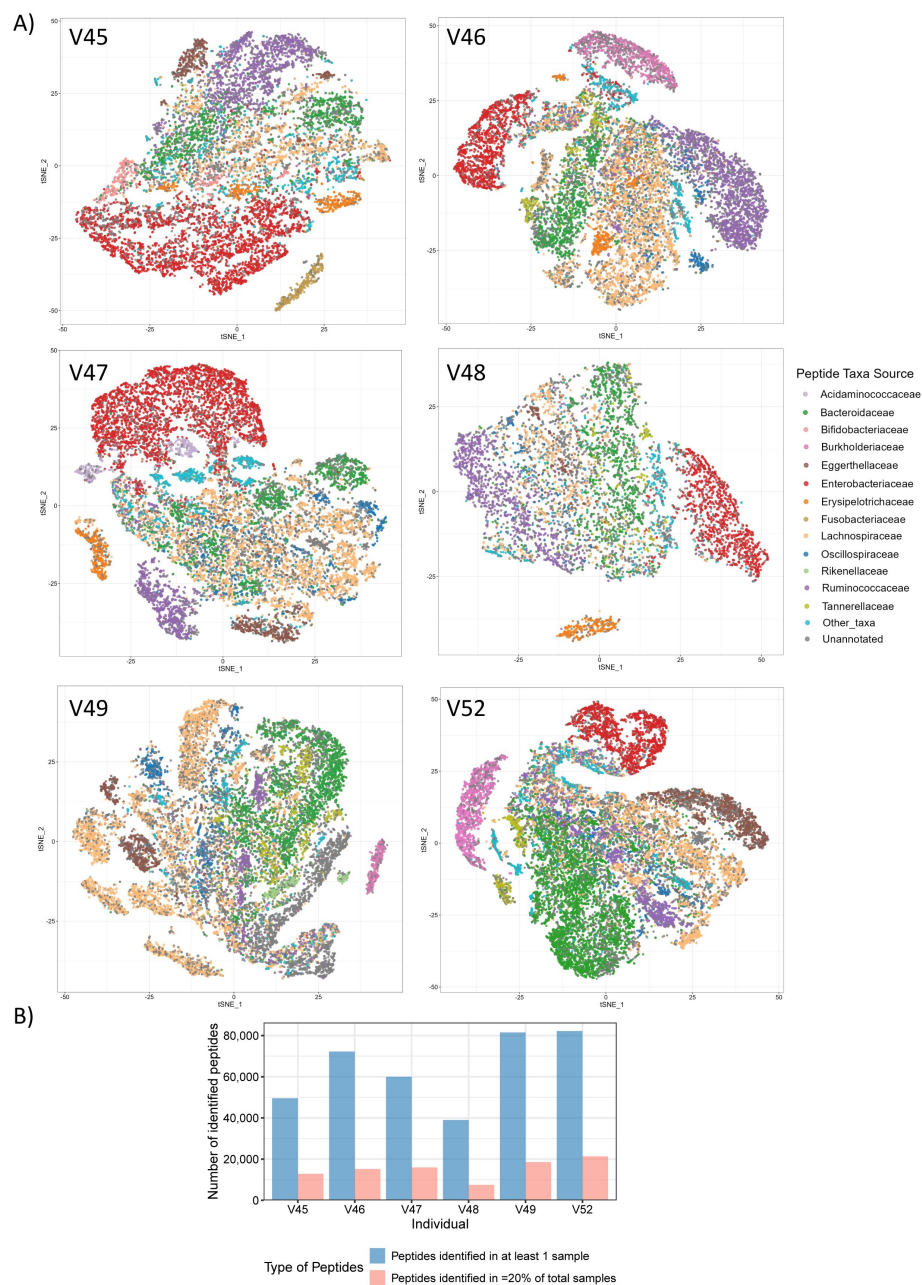


Figure 4.2. Peptide identification and abundance correlation across individuals. A) Peptide correlation maps generated using tSNE of all six individuals, colored by peptide family-level taxonomic annotations. B) Number of peptides identified from different individuals. For each individual, peptides identified in $\geq 20\%$ of total samples were used for abundance correlation analysis and displayed in the peptide correlation map.

4.4. RESULTS

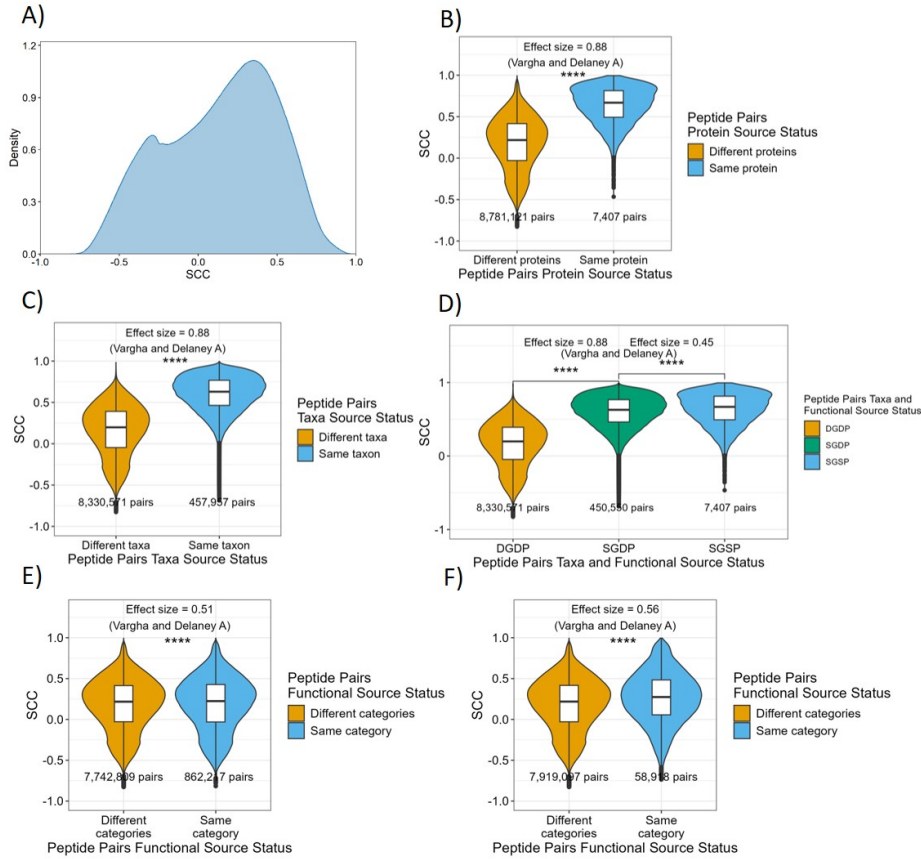


Figure 4.3. Peptide abundance correlations in the metaproteomics dataset of V52. A) Distribution of pairwise Spearman correlation coefficients (SCC) of peptide abundance fold changes for all peptide pairs. B) Comparison of the SCC of peptide pairs with both peptides from the same protein and peptide pairs of peptides from different proteins. C) Comparison of the SCC of peptide pairs of both peptides from the same taxon (genome) and peptide pairs of peptides from different taxa (genomes). D) Comparison of peptide pairs of both peptides from the same genome and same protein (SGSP), peptides from the same genome but different proteins (SGDP), and peptides from different genomes and different proteins (DGDP). E) Comparison of SCCs between peptide pairs from the same COG category and peptide pairs from different COG categories. F) Comparison of SCCs between peptide pairs from the same COG family and peptide pairs from different COG families. **** indicates statistical significance at the $p \leq 0.0001$ levels by two-sided Mann-Whitney U test.

Upon assigning both protein sources and taxonomic sources to the peptides, we observed that peptides from the same protein and peptides from the same taxon exhibited higher abundance correlation. First, peptide pairs derived from the same protein (7,407 pairs) exhibited higher SCCs (0.63 ± 0.22) compared to peptide pairs from different proteins (8,781,121 pairs; $\text{SCC} = 0.18 \pm 0.32$, with $p \leq 0.0001$ by Mann-Whitney U test and a

4.4. RESULTS

large effect size with Vargha and Delaney’s A of 0.88, Figure 4.3B). To be more intuitive, peptides from the same protein displayed more similar abundance profiles across samples than those from different proteins (Supplementary Figure S4.2). In addition, peptide pairs derived from the same genome (457,957) had an average SCC of 0.60 ± 0.22 , while those from different genomes (8,330,571) averaged an SCC of 0.16 ± 0.31 (Figure 4.3C, $p \leq 0.0001$ by Mann-Whitney U test and a large effect size with Vargha and Delaney’s A of 0.88). It is worth noting that even after excluding peptide pairs from the same protein, we found that peptides from different proteins within the same genome (450,550 pairs; $\text{SCC} = 0.60 \pm 0.22$) still had higher SCCs than those from different genomes (8,330,571 pairs; $\text{SCC} = 0.16 \pm 0.31$, $p \leq 0.0001$ by Mann-Whitney U test, Vargha and Delaney’s A of 0.88), indicating that sourcing from the same genome contributes significantly to higher SCCs of peptide abundance changes (Figure 4.3D). We also assigned functional annotations to the peptides, however, the difference in SCCs between peptides from the same functional category and those from different functional categories was relatively minor at both high-level COG category (Figure 4.3E) and refined COG family level (Figure 4.3F), with negligible effect size, Vargha and Delaney’s A of 0.51 and 0.56, respectively.

Higher SCC of peptides from the same protein and peptides from the same taxon collectively suggested that studying peptide abundance correlations in metaproteomics datasets is biologically meaningful and can provide valuable insights into metaproteomics analysis.

4.4.2 A GLOBAL PEPTIDE ABUNDANCE CORRELATION MAP

The SCC matrix of selected 21,363 peptides from V52 recorded how strongly or weakly each peptide is correlated with all other peptides. Although it is theoretically possible to be represented as a peptide interaction network with edges indicating strong correlations (Supplementary Figure S4.3). We visualized all 228,178,203 peptide-peptide correlations

4.4. RESULTS

from 21,363 peptides as a global peptide correlation map using the t-SNE algorithm by embedding the peptides in a low-dimensional space (Figure 4.4A). In this map, the distance between peptides reflects the similarity in their abundance changes under various drug treatments. Notably, the global map was generated based on all pairwise peptide abundance changes under different perturbations in the metaproteomics datasets per individual, without any filtering on the SCC.

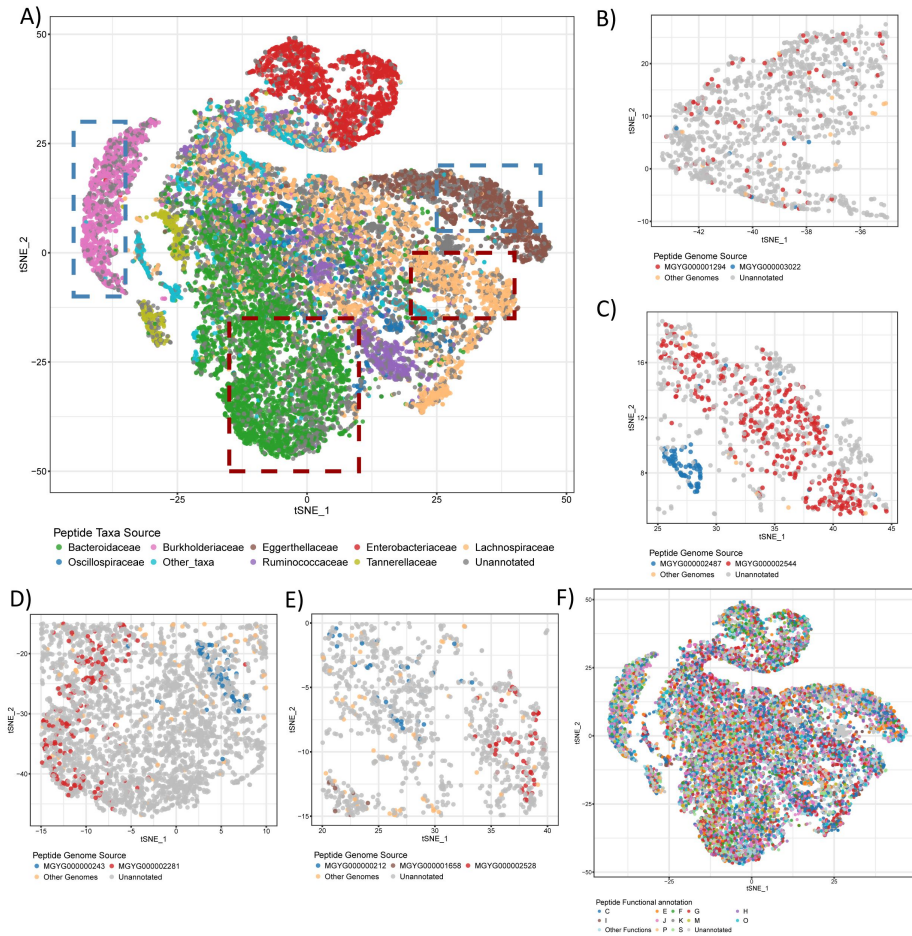


Figure 4.4. The global peptide correlation map of V52 generated using t-SNE. A) Peptide correlation map colored by family-level taxonomic annotations, showing the map broadly corresponds to taxa sources of peptides. The areas highlighted in panels B-E are boxed with dashed lines. B-E) Zoomed-in sections of the peptide correlation map for peptides from the different families: B) Burkholderiaceae, C) Eggerthellaceae, D) Bacteroidaceae, and E) Lachnospiraceae. Peptides are colored by genome-level taxonomic annotations. F) Peptide correlation map colored by functional annotations (COG categories), showing no overall clustering based on functional associations.

4.4. RESULTS

The peptide correlation map revealed a strong correlation between peptide abundance changes and their taxonomic sources. Peptides from the same taxon tend to cluster together in the correlation map across various taxonomic levels (Figure 4.4A and Supplementary Figure S4.4). Specifically, at the family level, distinct clusters emerged, particularly for families such as Burkholderiaceae, Eggerthellaceae, and Enterobacteriaceae. Zooming into these families, most peptides with genome-level annotations were either assigned to a specific genome or grouped into sub-clusters corresponding to different genomes (Figure 4.4B and C). For other families, such as Bacteroidaceae and Lachnospiraceae, the cluster of peptides was less obvious. However, peptides were still grouped at the genome level in the zoomed-in maps (Figure 4.4D and E).

Although there is a clear clustering of peptides from the same taxon, the global peptide abundance correlation map showed less obvious clustering patterns based on protein function (Figure 4.4F and Supplementary Figure S4.5). However, when zooming into a single species, more distinct clusters emerged, representing peptides from proteins with similar or related functions (Supplementary Figure S4.6). This suggests that studying protein functional linkage through peptide abundance correlations may be more effective at the single-species level, as inter-species functional correlations are not easily discernible in the global map.

Overall, this global peptide correlation map provides an overall trend of peptide abundance correlations between different peptides and indicates that changes in taxa abundance are the primary driver of peptide abundance correlations across different perturbations.

4.4.3 PEPTIDE ABUNDANCE CORRELATIONS PROVIDE ADDITIONAL INFORMATION ON ASSIGNING PEPTIDE TAXONOMIC SOURCE

Our results have shown that peptides from the same family clustered together in the global peptide correlation map (Figure 4.4A). However, within each family-level cluster,

4.4. RESULTS

only a small proportion of peptides were annotated to specific genomes, while a larger proportion were only annotated at the family level, leaving their genome-level sources unclear (Figure 4.4B-E and Supplementary Figure S4.4). Also as mentioned above, peptides from the same genome exhibit high SCCs in their abundance changes across different drug treatments (Figure 4.3C). Given a specific sample from the family Bacteroidaceae, genome-distinct peptides from the same genome were not grouped into a single cluster in the heatmap of the peptides SCC matrix. However, these peptides can still be grouped into several modules, each with high correlation coefficients (Figure 4.5A). Similar genome cluster modules were also observed in other families, such as Eggerthellaceae, Lachnospiraceae, and Burkholderiaceae (Supplementary Figure S4.7-Supplementary Figure S4.9). This observation suggests that the SCC matrix can serve as a valuable input for machine learning models to assign peptide taxonomic sources.

4.4. RESULTS

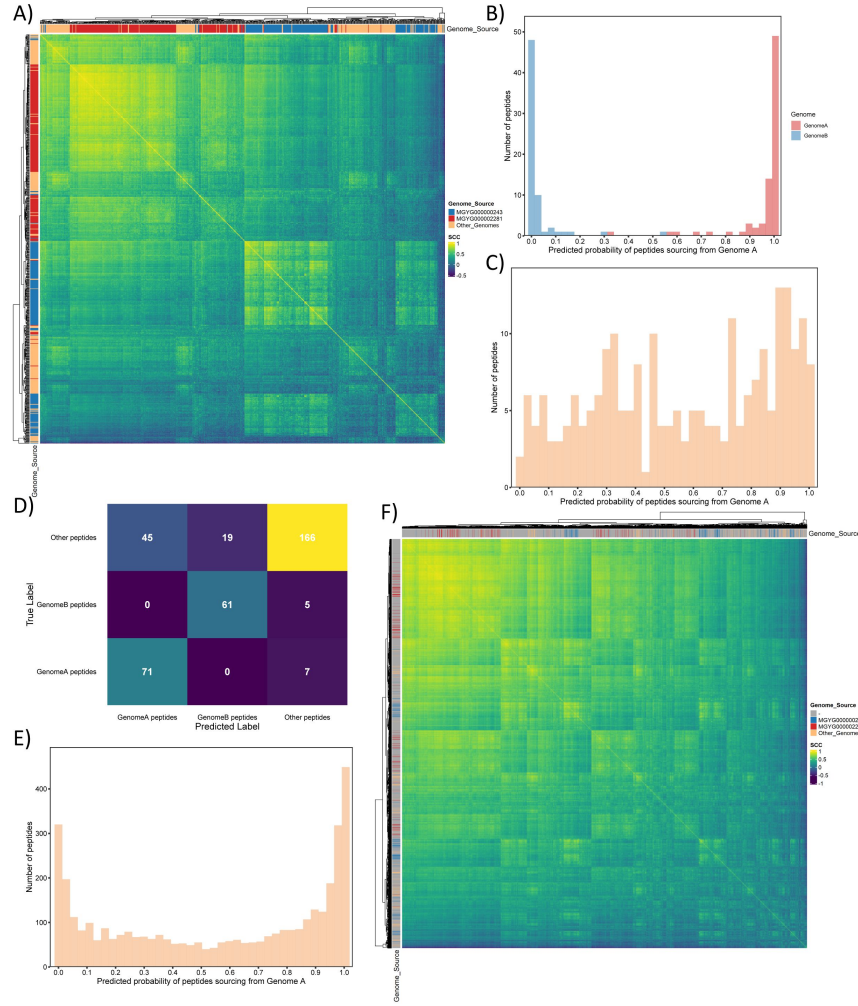


Figure 4.5. Applying peptide abundance correlations for peptide taxonomic assignments using the Bacteroidaceae family as an example. A) Heatmap of the SCC matrix for genome-distinct peptides from different genomes within the Bacteroidaceae family. B) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on genome-distinct peptides in the test dataset. C) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on peptides from other genomes within the same family. D) Confusion matrix from the Random Forest model applied to the combined dataset of genome-distinct peptides in the test dataset and peptides from other genomes within the family. E) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on genome-unannotated peptides within the Bacteroidaceae family. F) Heatmap of the SCC matrix for both genome-distinct peptides and genome-unannotated peptides within the Bacteroidaceae family. Genome-unannotated peptides were colored with grey in the color strip. Genome A, MGYG000002281; Genome B, MGYG000000243.

Using the trained Random Forest model for the family Bacteroidaceae, we found that most genome-distinct peptides from Genome A (MGYG000002281) and Genome B (MGYG000000243) in the test set were correctly classified with probabilities exceeding 90%

(Figure 4.5B). Specifically, the model assigned these peptides to their respective genome sources with high confidence. In contrast, when applying the same 90% probability threshold to peptides from other genomes within the same family and peptides from other families, the model did not tend to attribute most of the peptides from other genomes within the same family (72.2%) and the peptides from other families (87.6%) to either Genome A or Genome B (Figure 4.5C and Supplementary Figure S4.10). The confusion matrices generated from the two test sets demonstrated that the trained model achieved high sensitivity (Figure 4.5D and Supplementary Figure S4.10). For peptides without genome-level annotations from the same family, the model classified a large amount of these peptides (48.9%) into either Genome A or Genome B (Figure 4.5E). This classification was further supported by the SCC heatmap, where unannotated peptides clustered closely with genome-distinct peptides from a specific genome (Figure 4.5F). In addition, models trained using the same method also effectively predicted the genome sources of peptides lacking genome-level annotations in other families, such as Eggerthellaceae, Lachnospiraceae, and Burkholderiaceae (Supplementary Figure S4.7-Supplementary Figure S4.9).

4.4.4 FUNCTIONALLY RELATED PEPTIDES WERE CONNECTED IN SINGLE-SPECIES PEPTIDE ABUNDANCE CORRELATION NETWORKS

Our results revealed limited abundance correlations between peptides from proteins with the same or related functions (Figure 4.3E-F, and Supplementary Figure S4.5). Additionally, we found that taxa abundance changes are the major driver of peptide abundance changes across different perturbations or drug treatments. Consequently, peptides from proteins with the same or related functions were less likely to be correlated across different taxa compared to peptides originating from the same taxon. Therefore, it is more appropriate to analyze peptide functional linkages at the single-species level by focusing on single-species subsets of the metaproteomics dataset.

4.4. RESULTS

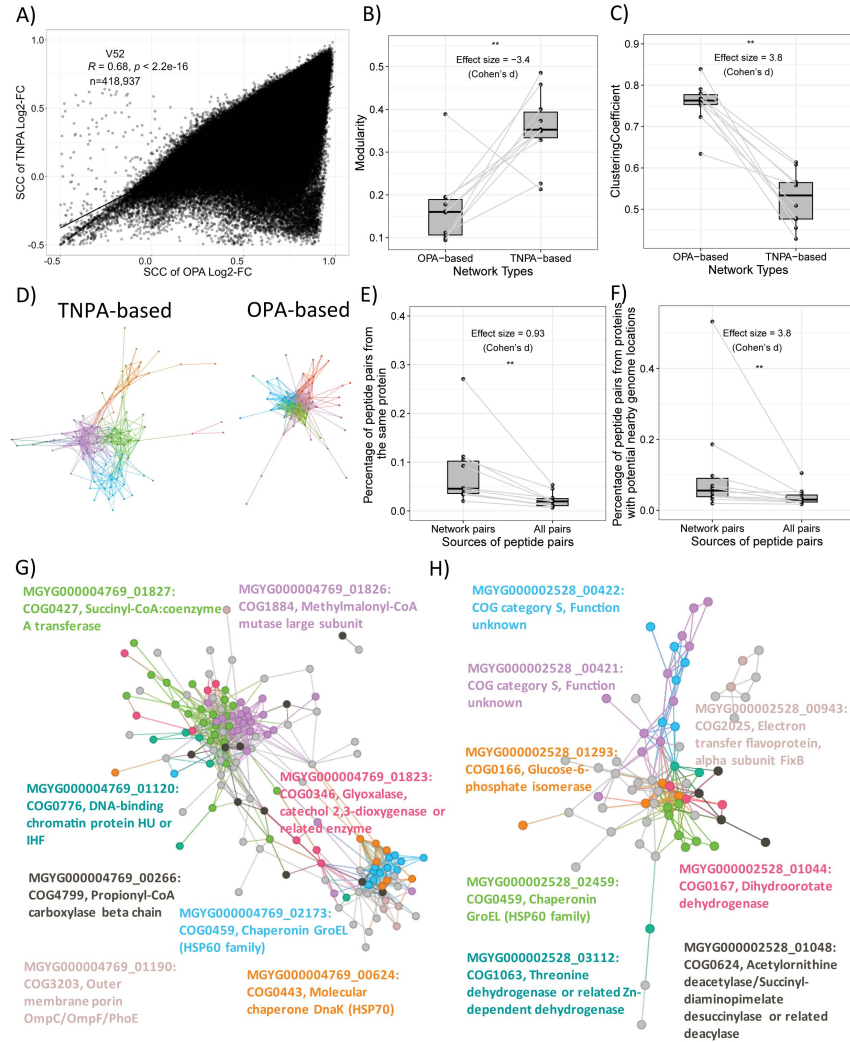


Figure 4.6. Functionally related peptides were connected in V52 single-species peptide abundance correlation networks constructed with SCCs of TNPA log2-FC. A) Correlation between SCCs of peptide log2-FC based on TNPA and SCCs based on OPA (original peptide abundance) in the top ten genomes from V52. B)-C) Higher modularity and lower clustering coefficient of peptide correlation networks constructed with SCC of TNPA log2-FC than networks constructed with SCC of OPA log2-FC. D) Comparison of representative single-species peptide correlation networks constructed with TNPA and OPA. Nodes are colored according to their modularity class from *Gephi*. E) Proportion of peptide pairs from the same protein in single-species peptide correlation networks compared to the proportion in all peptide pairs from the corresponding species. F) Proportion of peptide pairs from proteins potentially located near each other in the genome (≤ 10 gene ID difference) in single-species peptide correlation networks compared to the proportion in all peptide pairs from the corresponding species. ** indicates statistical significance at the $p \leq 0.01$ levels by two-sided Mann-Whitney U test. G)-H) Single-species peptide abundance correlation networks for MGYG000004769 and MGYG000002528. Peptides from proteins with only one peptide in the network were removed. Peptides from the top eight proteins with the most peptides in each network are annotated with distinct colors, and peptides from other proteins are colored grey.

4.4. RESULTS

The strong effect of taxonomic abundance changes could cause peptides from the same taxon to show high correlations, even if they are involved in different functions (Supplementary Figure S4.11). To minimize this effect and better study the correlations between functionally related peptides, we calculated taxon-based normalized peptide abundance (TNPA) from original peptide abundance (OPA) for the top 10 species/genomes with the highest number of identified peptides, as described in the methods section. TNPA was mostly moderately or weakly positively correlated with OPA in most studied species (Supplementary Figure S4.12), and we observed that some peptides with high SCCs for OPA log2-FC exhibited low SCCs for TNPA log2-FC (Figure 4.6A and Supplementary Figure S4.13). This suggests that many peptide abundance correlations driven by taxa abundance changes were not significantly correlating at the TNPA level as shown in Supplementary Figure S4.10.

To reveal functional linkages, single-species peptide abundance correlation networks were constructed using peptide pairs with the top 5% SCC from each of the top 10 species with the most identified peptides. Two sets of networks were constructed using SCCs derived from OPA and TNPA, respectively (Supplementary Figure S4.14 and Supplementary Figure S4.15). Networks constructed with TNPA had higher modularity (Figure 4.6B) and lower clustering coefficients (Figure 4.6C), indicating that these networks had dense connections within modules but sparse connections between different modules. The lower clustering coefficients also suggested that TNPA-based networks were less tightly connected as shown in the Figure 4.6D.

Single-species peptide abundance correlation networks constructed with TNPA revealed functional linkages. Peptides from the same protein and those from proteins potentially located near each other in the genomes (with ≤ 10 gene ID differences in the UHGG) were considered functionally related. Peptide pairs in the network showed a higher percentage of these types of relationships compared to all peptide pairs (Figure 4.6E-F). It is worth noting that gene ID differences do not always perfectly reflect physical genome dis-

4.4. RESULTS

tance; however, they generally indicate proximity when genes are located on the same contig. Here, we provide examples of connections between peptides from functionally related proteins. In the peptide correlation network of genome MGYG000004769 (*Phascolarctobacterium faecium*), peptides from MGYG000004769_01826 (Methylmalonyl-CoA mutase large subunit) and MGYG000004769_01827 (Succinyl-CoA: coenzyme A transferase) were densely connected (Figure 4.6G). Genes encoding these two proteins are located near each other in the genome, with adjacent gene ID numbers and only a 218bp intergenic region on the same contig. These proteins also have closely related functions in the TCA cycle. Methylmalonyl-CoA mutase catalyzes the conversion of methylmalonyl-CoA to succinyl-CoA, while Succinyl-CoA:coenzyme A transferase then catalyzes the reversible reaction: $\text{succinyl-CoA} + L\text{-malate} \rightleftharpoons \text{succinate} + L\text{-malyl-CoA}$. In the same network, peptides from MGYG000004769_00624 (Molecular chaperone DnaK) and MGYG000004769_02173 (Chaperonin GroEL), which have similar functional annotations, were also connected. In another peptide correlation network of genome MGYG000002528 (*Anaerostipes hadrus*), peptides from proteins MGYG000002528_01044 (NAD-dependent dihydropyrimidine dehydrogenase subunit PreA) and MGYG000002528_01048 (*Allantoate amidohydrolase*) were densely connected (Figure 4.6H). These proteins are also encoded by genes with adjacent gene numbers which are separated by a 2,717 bp intergenic region.

In summary, single-species peptide abundance correlation networks constructed using TNPA revealed connections between functionally related peptides. These connections demonstrated the potential for using peptide abundance correlations to predict protein functions.

4.4.5 APPLYING SINGLE-SPECIES PEPTIDE CORRELATION NETWORKS FOR PREDICTING UNKNOWN PROTEIN FUNCTIONS

An average of 5.2% of peptides in the 10 constructed single-species peptide correlation networks originated from proteins of unknown functions (PUFs). This percentage was comparable to the overall proportion of peptides from unannotated proteins in each species (Figure 4.7A). To predict functions for PUFs we calculated peptide abundance correlations between peptides from PUFs and peptides from annotated proteins. The proportion of connections between peptides from PUFs and peptides from annotated proteins varied across species (Figure 4.7B). For genome MGYG000002478 (*Phocaeicola dorei*), the proportion reached 24%, while in other genomes, the lowest was 2%. Here, we provide examples from the top two genomes with the highest percentage of connections between peptides from PUFs and peptides from annotated proteins to show the feasibility of predicting microbial protein functions using peptide abundance correlations.

4.4. RESULTS

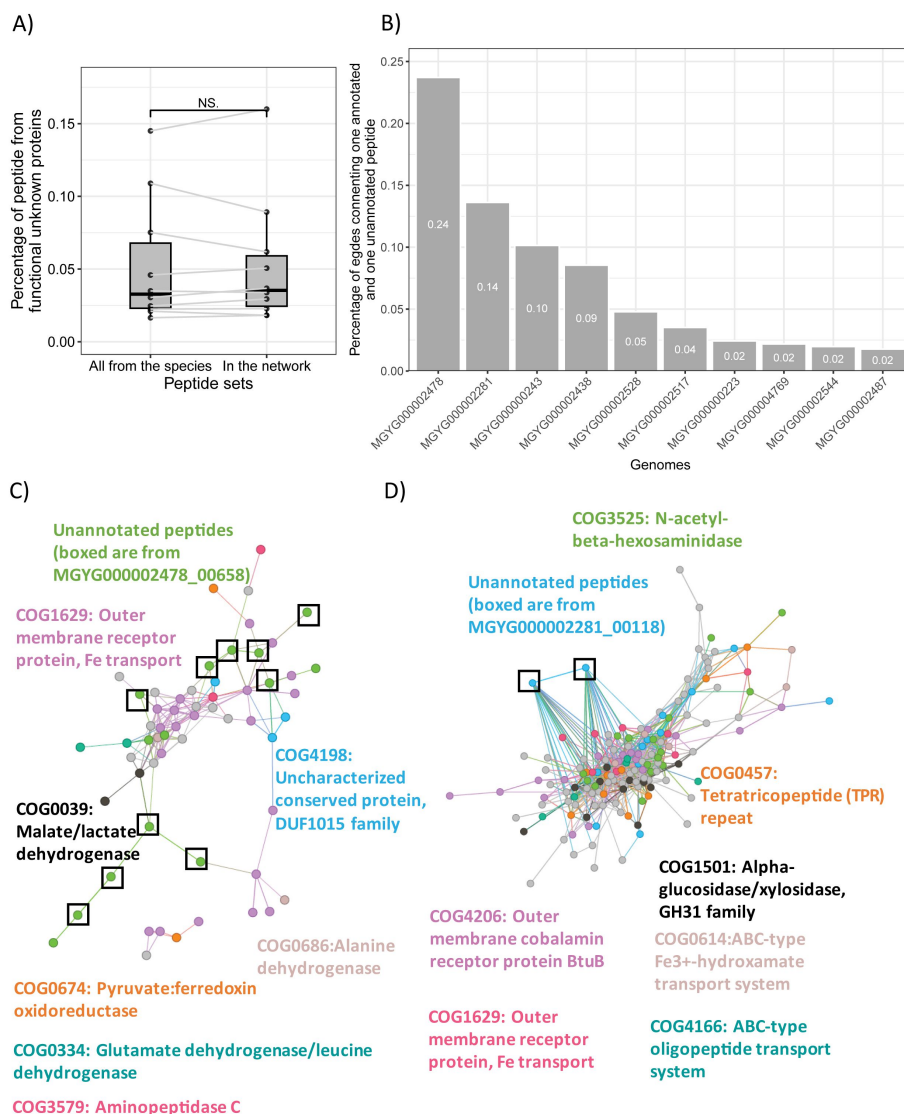


Figure 4.7. Predicting functions of proteins of unknown functions (PUFs). A) Proportion of peptides from PUFs in each single species peptide correlation network and proportion of peptides from PUFs in all identified peptides from the corresponding species. NS. indicates statistical significance at the $p > 0.05$ levels by two-sided Mann-Whitney U test. B) Distribution of the proportions of edges between peptides from PUFs and peptides from proteins with known function in the peptide correlation network of each species. C)-D) Peptide abundance correlation networks for the two species, MGYG000002478 and MGYG000002281, with the highest percentage of edges between peptides from PUFs and peptides from proteins with known function. Networks were constructed based on peptide pairs with the top 5% SCCs from each species. Peptides from proteins with only one peptide in the network were removed. Peptides from the top eight COG families in each network are annotated with distinct colors, and peptides from other COG families are colored grey.

In the network of genome MGYG000002478 (Figure 4.7C), *Phocaeicola dorei*, peptides

4.4. RESULTS

from uncharacterized protein MGYG000002478_00658 had the largest number of connections (31) to annotated peptides in the network. Of these, 10 connections were to peptides annotated to COG1629 (Outer membrane receptor protein, Fe transport), with 7 linked to peptides from MGYG000002478_00657, an iron complex outer-membrane receptor protein (K02014). This indicates that the uncharacterized protein is likely to be related to transmembrane transport. This prediction was further supported by the InterProScan of the uncharacterized protein, showing it has a SusD-like domain, a typical outer membrane protein feature. In addition, most parts of the protein were assigned to a non-cytoplasmic domain, also indicating its extracellular location.

In the network of genome MGYG000002281 (Figure 4.7D), *Bacteroides faecis*, peptides from uncharacterized protein MGYG000002281_00118 had the largest number of connections (52) to annotated peptides in the network. Excluding 9 peptide connections to functionally ambiguous COG0457 (Tetratricopeptide repeat protein), the second-largest group comprises 7 connections to COG3525 (N-acetyl-beta-hexosaminidase), including 3 connections to MGYG000002281_02589 (hyaluronoglucosaminidase, K01197), 2 to MGYG000002281_04635 (hexosaminidase, K12373), and 2 to MGYG000002281_01369 (hexosaminidase, K12373). These connections suggest the uncharacterized protein is likely to be involved in carbohydrate metabolism. An InterProScan analysis of the protein revealed a non-cytoplasmic domain, indicating that it is a membrane-bound protein predicted to be outside the membrane, possibly acting as a signal protein for carbohydrate metabolism. In contrast, searching this protein in the AlphaFold Protein Structure Database (Jumper et al., 2021; Varadi et al., 2024) matched UniProt ID A0A1E9BZD4 (100% identity), an uncharacterized protein without known biological functions. These findings highlight the value of peptide correlation networks in providing complementary insights for predicting the potential biological role of PUFs.

4.5 DISCUSSION

We analyzed peptide abundance correlations in a metaproteomics dataset of *in vitro* cultured human gut microbiomes with different drug treatments/perturbations, demonstrating the feasibility of applying peptide abundance correlations for peptide taxonomic assignments as well as revealing protein functional linkages in single-species subsets.

Protein abundance correlations have been applied to study protein functional linkage and to predict functions of functional unknown proteins in single-species proteomics datasets (Kustatscher et al., 2019; Mateus et al., 2020; Messner et al., 2023). In these single-species studies, functionally associated proteins have coordinated changes in abundance across perturbations (Kustatscher et al., 2019; Mateus et al., 2020; Messner et al., 2023), and proteins with identical subcellular localization also exhibited coordinated abundance changes (Kustatscher et al., 2019). However, in our metaproteomics dataset, the abundances of peptides from proteins of the same taxon instead of peptides from proteins with associated functions were more likely to have correlated abundance changes.

Peptides from the same taxon showing high abundance correlation are largely due to the significant impact of environmental disturbances on the composition of microbial communities (Nguyen et al., 2021). In contrast, the limited correlation of peptides from proteins with related functions can be attributed to the high functional redundancy of human gut microbial communities (Li et al., 2023). Specifically, a decrease in peptides from proteins with a particular function could lead to an increase in peptides from proteins performing a similar function in a phylogenetically unrelated species. This compensatory mechanism helps maintain the overall functional stability of the microbial community, resulting in limited correlations among peptides from proteins with related functions.

Our study is the first to investigate peptide abundance correlations in a gut microbiome metaproteomic dataset. However, similar studies have been conducted using metagenomic

data, such as genetic correlation networks from soil metagenomes which have revealed a hierarchical functional structure (Ma et al., 2018), similar to findings in single-species cellular genetic correlation networks (Costanzo et al., 2016). In contrast, our analysis of peptide abundance correlations primarily showed a clear clustering of peptides from the same taxon, with a less obvious functional hierarchy. This underscores that functional insights derived from protein abundance measurements are different from those inferred from genetic materials through metagenomics or metatranscriptomics (Armengaud, 2023).

Our global peptide abundance correlation map also provided clues for studying microbial responses to drugs. In the constructed peptide correlation maps (Figure 4.2A), peptides from the family Enterobacteriaceae were usually separated from other taxa, indicating unique response patterns for this taxonomic group. This is supported by previous experimental studies. For example, psychotropic drugs like fluoxetine have shown antimicrobial activity against *Escherichia coli* (family Enterobacteriaceae), while the abundance of other human gut species increased in the in vitro microbiome after the treatment of the same drug (Cussotto et al., 2019). In addition, microbial genera have been classified into three distinct taxonomic clusters, each representing a different pattern of drug response (Li et al., 2025). Patterns resembling these taxonomic clusters can also be observed in our peptide correlation maps. For instance, peptides from *Parasutterella* (Burkholderiaceae family) and *Eggerthella* (Eggerthellaceae family), which belong to two clusters exhibiting distinct response patterns, are also well separated in our peptide correlation map (Figure 4.2A and Supplementary Figure S4.4).

Higher correlations of peptides from the same taxon also provide a solution for peptide taxonomic source assignment which remains a challenge for peptide-centric metaproteomics analysis. Although most of the in-silico digested peptides were genome-distinct peptides (Mesuere et al., 2012), most of the peptides identified from real metaproteomics datasets were shared by different genomes (Sun et al., 2023). In this study, even after refining the potential genome sources, a large number of peptides could only be assigned to a

family-level LCA, and the actual taxonomic sources of these peptides remain unknown (Figure 4.4B-E). By investigating peptide abundance correlations, additional information was provided for peptide taxonomic source assignment. This is expected to provide a microbial biomass profile with higher accuracy using metaproteomics data (Kleiner et al., 2017).

Although taxa-abundance changes were major contributors to peptide abundance correlations, to reveal peptide functional linkages, peptide abundance correlations were further analyzed at the single-species level by extracting subsets of genome-distinct peptides of each genome from the metaproteomics dataset. In addition, to reduce the impact of taxa-abundance changes, TNPA was calculated. The principle of the calculation of TNPA was similar to a normalization method named LFQRatio (Shi et al., 2024) which divided protein LFQ intensity by the sum of all protein LFQ intensities for its respective strain in a microbial coculture system. LFQRatio has shown its ability to transform absolute protein quantification data into accurate and biologically meaningful protein abundance values for samples with multiple species at variable cell ratios. Similarly, in our study, we also found that taxon-based normalized peptide abundance (TNPA) showed a good ability to reveal functional linkages (Figure 4.6E-F). Peptides from proteins with related functions were connected in the single-species peptide abundance correlation network constructed with TNPA.

A limitation of this study comes partly from metaproteomics itself. Current metaproteomics has limited proteome coverage (Lohmann et al., 2020; Sun et al., 2024). It is estimated that around 200 bacterial species reside in the human gut microbial community (Lloyd-Price et al., 2016). However, bacterial species with an abundance lower than 0.5% were hard to detect with current metaproteomics techniques (Duan et al., 2022), and the number of peptides and proteins identified in low-abundant species is limited (Sun et al., 2024). To that extent, only the top ten genomes were investigated to study intra-species functional relations in this work. Even in these genomes, the number of peptides was still

4.6. CONCLUSION

limited. This results in the limited scale of constructed single-species peptide correlation networks. A lot of peptides from proteins with interesting or unknown functions were not identified and were not included in the network. In addition to considering proteome coverage, it is also worth noting that, a high-resolution LC-MS/MS for identification and quantification is also needed. Our analysis in another metaproteomics dataset from Q Exactive mass spectrometer with the same procedure showed a weak abundance correlation of same-taxon peptides (Supplementary Figure S4.16). Fortunately, DIA-based metaproteomics has proven to be able to significantly increase the number of identified peptides as well as to improve quantification reproducibilities (Aakko et al., 2020; Zhao et al., 2023) which has the potential to expand and strengthen applying peptide abundance correlations to reveal more interesting findings.

Another aspect that needs improvement is distinguishing true functional correlations from incidental ones in single-species peptide correlation networks. First, the number of samples used to calculate Spearman correlation coefficients should be carefully considered, as these coefficients require sufficient sample sizes to achieve high confidence in the correlations detected (Bonett & Wright, 2000). Additionally, applying a strict threshold or employing more refined methods to filter out functionally related interactions from correlations driven by other factors is crucial, since abundance correlations do not always indicate functional relationships (Chase L. S. & Priya S., 2024). Ultimately, the field would benefit from larger, more reproducible datasets and refined protein association reference methods to advance the study of peptide/protein abundance correlations.

4.6 CONCLUSION

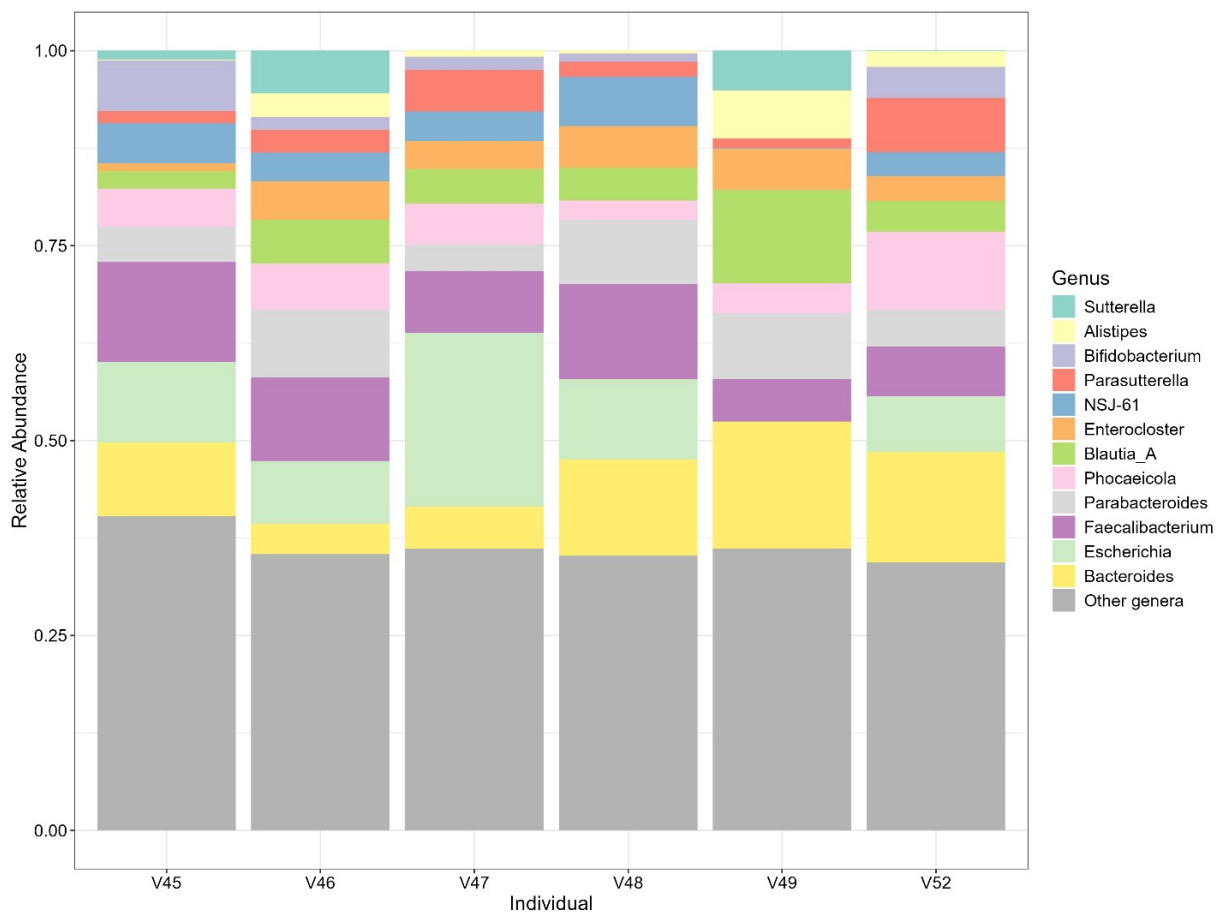
In summary, fluctuations in peptide abundance contain wealthy information on the microbiome’s response to perturbations. These fluctuations can be harnessed to assign peptides to their protein and taxon of origin as well as to predict function for functional unknown

4.6. CONCLUSION

proteins. In the metaproteomics dataset, peptides from the same taxon were clustered in the peptide correlation map, suggesting that microbiome taxonomic abundance change is the major contributor to peptide abundance changes. We anticipate that the concept of peptide/protein abundance correlations needs further investigation to deepen taxonomic and functional understanding of metaproteomics data.

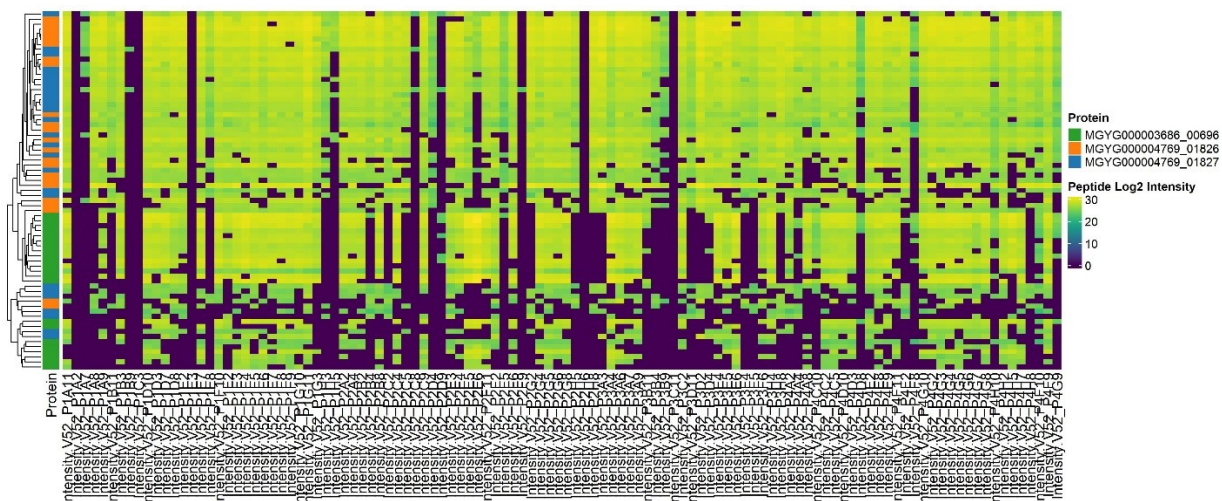
4.7 SUPPLEMENTARY MATERIALS

4.7.1 SUPPLEMENTARY FIGURES

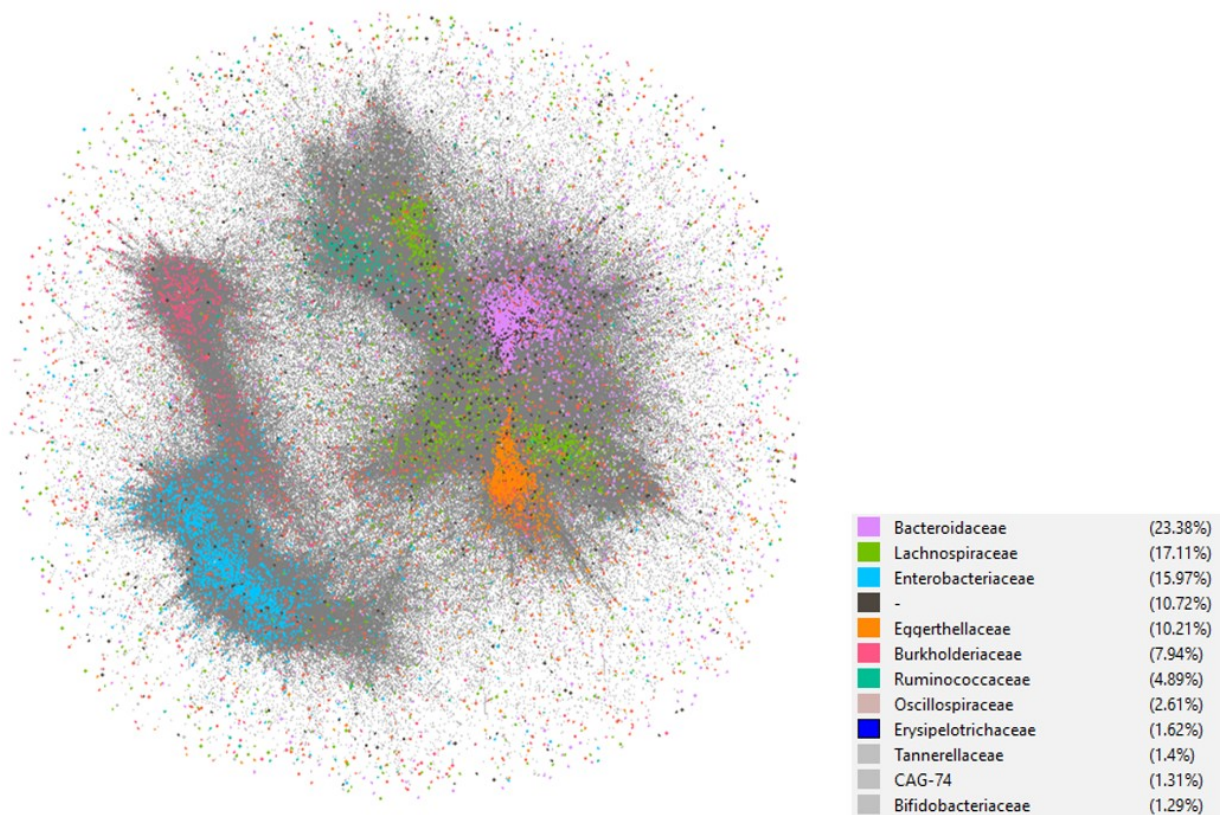


Supplementary Figure S4.1. The relative abundance of peptides assigned to different genera across individuals. Only peptides with genus-level taxonomic annotations were included in the plot. The top 12 genera, ranked by the total number of peptides across all individuals, are shown in distinct colors, while peptides assigned to other genera are grouped under the "Other genera" category.

4.7. SUPPLEMENTARY MATERIALS

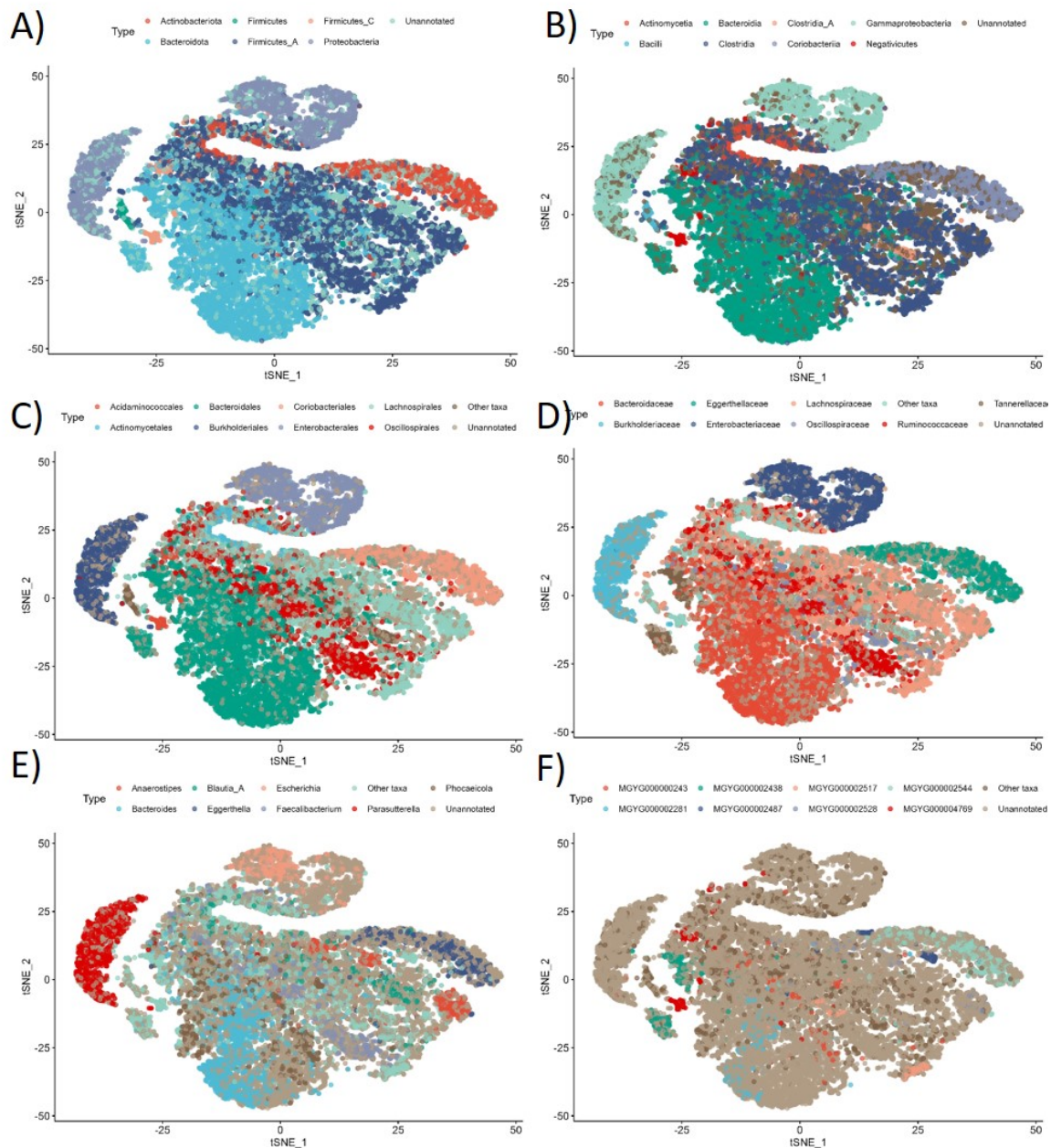


Supplementary Figure S4.2. Peptide abundance profiles from three distinct proteins. Peptides from the same proteins showed similar abundance profiles across samples under different treatments/perturbations.

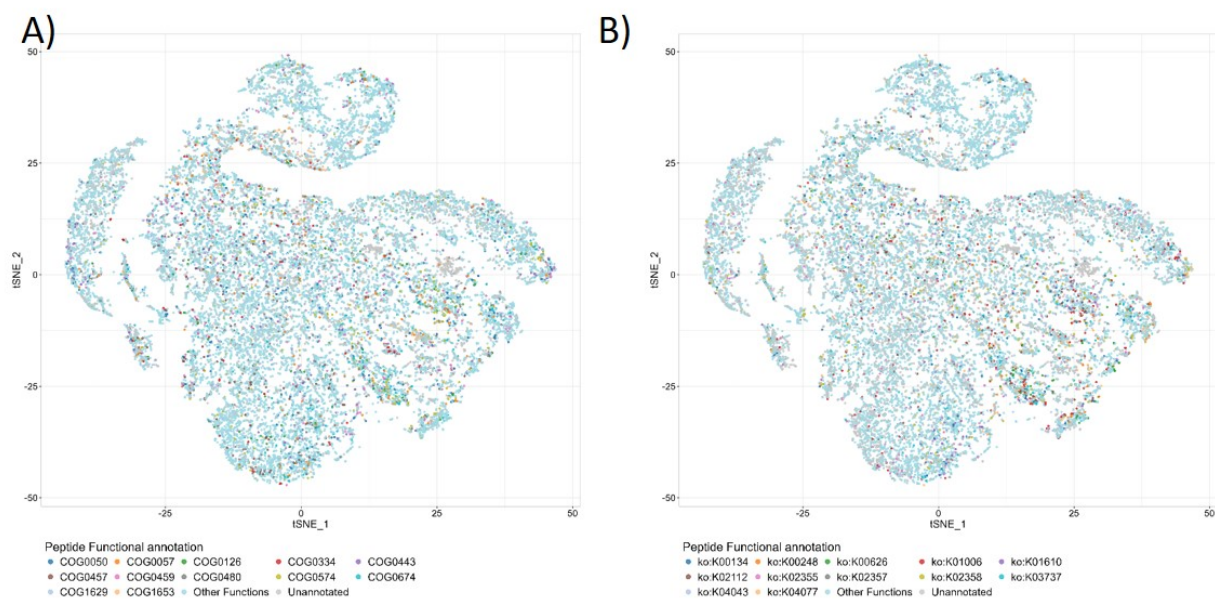


Supplementary Figure S4.3. Global peptide abundance correlation network of V52. The network comprises peptide pairs with the top 1% highest Spearman correlation coefficients (SCC). It consists of 12,549 nodes (representing individual peptides) and 2,281,783 edges (representing peptide pairs with high SCC).

4.7. SUPPLEMENTARY MATERIALS

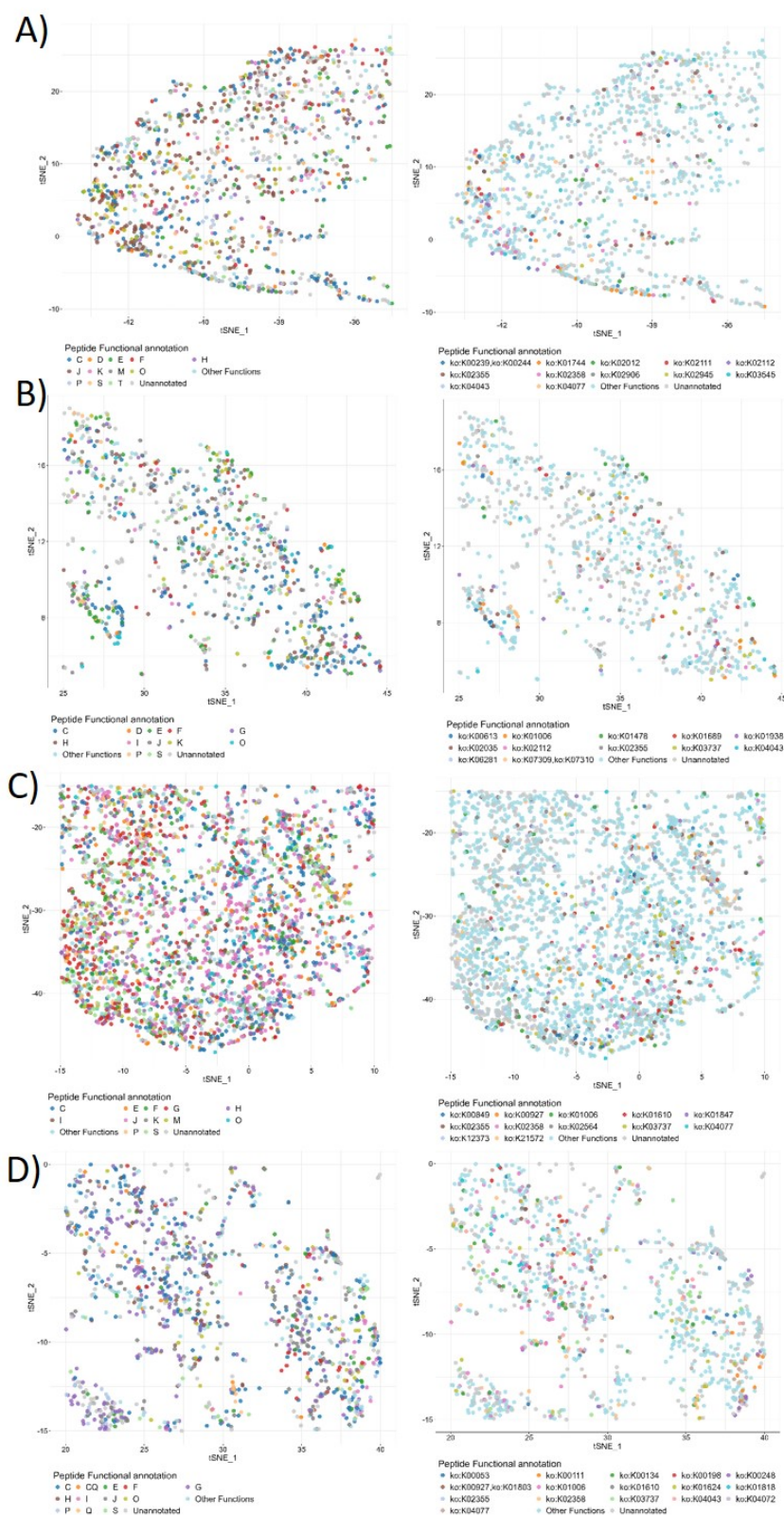


Supplementary Figure S4.4. Global peptide correlation maps of V52 colored at different taxa levels. A) Phylum, B) Class, C) Order, D) Family, E) Genus, and F) Species (Genome).



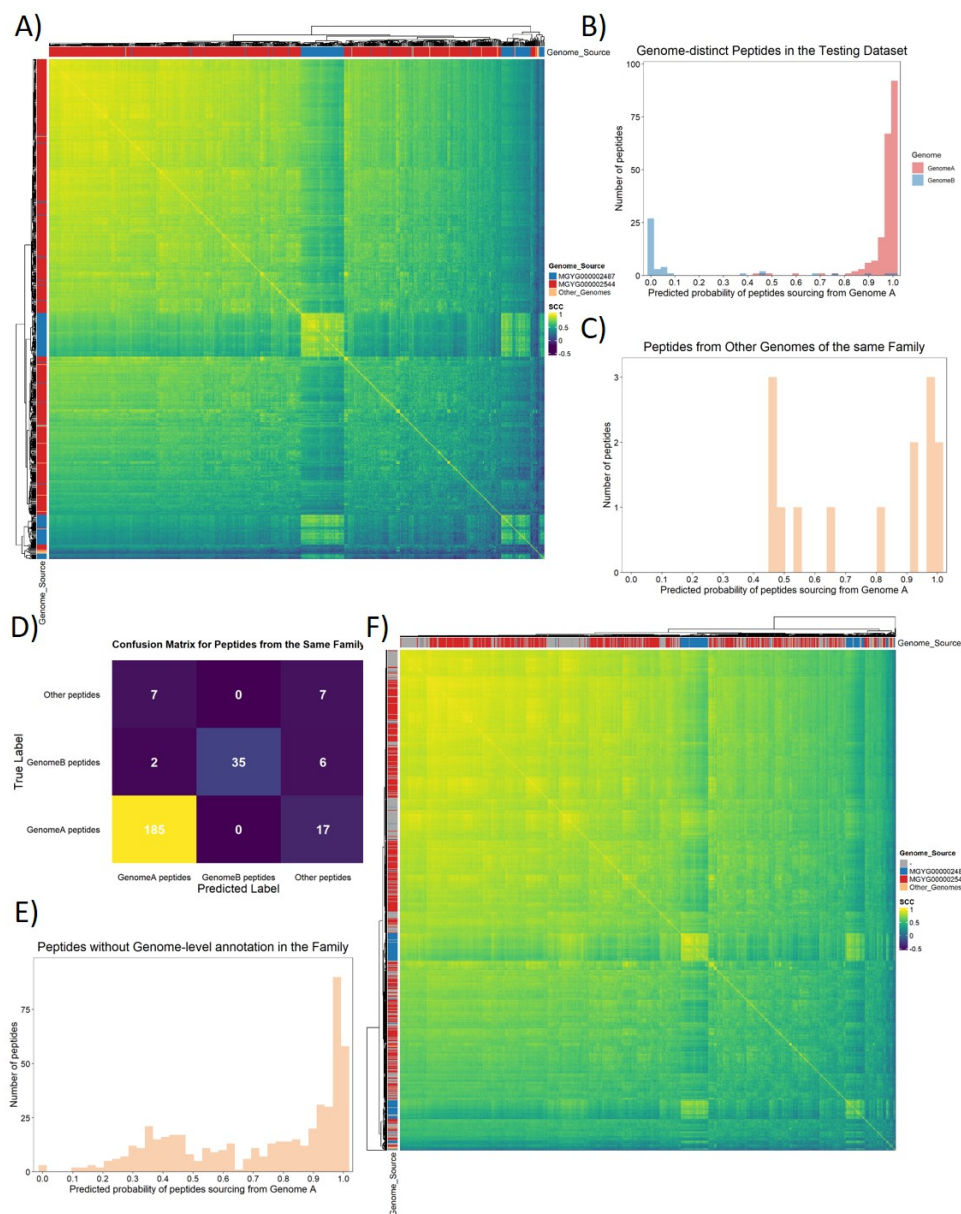
Supplementary Figure S4.5. Global peptide abundance correlation maps of V52 colored at different functional categories. A) COG family and B) KEGG ko.

4.7. SUPPLEMENTARY MATERIALS

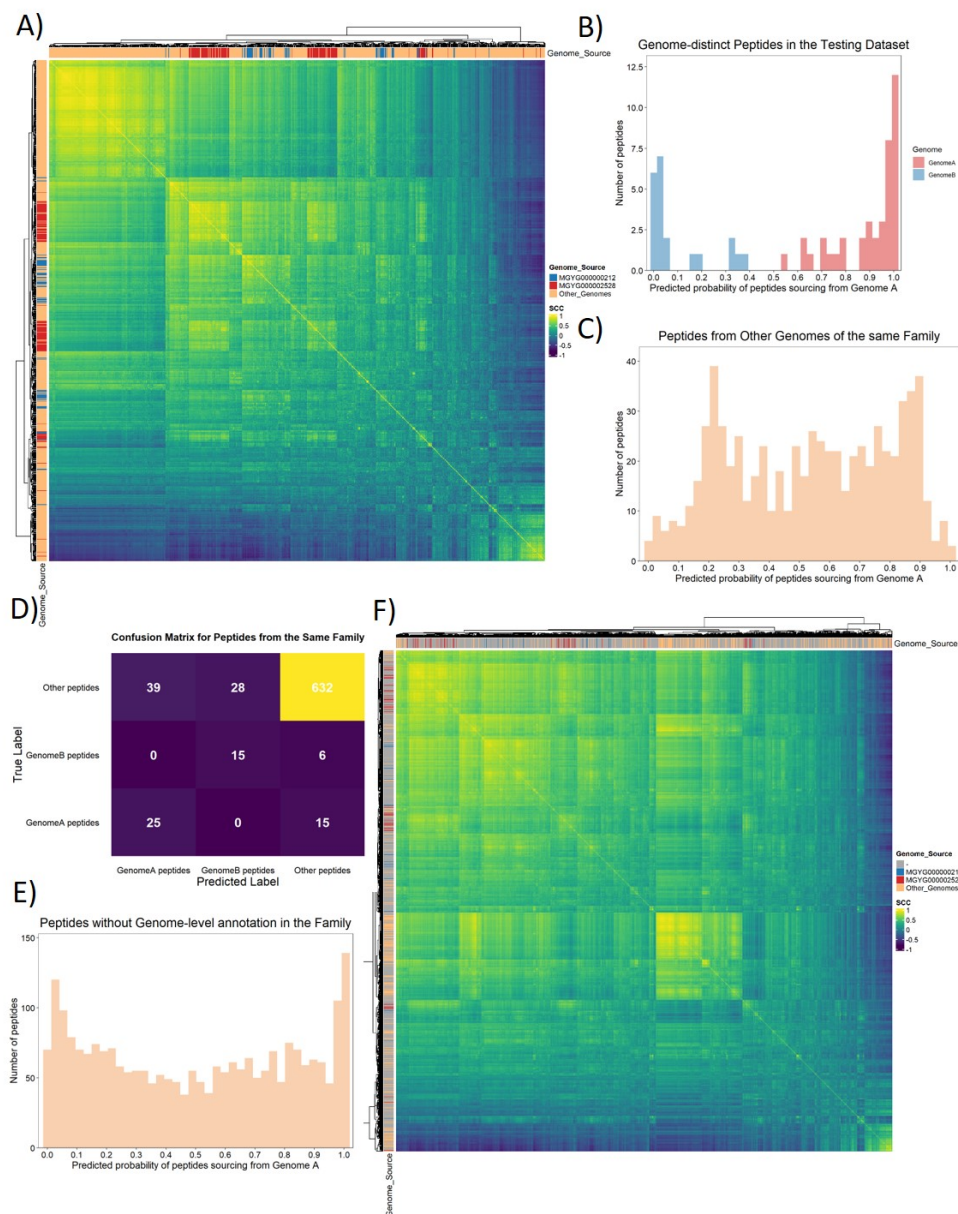


Supplementary Figure S4.6. Zoomed-in peptide abundance correlation maps of V52 colored at functional categories: COG family (left) and KEGG ko (right). Zoomed in maps for bacteria families: A) Burkholderiaceae, B) Eggerthellaceae, C) Bacteroidaceae, and D) Lachnospiraceae.

4.7. SUPPLEMENTARY MATERIALS

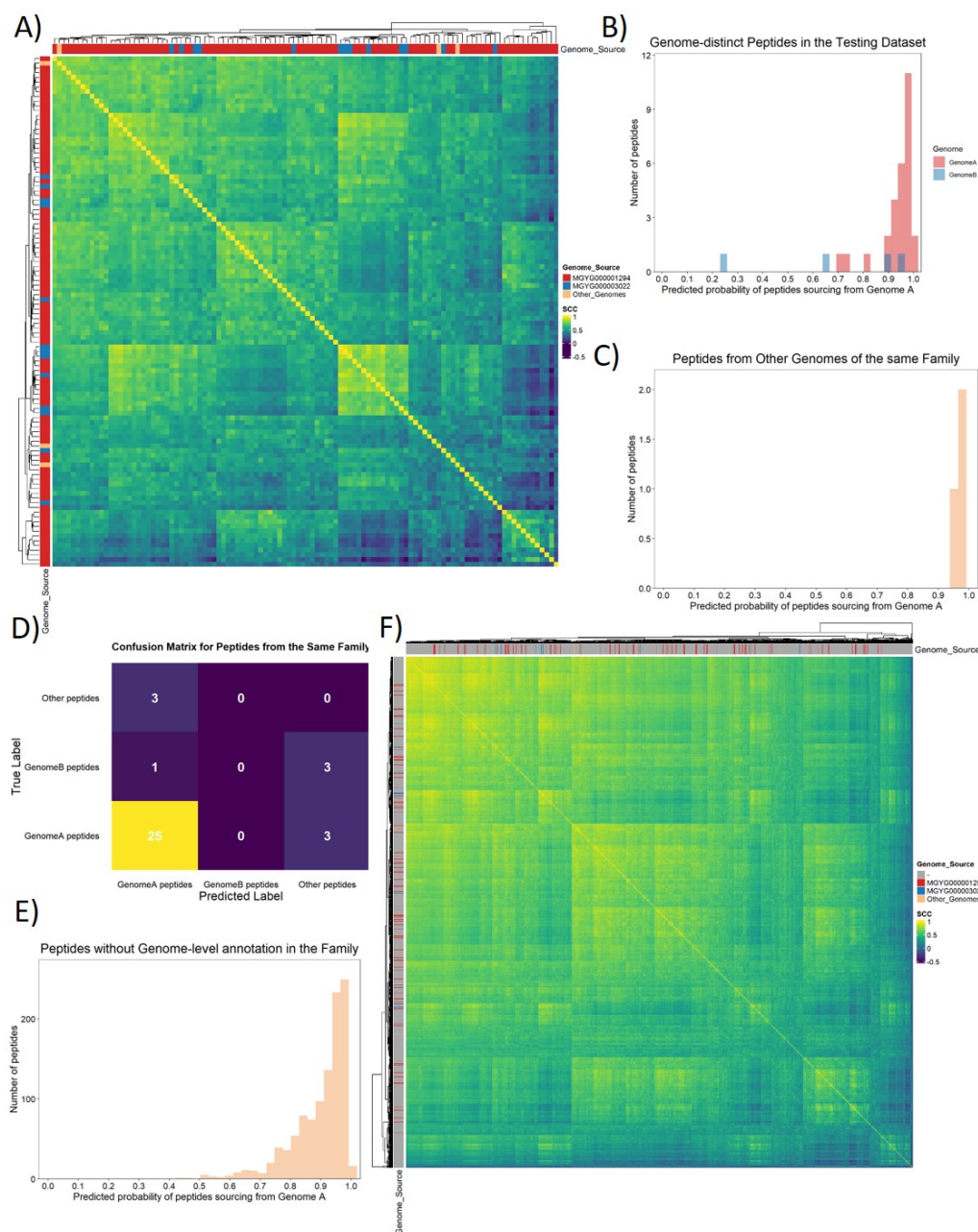


Supplementary Figure S4.7. Applying peptide abundance correlation for peptide taxonomic assignments in the family Eggerthellaceae. A) Heatmap of the SCC matrix for genome-distinct peptides from different genomes within the family. B) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on genome-distinct peptides in the test dataset. C) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on peptides from other genomes within the same family. D) Confusion matrix from the Random Forest model applied to the combined dataset of genome-distinct peptides in the test dataset and peptides from other genomes within the family. E) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on genome-unannotated peptides within the family. F) Heatmap of the SCC matrix for both genome-distinct peptides and genome-unannotated peptides within the family. Genome A, MGYG000002544; Genome B, MGYG000002487.

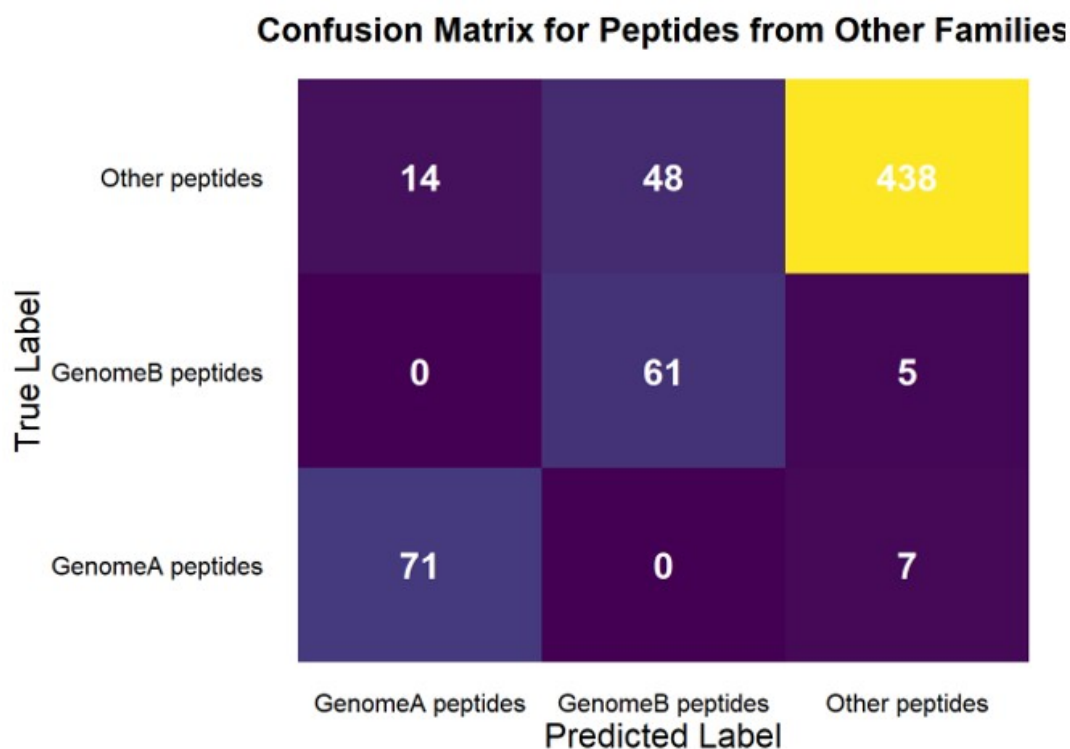
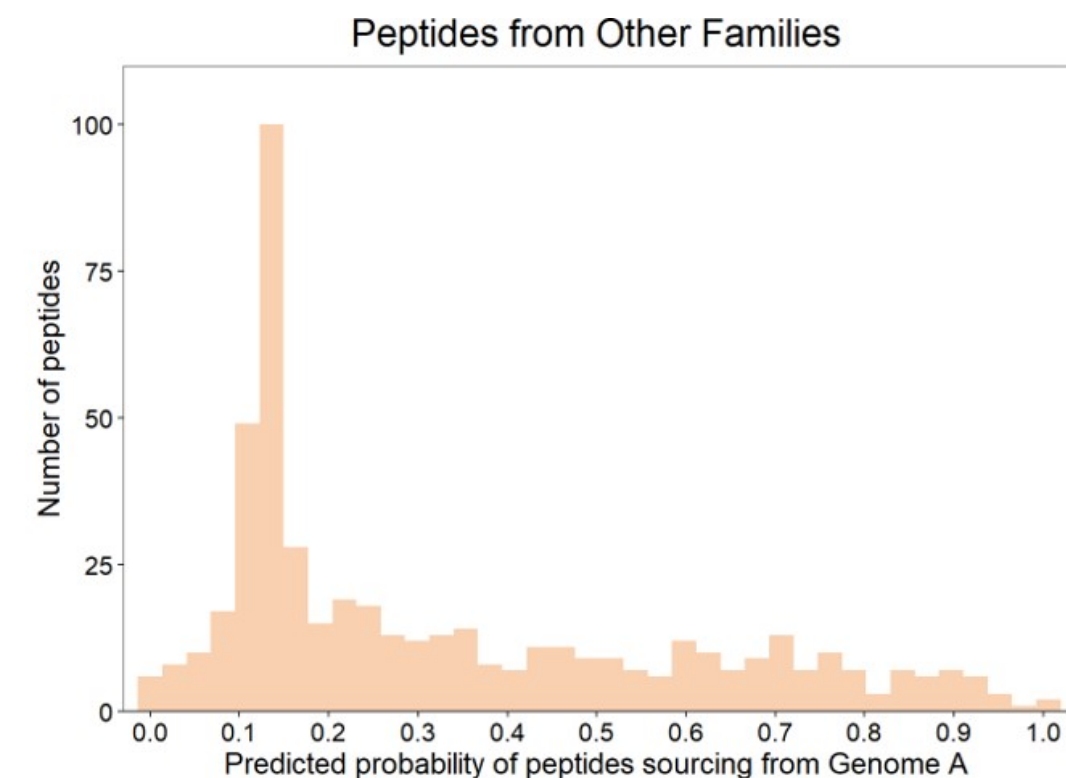


Supplementary Figure S4.8. Applying peptide abundance correlation for peptide taxonomic assignments in the family Lachnospiraceae. A) Heatmap of the SCC matrix for genome-distinct peptides from different genomes within the family. B) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on genome-distinct peptides in the test dataset. C) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on peptides from other genomes within the same family. D) Confusion matrix from the Random Forest model applied to the combined dataset of genome-distinct peptides in the test dataset and peptides from other genomes within the family. E) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on genome-unannotated peptides within the family. F) Heatmap of the SCC matrix for both genome-distinct peptides and genome-unannotated peptides within the family. Genome A, MGYG000002528; Genome B, MGYG000000212.

4.7. SUPPLEMENTARY MATERIALS

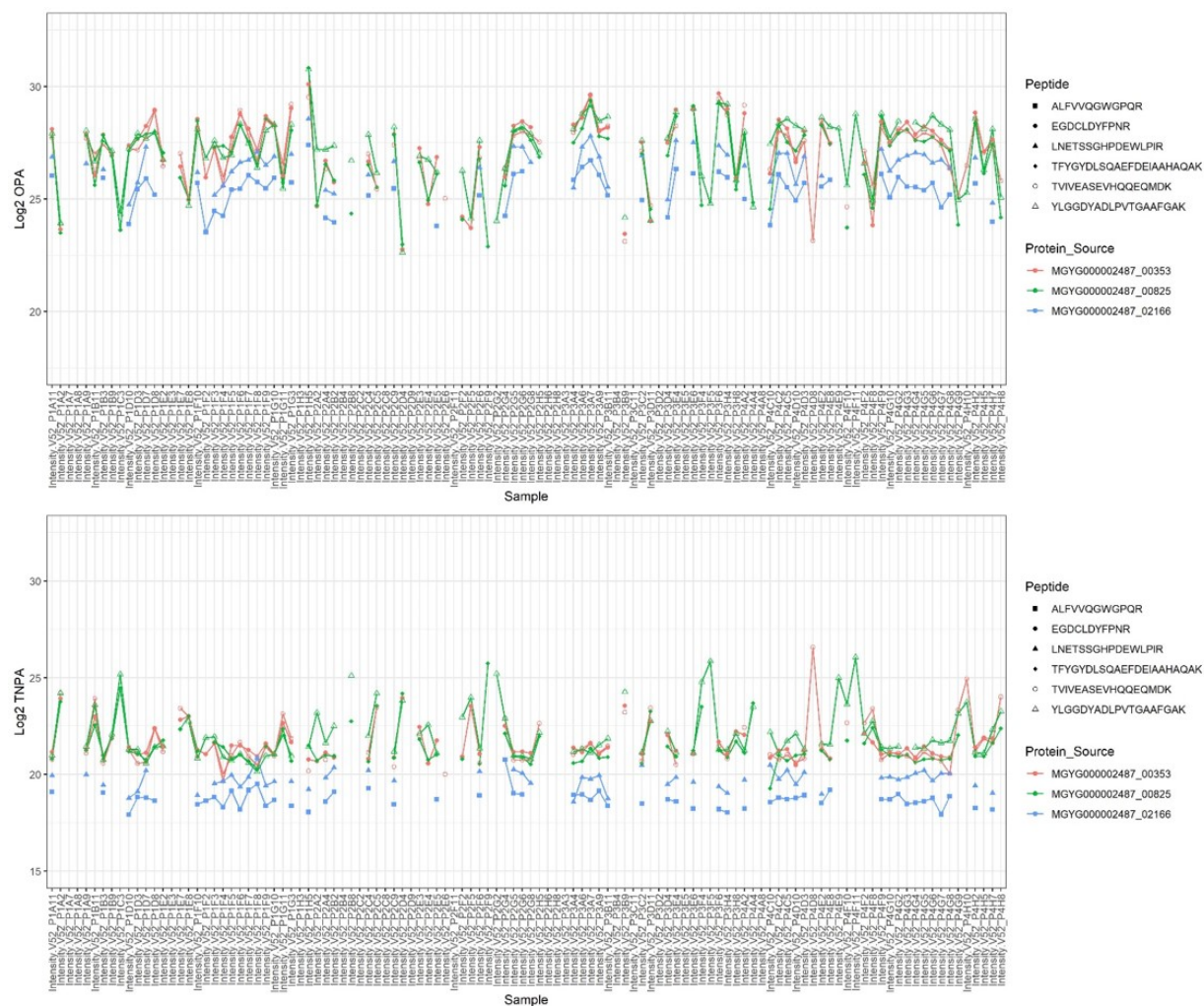


Supplementary Figure S4.9. Applying peptide abundance correlation for peptide taxonomic assignments in the family Burkholderiaceae. A) Heatmap of the SCC matrix for genome-distinct peptides from different genomes within the family. B) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on genome-distinct peptides in the test dataset. C) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on peptides from other genomes within the same family. D) Confusion matrix from the Random Forest model applied to the combined dataset of genome-distinct peptides in the test dataset and peptides from other genomes within the family. E) Distribution of predicted probabilities for peptides sourcing from Genome A, focusing on genome-unannotated peptides within the family. F) Heatmap of the SCC matrix for both genome-distinct peptides and genome-unannotated peptides within the family. Genome A, MGYG000001294; Genome B, MGYG000003022.

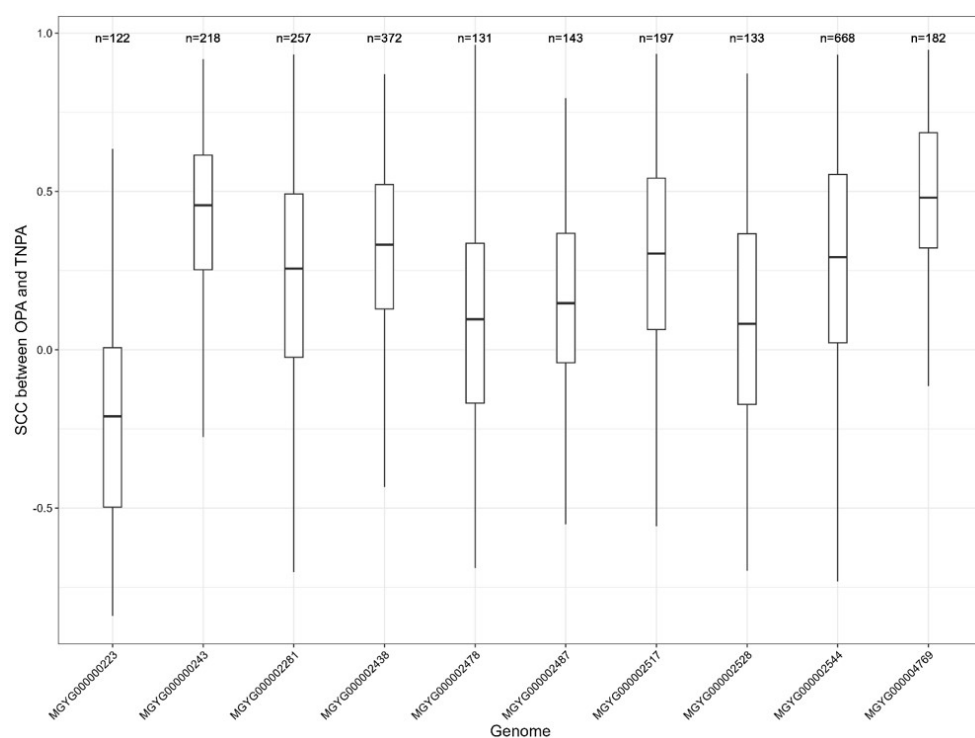


Supplementary Figure S4.10. Peptide taxonomic source prediction results for testing peptides from families other than Bacteroidaceae. The prediction model was trained using genome-distinct peptides from Bacteroidaceae. Genome A, MGYG000002281; Genome B, MGYG000000243.

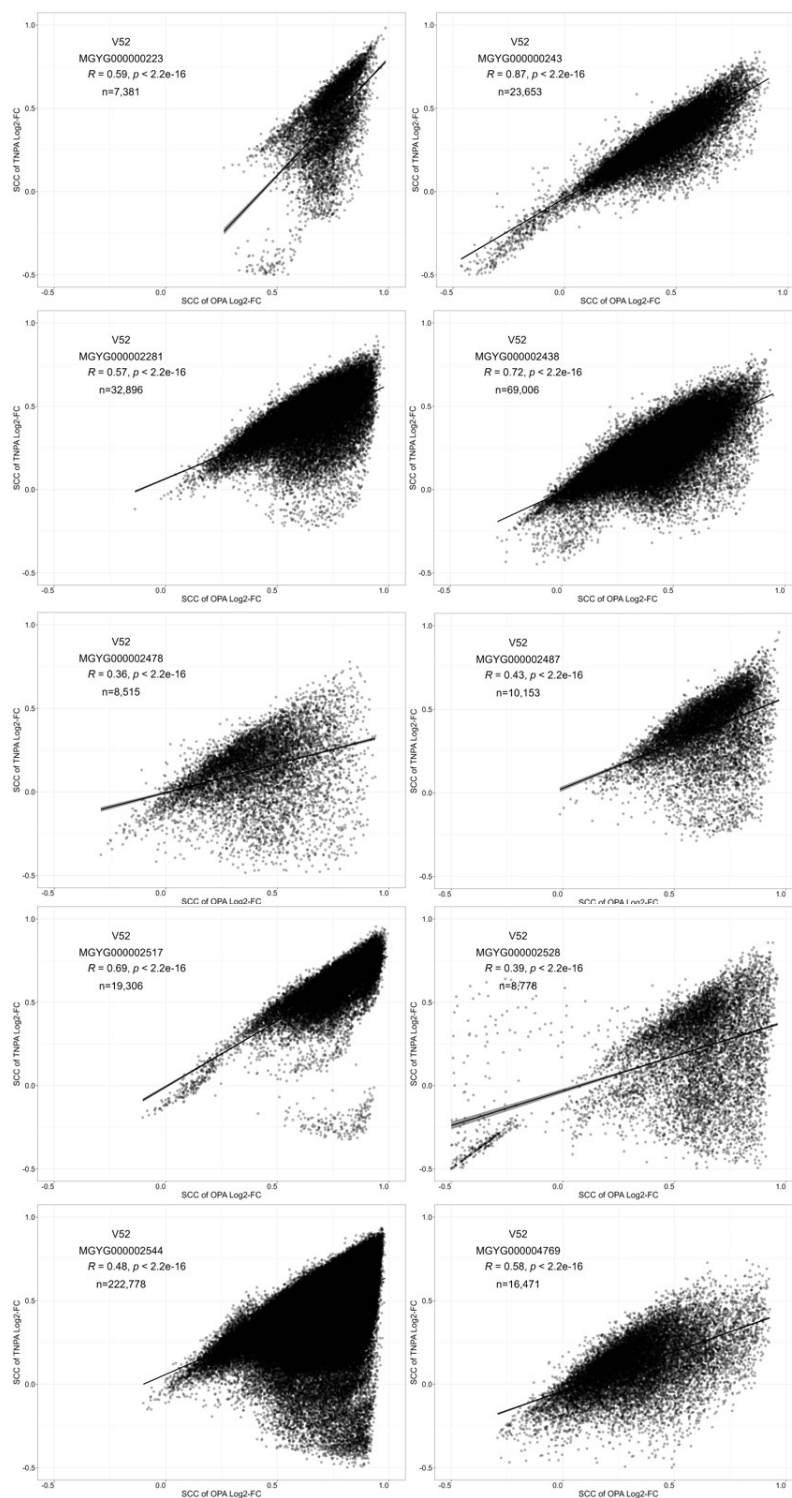
4.7. SUPPLEMENTARY MATERIALS



Supplementary Figure S4.11. Changes in original peptide abundance (OPA, top panel) and taxa-normalized peptide abundance (TNPA, bottom panel) across different treatments for peptides from three proteins. These proteins, originating from the same genome, have distinct functions: MGYG000002487_00353 (inosamine-phosphate amidinotransferase 1), MGYG000002487_00825 (putative protein), and MGYG000002487_02166 (putative dimethyl sulfoxide reductase chain YnfE).



Supplementary Figure S4.12. Distribution of SCC between OPA and TNPA for all peptides from each of the top 10 genomes. n represents the number of peptides from each species.

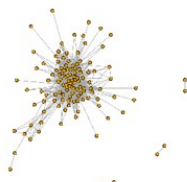


Supplementary Figure S4.13. The correlation between the SCC of peptide abundance log2 FC based on TNPA and SCC of peptide abundance log2 FC based on OPA.

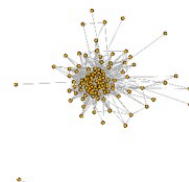
MGYG000000223



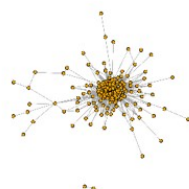
MGYG000000243



MGYG0000002281



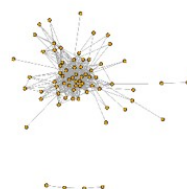
MGYG0000002438



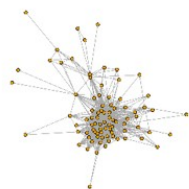
MGYG0000002478



MGYG0000002487



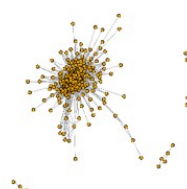
MGYG0000002517



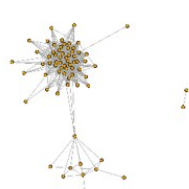
MGYG0000002528



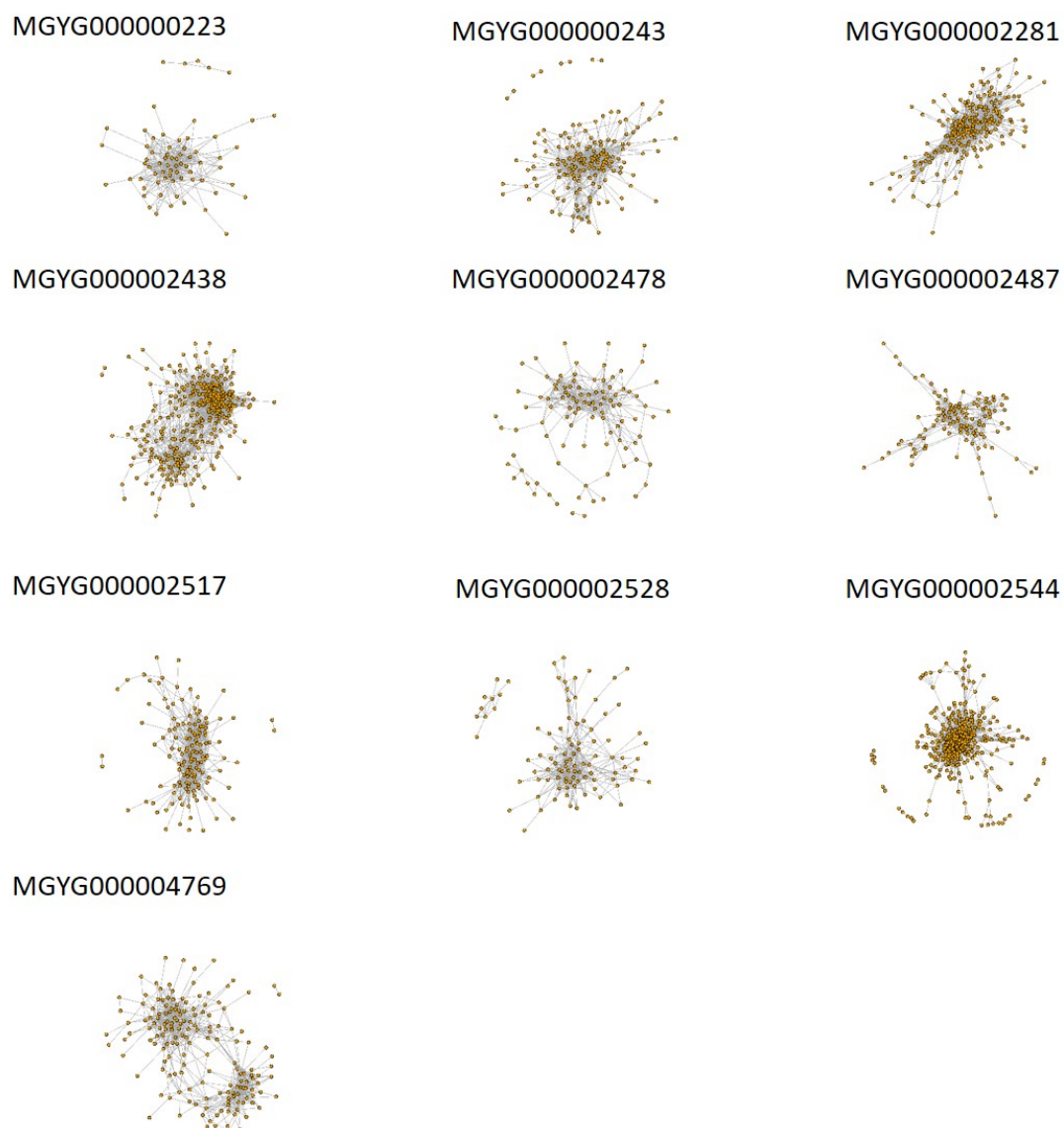
MGYG0000002544



MGYG0000004769

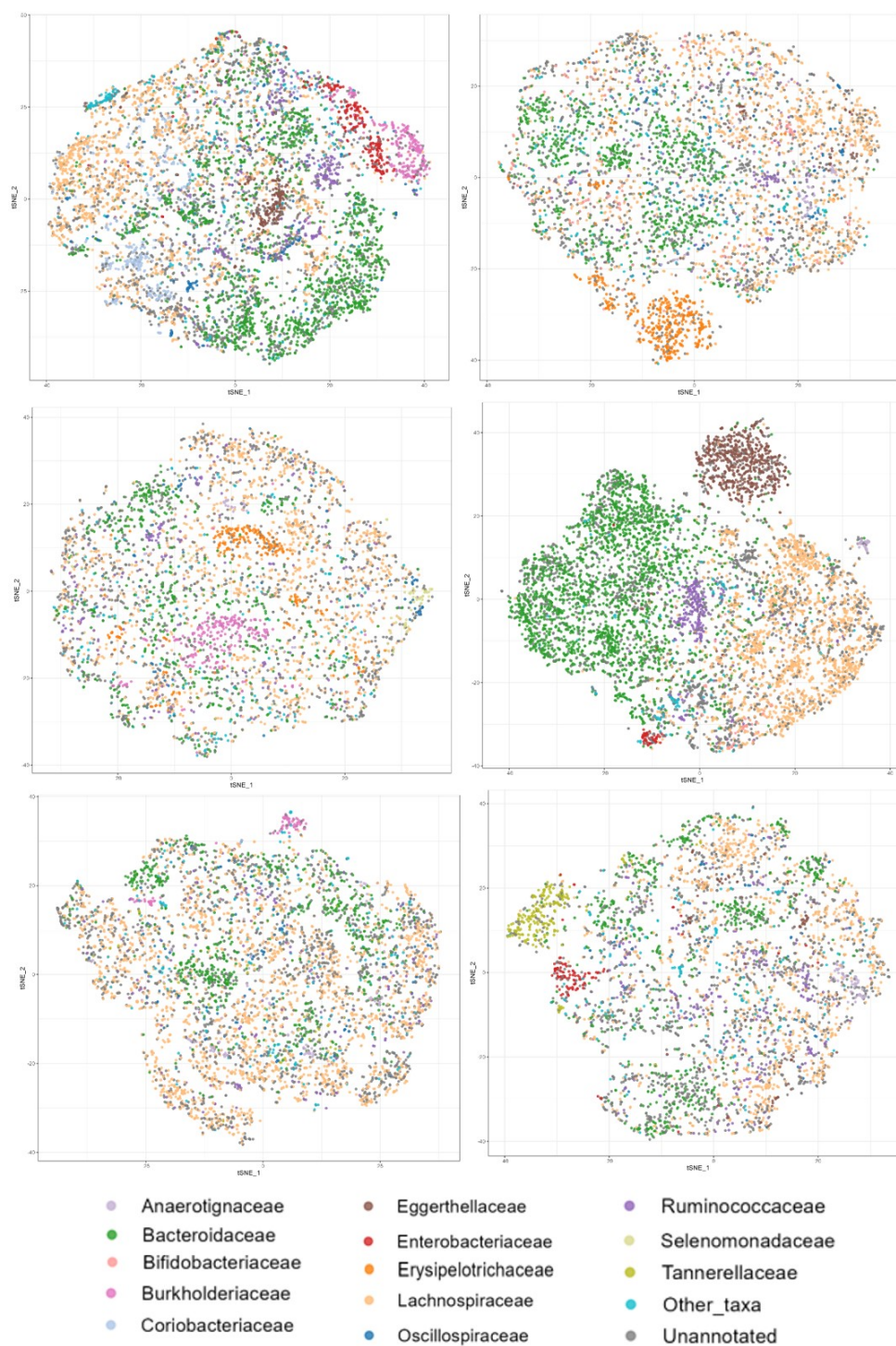


Supplementary Figure S4.14. Single-species peptide abundance correlation networks based on OPA for the ten genomes with the most identified peptides from V52. Each network was constructed using peptide pairs with the top 5% SCC from each species.



Supplementary Figure S4.15. Single-species peptide abundance correlation networks based on TNPA for the ten genomes with the most identified peptides from V52. Each network was constructed using peptide pairs with the top 5% SCC from each species.

4.7. SUPPLEMENTARY MATERIALS



Supplementary Figure S4.16. Global peptide correlation maps colored by peptide family-level taxonomic annotations from six individuals in another metaproteomics dataset (PXD012724).

4.7. SUPPLEMENTARY MATERIALS

4.7.2 SUPPLEMENTARY TABLES

Supplementary Table S4.1. The list of metaproteomics raw files analyzed in this study. The table could be accessed by: <https://doi.org/10.1101/2025.03.13.642886>

GENERAL DISCUSSION

Summary of Major Findings

Metaproteomics has emerged as a pivotal tool for investigating the human gut microbiome, which is closely associated with human health and disease. However, MS-based (meta) proteomics identifies and quantifies peptides directly, rather than proteins. This thesis presents a comprehensive exploration of human gut metaproteomics, transitioning from traditional protein-level analyses to innovative peptide-level approaches. Specifically, in addition to a traditional protein-level metaproteomics study to evaluate the functional effects of commonly used sweeteners on *ex vivo* human gut microbiome (Chapter 2), the construction of a core peptide database (Chapter 3) and application of peptide abundance correlation (Chapter 4) enhance peptide-level human gut metaproteomics analysis and offer new approaches for human gut microbiome studies.

Protein-Level Sweetener Study: Achievements and Challenges

In Chapter 2, the effects of 21 sugar substitute sweeteners on *ex vivo* human gut microbiomes were investigated using traditional protein-level metaproteomic analyses. The study revealed that seven sweeteners, primarily sugar alcohols, along with three positive controls, significantly altered the metaproteomic profiles of the microbiomes. Hierarchical clustering based on functional responses categorized the sweeteners into two distinct clusters, the noncaloric artificial sweetener (NAS) cluster and the carbohydrate (CHO) cluster. This work represents the first comprehensive metaproteomic assessment of the functional impacts of such a broad range of sweeteners on the human gut microbiome.

While these findings provide important insights, further validation in animal models would help assess the physiological relevance of the observed microbial responses. In addition, several limitations were also noted in the current study. Most notably, proteome coverage was limited, with only approximately 8,300 peptides and 3,400 protein groups quantified on average per sample. These limitations could potentially be addressed through more advanced metaproteomic approaches, including the use of improved mass spectrometry.

try instrumentation and acquisition strategies such as data-independent acquisition (DIA), which could enhance both proteome coverage and quantification reproducibility.

Moreover, the use of peptide-level analysis may provide complementary insights. Although the taxonomic analysis in this study was already based on distinctive peptides, the broader interpretation remained constrained within a protein-centric framework. For example, in functional analysis, the function of the top protein in a protein group was used to represent the entire group, which inherently limits the resolution of protein-level functional annotation. To address the limitations of protein-level analysis, the subsequent chapters of this thesis introduce peptide-centric metaproteomic strategies aimed at reducing information loss and improving the interpretability of human gut metaproteomics data.

MetaPep Core Peptide Database: Achievements and Challenges

In Chapter 3, a core peptide database, MetaPep, was constructed by compiling peptide-spectrum matches (PSMs) from publicly available human gut metaproteomics datasets. This study demonstrated the feasibility of using a curated core peptide database as an alternative to traditional protein databases in metaproteomic workflows. MetaPep enabled rapid peptide identification, achieving a comparable number of identified peptides and taxonomic profiling results when benchmarked against comprehensive protein database searches using MetaLab-MAG (Cheng et al., 2023).

In addition to peptide-level functional and taxonomic analysis methods, such as UniPept (Gurdeep Singh et al., 2019) and pepFunk (Simopoulos et al., 2020), MetaPep advances full-stage peptide-level metaproteomic analysis by supporting database searching at the peptide level—an initial and critical step in metaproteomic data processing. Moreover, the construction of MetaPep also provided an opportunity to systematically assess peptide-level proteome coverage and to reflect on the current landscape of human gut metaproteomics research, as further discussed in a related perspective article (Sun et al., 2024).

Nevertheless, the current version of MetaPep incorporates data from only 15 studies available prior to early 2022. With the rapid accumulation of new datasets and the application of more advanced mass spectrometry technologies, the number of confidently identified peptides is expected to increase substantially (Supplementary Figure S3.2). Consequently, the database will require continuous updates and optimization to balance comprehensiveness with search efficiency. Future versions of MetaPep should also be regularly updated with taxonomic and functional annotations as the database expands.

Peptide Abundance Correlations: Achievements and Challenges

In Chapter 4, peptide abundance correlations within a metaproteomics dataset derived from *in vitro* cultured human gut microbiomes treated with over 100 different drugs were investigated. The analysis revealed that peptides from the same taxon exhibited higher abundance correlations across various perturbations. This approach enabled genome-level taxonomic assignments for a larger number of peptides and provided insights into the functional relationships between peptides, including those derived from proteins of previously unknown function.

Interestingly, taxa exhibiting similar peptide abundance patterns across perturbations may reflect co-regulated behavior within microbial guilds—groups of bacteria that display consistent co-abundance and collectively contribute to the same ecological functions (Wu et al., 2024, 2021). In contrast, limited correlations between peptides from functionally related proteins across different taxa may result from microbial functional redundancy (Li et al., 2023). For example, a decrease in a specific function in one taxon may be compensated by increased expression of the same function in another, thereby masking peptide-level correlations despite shared functionality. Together, these findings highlight peptide abundance correlation as a complementary approach for unraveling microbial community dynamics and functional relationships in the human gut.

Nonetheless, several limitations remain. First, stringent correlation thresholds are needed to identify meaningful peptide associations, and these thresholds depend heavily on the size and depth of the dataset. Second, functional prediction of previously uncharacterized proteins remains constrained by the limited availability of larger and deeper metaproteomic datasets. To support broader application, this approach requires further validation across diverse datasets and experimental conditions, which will improve its robustness, scalability, and generalizability.

Future Directions and Summary

Looking ahead, although this thesis has advanced metaproteomics from protein-level to peptide-level analyses, it is worth noting that some information is still lost at the peptide-spectrum match (PSM) level. Spectra-level analyses, which retain the full complexity of the MS/MS data, may uncover novel insights that are missed when aggregating data at higher levels. However, interpreting such complex spectral data will likely require the application of advanced deep learning techniques capable of capturing and disentangling informative patterns.

In parallel, the ongoing development of high-resolution mass spectrometry technologies and data-independent acquisition (DIA) methods will continue to improve proteome coverage, quantification accuracy, and reproducibility. When combined with increasingly large and diverse metaproteomic datasets, these improvements will support the continued expansion and refinement of the core peptide database. They will also enable functional annotation of a larger number of previously uncharacterized proteins through analysis of peptide abundance correlations.

Moreover, improvements in sampling strategies, particularly the ability to collect samples from multiple sites along the human gastrointestinal tract, will facilitate more spatially resolved analyses of the gut microbiome. Such efforts will provide a more comprehensive understanding of host–microbiome interactions across different gut regions.

In summary, this thesis demonstrates that, beyond traditional protein-level metaproteomic analysis, peptide-level approaches can significantly accelerate database search, maintain comparable peptide identification rates, and extract additional biological insights by analyzing peptide abundance correlations. Collectively, these findings position metaproteomics, especially through peptide-level analysis, as a valuable and evolving tool for advancing human gut microbiome research.

REFERENCES

- Aakko, J., Pietila, S., Suomi, T., Mahmoudian, M., Toivonen, R., Kouvonen, P., Rokka, A., Hanninen, A., & Elo, L. L. (2020). Data-independent acquisition mass spectrometry in metaproteomics of gut microbiota-implementation and computational analysis. *Journal of Proteome Research*, *19*(1), 432–436. <https://doi.org/10.1021/acs.jproteome.9b00606>
- Abou-Donia, M. B., El-Masry, E. M., Abdel-Rahman, A. A., McLendon, R. E., & Schiffman, S. S. (2008). Splenda alters gut microflora and increases intestinal p-glycoprotein and cytochrome p-450 in male rats. *Journal of Toxicology and Environmental Health, Part A*, *71*(21), 1415–1429. <https://doi.org/10.1080/15287390802328630>
- Abuhelwa, A. Y., Foster, D. J. R., & Upton, R. N. (2016). A quantitative review and meta-models of the variability and factors affecting oral drug absorption—part i: Gastrointestinal pH. *The AAPS Journal*, *18*(5), 1309–1321. <https://doi.org/10.1208/s12248-016-9952-8>
- Aebersold, R., Blum, L., Gillet, L., Malmström, L., Navarro, P., Röst, H., Rosenberger, G., & Sigurdur, S. (2016). Msproteomicstools. <https://github.com/msproteomicstools/msproteomicstools>
- Aguilar, F., Charrondiere, U. R., Dusemund, B., Galtier, P., Gilbert, J., Gott, D. M., Grilli, S., Guertler, R., Koenig, J., Lambre, C., Larsen, J.-C., Leblanc, J.-C., Mortensen, A., Parent-Massin, D., Pratt, I., Rietjens, I. M. C. M., Stankovic, I., Tobback, P., Verguieva, T., ... Nutrient, E. P. F. A. (2009). Scientific opinion on the use of polyglycitol syrup as a food additive. *EFSA Journal*, *7*(12). <https://doi.org/10.2903/j.efsa.2009.1413>
- Aguilar, F., Crebelli, R., & Alessandro, D. (2015). Scientific opinion on the safety of the proposed extension of use of erythritol (e 968) as a food additive. *EFSA Journal*, (13). <https://doi.org/10.2903/j.efsa.2015.4033>
- Aguilar, F., Autrup, H., Barlow, S., Castle, L., & Crebelli, R. (2007). Neotame as a sweetener and flavour enhancer - scientific opinion of the panel on food additives, flavourings, processing aids and materials in contact with food. *EFSA Journal*, *581*, 1–43. <https://doi.org/10.2903/j.efsa.2007.581>
- Aguilar, F., Crebelli, R., Dusemund, B., Galtier, P., Gott, D., Gundert-Remy, U., Koenig, J., lambre, C., Leblanc, J.-C., Mosesso, P., Mortensen, A., Oskarsson, A., Parent-Massin, D., Rose, M., Stankovic, I., Tobback, P., Waalkens-Berendsen, I., Woutersen, R., Wright, M., & Nutrient, E. P. F. A. (2013). Scientific opinion on the safety of advantame for the proposed uses as a food additive. *EFSA Journal*, *11*(7). <https://doi.org/10.2903/j.efsa.2013.3301>
- Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1), 105–114. <https://doi.org/10.1038/s41587-020-0603-3>
- Armengaud, J. (2023). Metaproteomics to understand how microbiota function: The crystal ball predicts a promising future. *Environmental Microbiology*, *25*(1), 115–125. <https://doi.org/10.1111/1462-2920.16238>
- Audain, E., Uszkoreit, J., Sachsenberg, T., Pfeuffer, J., Liang, X., Hermjakob, H., Sanchez, A., Eisenacher, M., Reinert, K., Tabb, D. L., Kohlbacher, O., & Perez-Riverol, Y. (2017). In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *Journal of Proteomics*, *150*, 170–182. <https://doi.org/10.1016/j.jprot.2016.08.002>
- Baquero, F., & Nombela, C. (2012). The microbiome as a human organ. *Clinical Microbiology and Infection*, *18*, 2–4. <https://doi.org/10.1111/j.1469-0691.2012.03916.x>
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *3*(1), 361–362. <https://doi.org/10.1609/icwsm.v3i1.13937>

REFERENCES

- Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L. G., . . . Consortium, U. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*, D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- Beards, E., Tuohy, K., & Gibson, G. (2010). A human volunteer study to assess the impact of confectionery sweeteners on the gut microbiota composition. *British Journal of Nutrition*, *104*(5), 701–708. <https://doi.org/10.1017/S0007114510001078>
- Belkaid, Y., & Hand, T. (2014). Role of the microbiota in immunity and inflammation. *Cell*, *157*(1), 121–141. <https://doi.org/10.1016/j.cell.2014.03.011>
- Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., . . . Schlöter, M. (2020). Microbiome definition re-visited: Old concepts and new challenges. *Microbiome*, *8*(1), 103. <https://doi.org/10.1186/s40168-020-00875-0>
- Beyter, D., Lin, M. S., Yu, Y., Pieper, R., & Bafna, V. (2018). ProteoStorm: An ultrafast metaproteomics database search framework. *Cell Systems*, *7*(4), 463–467.e6. <https://doi.org/10.1016/j.cels.2018.08.009>
- Bik, E. M. (2009). Composition and function of the human-associated microbiota. *Nutrition Reviews*, *67*, S164–S171. <https://doi.org/10.1111/j.1753-4887.2009.00237.x>
- Blakeley-Ruiz, J. A., & Kleiner, M. (2022). Considerations for constructing a protein sequence database for metaproteomics [Publisher: Elsevier BV]. *Computational and Structural Biotechnology Journal*, *20*, 937–952. <https://doi.org/10.1016/j.csbj.2022.01.018>
- Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, *65*(1), 23–28. <https://doi.org/10.1007/BF02294183>
- Brandman, O., & Meyer, T. (2008). Feedback loops shape cellular signals in space and time. *Science*, *322*(5900), 390–395. <https://doi.org/10.1126/science.1160617>
- Brown, J. M., & Hazen, S. L. (2018). Microbial modulation of cardiovascular disease. *Nature Reviews Microbiology*, *16*(3), 171–181. <https://doi.org/10.1038/nrmicro.2017.149>
- Caballero, B., Finglas, P., & Toldra, F. (2003). *Encyclopedia of food sciences and nutrition* (2nd ed). Academic Press.
- Calabrese, F. M., Porrelli, A., Vacca, M., Comte, B., Nimptsch, K., Pinart, M., Pischon, T., Pujos-Guillot, E., & De Angelis, M. (2021). Metaproteomics approach and pathway modulation in obesity and diabetes: A narrative review. *Nutrients*, *14*(1), 47. <https://doi.org/10.3390/nu14010047>
- Callister, S. J., Fillmore, T. L., Nicora, C. D., Shaw, J. B., Purvine, S. O., Orton, D. J., White, R. A., Moore, R. J., Burnet, M. C., Nakayasu, E. S., Payne, S. H., Jansson, J. K., & Paša-Tolić, L. (2018). Addressing the challenge of soil metaproteome complexity by improving metaproteome depth of coverage through two-dimensional liquid chromatography. *Soil Biology and Biochemistry*, *125*, 290–299. <https://doi.org/10.1016/j.soilbio.2018.07.018>
- Canfora, E. E., Jocken, J. W., & Blaak, E. E. (2015). Short-chain fatty acids in control of body weight and insulin sensitivity. *Nature Reviews Endocrinology*, *11*(10), 577–591. <https://doi.org/10.1038/nrendo.2015.128>
- Cani, P. D. (2018). Human gut microbiome: Hopes, threats and promises. *Gut*, *67*(9), 1716–1725. <https://doi.org/10.1136/gutjnl-2018-316723>
- Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The carbohydrate-active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Research*, *37*, D233–D238. <https://doi.org/10.1093/nar/gkn663>
- Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., . . . Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, *30*(10), 918–920. <https://doi.org/10.1038/nbt.2377>

REFERENCES

- Chase L. S., S., & Priya S., S. (2024). I'm walking into spiderwebs: Making sense of proteinprotein interaction data. *Journal of Proteome Research*, 23(8), 2723–2732. <https://doi.org/10.1021/acs.jproteome.3c00892>
- Chen, C., Liao, J., Xia, Y., Liu, X., Jones, R., Haran, J., McCormick, B., Sampson, T. R., Alam, A., & Ye, K. (2022). Gut microbiota regulate alzheimer's disease pathologies and cognitive disorders via PUFA-associated neuroinflammation. *Gut*, 71(11), 2233–2252. <https://doi.org/10.1136/gutjnl-2021-326269>
- Cheng, K., Ning, Z., Li, L., Zhang, X., Serrana, J. M., Mayne, J., & Figeys, D. (2023). MetaLab-MAG: A metaproteomic data analysis platform for genome-level characterization of microbiomes from the metagenome-assembled genomes database. *Journal of Proteome Research*, 22(2), 387–398. <https://doi.org/10.1021/acs.jproteome.2c00554>
- Cheng, K., Ning, Z., Zhang, X., Li, L., Liao, B., Mayne, J., & Figeys, D. (2020). MetaLab 2.0 enables accurate post-translational modifications profiling in metaproteomics. *Journal of the American Society for Mass Spectrometry*, 31(7), 1473–1482. <https://doi.org/10.1021/jasms.0c00083>
- Cheng, K., Ning, Z., Zhang, X., Li, L., Liao, B., Mayne, J., Stintzi, A., & Figeys, D. (2017). MetaLab: An automated pipeline for metaproteomic data analysis. *Microbiome*, 5. <https://doi.org/10.1186/s40168-017-0375-2>
- Chi, H., Liu, C., Yang, H., Zeng, W.-F., Wu, L., Zhou, W.-J., Wang, R.-M., Niu, X.-N., Ding, Y.-H., Zhang, Y., Wang, Z.-W., Chen, Z.-L., Sun, R.-X., Liu, T., Tan, G.-M., Dong, M.-Q., Xu, P., Zhang, P.-H., & He, S.-M. (2018). Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nature Biotechnology*, 36(11), 1059+. <https://doi.org/10.1038/nbt.4236>
- Cohen, L., Curhan, G., & Forman, J. (2012). Association of sweetened beverage intake with incident hypertension. *Journal of General Internal Medicine*, 27(9), 1127–1134. <https://doi.org/10.1007/s11606-012-2069-6>
- Colaert, N., Degroeve, S., Helsens, K., & Martens, L. (2011). Analysis of the resolution limitations of peptide identification algorithms. *Journal of Proteome Research*, 10(12), 5555–5561. <https://doi.org/10.1021/pr200913a>
- Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., Van Leeuwen, J., Van Dyk, N., Lin, Z.-Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., ... Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306), aaf1420. <https://doi.org/10.1126/science.aaf1420>
- Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), 1367–1372. <https://doi.org/10.1038/nbt.1511>
- Cussotto, S., Strain, C. R., Fouhy, F., Strain, R. G., Peterson, V. L., Clarke, G., Stanton, C., Dinan, T. G., & Cryan, J. F. (2019). Differential effects of psychotropic drugs on microbiome composition and gastrointestinal function. *Psychopharmacology*, 236(5), 1671–1685. <https://doi.org/10.1007/s00213-018-5006-5>
- Dalile, B., Van Oudenhove, L., Vervliet, B., & Verbeke, K. (2019). The role of short-chain fatty acids in microbiota–gut–brain communication. *Nature Reviews Gastroenterology & Hepatology*, 16(8), 461–478. <https://doi.org/10.1038/s41575-019-0157-3>
- Darzi, Y., Letunic, I., Bork, P., & Yamada, T. (2018). iPath3.0: Interactive pathways explorer v3. *Nucleic Acids Research*, 46, W510–W513. <https://doi.org/10.1093/nar/gky299>
- de Vos, W. M., Tilg, H., Van Hul, M., & Cani, P. D. (2022). Gut microbiome and health: Mechanistic insights. *Gut*, 71(5), 1020–1032. <https://doi.org/10.1136/gutjnl-2021-326789>
- Deehan, E. C., Mocanu, V., & Madsen, K. L. (2024). Effects of dietary fibre on metabolic health and obesity. *Nature Reviews Gastroenterology & Hepatology*, 21(5), 301–318. <https://doi.org/10.1038/s41575-023-00891-z>
- Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2020). DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, 17(1), 41+. <https://doi.org/10.1038/s41592-019-0638-x>

REFERENCES

- Deutsch, E. W., Bandeira, N., Sharma, V., Perez-Riverol, Y., Carver, J. J., Kundu, D. J., García-Seisdedos, D., Jarnuczak, A. F., Hewapathirana, S., Pullman, B. S., Wertz, J., Sun, Z., Kawano, S., Okuda, S., Watanabe, Y., Hermjakob, H., MacLean, B., MacCoss, M. J., Zhu, Y., . . . Vizcaíno, J. A. (2019). The ProteomeXchange consortium in 2020: Enabling ‘big data’ approaches in proteomics. *Nucleic Acids Research*, D1145–D1152. <https://doi.org/10.1093/nar/gkz984>
- Deutsch, E. W., Sun, Z., Campbell, D., Kusebauch, U., Chu, C. S., Mendoza, L., Shteynberg, D., Omenn, G. S., & Moritz, R. L. (2015). State of the human proteome in 2014/2015 as viewed through PeptideAtlas: Enhancing accuracy and coverage through the AtlasProphet. *Journal of Proteome Research*, 14(9), 3461–3473. <https://doi.org/10.1021/acs.jproteome.5b00500>
- Dixon, P. (2003). VEGAN, a package of r functions for community ecology. *Journal of Vegetation Science*, 14(6), 927–930. [https://doi.org/10.1658/1100-9233\(2003\)014\[0927:VAPORF\]2.0.CO;2](https://doi.org/10.1658/1100-9233(2003)014[0927:VAPORF]2.0.CO;2)
- Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., & Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, 107(26), 11971–11975. <https://doi.org/10.1073/pnas.1002601107>
- Donaldson, G. P., Lee, S. M., & Mazmanian, S. K. (2016). Gut biogeography of the bacterial microbiota. *Nature Reviews Microbiology*, 14(1), 20–32. <https://doi.org/10.1038/nrmicro3552>
- Dos Santos, S. J., Pakzad, Z., Albert, A. Y. K., Elwood, C. N., Grabowska, K., Links, M. G., Hutcheon, J. A., Maan, E. J., Manges, A. R., Dumonceaux, T. J., Hodgson, Z. G., Lyons, J., Mitchell-Foster, S. M., Gantt, S., Joseph, K., Van Schalkwyk, J. E., Hill, J. E., & Money, D. M. (2023). Maternal vaginal microbiome composition does not affect development of the infant gut microbiome in early life. *Frontiers in Cellular and Infection Microbiology*, 13, 1144254. <https://doi.org/10.3389/fcimb.2023.1144254>
- Duan, H., Cheng, K., Ning, Z., Li, L., Mayne, J., Sun, Z., & Figeys, D. (2022). Assessing the dark field of metaproteome. *Analytical Chemistry*, 94(45), 15648–15654. <https://doi.org/10.1021/acs.analchem.2c02452>
- Duan, H., Zhang, X., & Figeys, D. (2023). An emerging field: Post-translational modification in microbiome. *Proteomics*, 23(3), 2100389. <https://doi.org/10.1002/pmic.202100389>
- EFSA, E. C. (2018). Sugars and sweeteners. https://knowledge4policy.ec.europa.eu/health-promotion-knowledge-gateway/sugars-sweeteners_en
- Fan, Y., & Pedersen, O. (2021). Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, 19(1), 55–71. <https://doi.org/10.1038/s41579-020-0433-9>
- Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D. T., Manara, S., Zolfo, M., Beghini, F., Bertorelli, R., De Sanctis, V., Bariletti, I., Canto, R., Clementi, R., Cologna, M., Crifò, T., Cusumano, G., . . . Segata, N. (2018). Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host & Microbe*, 24(1), 133–145.e5. <https://doi.org/10.1016/j.chom.2018.06.005>
- Flynn, K. J., Ruffin, M. T., Turgeon, D. K., & Schloss, P. D. (2018). Spatial variation of the native colon microbiota in healthy adults. *Cancer Prevention Research*, 11(7), 393–402. <https://doi.org/10.1158/1940-6207.CAPR-17-0370>
- Foppa, C., Rizkala, T., Repici, A., Hassan, C., & Spinelli, A. (2024). Microbiota and IBD: Current knowledge and future perspectives. *Digestive and Liver Disease*, 56(6), 911–922. <https://doi.org/10.1016/j.dld.2023.11.015>
- Fowler, S. P., Williams, K., Resendez, R. G., Hunt, K. J., Hazuda, H. P., & Sterns, M. P. (2008). Fueling the obesity epidemic? artificially sweetened beverage use and long-term weight gain. *Obesity*, 16(8), 1894–1900. <https://doi.org/10.1038/oby.2008.284>
- Franzosa, E. A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X. C., & Huttenhower, C. (2015). Sequencing and beyond: Integrating molecular ‘omics’ for microbial community profiling. *Nature Reviews Microbiology*, 13(6), 360–372. <https://doi.org/10.1038/nrmicro3451>
- Gardner, C., Wylie-Rosett, J., Gidding, S. S., Johnson, R. K., Reader, D., Lichtenstein, A. H., Young, C. C. D., & Assoc, A. D. (2012). Nonnutritive sweeteners: Current use and health perspectives a scientific statement from the american heart association and the american diabetes association. *Circulation*, 126(4), 509–519. <https://doi.org/10.1161/CIR.0b013e31825c42ee>

REFERENCES

- Garrido-Rodriguez, M., Zirngibl, K., Ivanova, O., Lobentanzer, S., & Saez-Rodriguez, J. (2022). Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. *Molecular Systems Biology*, *18*(7), e11036. <https://doi.org/10.15252/msb.202211036>
- Gavin, P. G., Mullaney, J. A., Loo, D., Cao, K.-A. L., Gottlieb, P. A., Hill, M. M., Zipris, D., & Hamilton-Williams, E. E. (2018). Intestinal metaproteomics reveals host-microbiota interactions in subjects at risk for type 1 diabetes. *Diabetes Care*, *41*(10), 2178–2186. <https://doi.org/10.2337/dc18-0777>
- Gomaa, E. Z. (2020). Human gut microbiota/microbiome in health and diseases: A review. *Antonie van Leeuwenhoek*, *113*(12), 2019–2040. <https://doi.org/10.1007/s10482-020-01474-7>
- Grenga, L., Pible, O., Miotello, G., Culotta, K., Ruat, S., Roncato, M., Gas, F., Bellanger, L., Claret, P., Dunyach-Remy, C., Laureillard, D., Sotto, A., Lavigne, J., & Armengaud, J. (2022). Taxonomical and functional changes in COVID-19 faecal microbiome could be related to SARS-CoV-2 faecal load. *Environmental Microbiology*, *24*(9), 4299–4316. <https://doi.org/10.1111/1462-2920.16028>
- Gribun, A., Nitzan, Y., Pechatnikov, I., Hershkovits, G., & Katcoff, D. (2003). Molecular and structural characterization of the HMP-AB gene encoding a pore-forming protein from a clinical isolate of *acinetobacter baumannii*. *Current Microbiology*, *47*(5), 434–443. <https://doi.org/10.1007/s00284-003-4050-4>
- Guo, P., Zhang, K., Ma, X., & He, P. (2020). Clostridium species as probiotics: Potentials and challenges. *Journal of Animal Science and Biotechnology*, *11*(1). <https://doi.org/10.1186/s40104-019-0402-1>
- Gurdeep Singh, R., Tanca, A., Palomba, A., Van der Jeugt, F., Verschaffelt, P., Uzzau, S., Martens, L., Dawyndt, P., & Mesuere, B. (2019). Unipept 4.0: Functional analysis of metaproteome data. *Journal of Proteome Research*, *18*(2), 606–615. <https://doi.org/10.1021/acs.jproteome.8b00716>
- Gábor, C., Tamás, N., Vincent, T., Szabolcs, H., Fabio, Z., Daniel, N., Kirill, M., Maëlle, S., & Michael, A. (2025). Igraph: Network analysis and visualization. <https://doi.org/10.32614/CRAN.package.igraph>
- Gómez-Varela, D., Xian, F., Grundtner, S., Sondermann, J. R., Carta, G., & Schmidt, M. (2023a, March 16). Expanding the characterization of microbial ecosystems using DIA-PASEF metaproteomics (bioRxiv). *bioRxiv*. <https://doi.org/10.1101/2023.03.16.532922>
- Gómez-Varela, D., Xian, F., Grundtner, S., Sondermann, J. R., Carta, G., & Schmidt, M. (2023b). Increasing taxonomic and functional characterization of host-microbiome interactions by DIA-PASEF metaproteomics. *Frontiers in Microbiology*, *14*, 1258703. <https://doi.org/10.3389/fmicb.2023.1258703>
- Hamer, H. M., Jonkers, D., Venema, K., Vanhoutvin, S., Troost, F. J., & Brummer, R.-J. (2007). Review article: The role of butyrate on colonic function. *Alimentary Pharmacology & Therapeutics*, *27*(2), 104–119. <https://doi.org/10.1111/j.1365-2036.2007.03562.x>
- Haynes, W. A., Tomczak, A., & Khatri, P. (2018). Gene annotation bias impedes biomedical research. *Scientific Reports*, *8*(1), 1362. <https://doi.org/10.1038/s41598-018-19333-x>
- He, F., Zhang, T., Xue, K., Fang, Z., Jiang, G., Huang, S., Li, K., Gu, Z., Shi, H., Zhang, Z., Zhu, H., Lin, L., Li, J., Xiao, F., Shan, H., Yan, R., Li, X., & Yan, Z. (2021). Fecal multi-omics analysis reveals diverse molecular alterations of gut ecosystem in COVID-19 patients. *Analytica Chimica Acta*, *1180*, 338881. <https://doi.org/10.1016/j.aca.2021.338881>
- Health Canada, H. C. (2017). List of permitted sweeteners. <https://www.canada.ca/en/health-canada/services/food-nutrition/food-safety/food-additives/lists-permitted/9-sweeteners>
- GRAS Notice 627: *Siraitia grosvenorii* Swingle (Luo Han Guo) fruit juice concentrate (2016). <https://www.hfpappexternal.fda.gov/scripts/fdcc/index.cfm?set=grasnotices&id=627>
- Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., Wampach, L., Schneider, J. G., Hogan, A., De Beaufort, C., & Wilmes, P. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology*, *2*(1), 16180. <https://doi.org/10.1038/nmicrobiol.2016.180>
- Heintz-Buschart, A., & Wilmes, P. (2018). Human gut microbiome: Function matters. *Trends in Microbiology*, *26*(7), 563–574. <https://doi.org/10.1016/j.tim.2017.11.002>
- Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271. <https://doi.org/10.1016/j.csda.2006.11.025>
- Hillman, E. T., Lu, H., Yao, T., & Nakatsu, C. H. (2017). Microbial ecology along the gastrointestinal tract. *Microbes and environments*, *32*(4), 300–313. <https://doi.org/10.1264/jsme2.ME17017>

REFERENCES

- Ho, H.-e., & Bunyavanich, S. (2018). Role of the microbiome in food allergy. *Current Allergy and Asthma Reports*, 18(4), 27. <https://doi.org/10.1007/s11882-018-0780-z>
- Hou, K., Wu, Z.-X., Chen, X.-Y., Wang, J.-Q., Zhang, D., Xiao, C., Zhu, D., Koya, J. B., Wei, L., Li, J., & Chen, Z.-S. (2022). Microbiota in health and diseases. *Signal Transduction and Targeted Therapy*, 7(1), 135. <https://doi.org/10.1038/s41392-022-00974-4>
- Hou, M., Huang, J., Jia, T., Guan, Y., Yang, F., Zhou, H., Huang, P., Wang, J., Yang, L., & Dai, L. (2023). Deep profiling of the proteome dynamics of *Pseudomonas aeruginosa* reference strain PAO1 under different growth conditions. *Journal of Proteome Research*, 22(6), 1747–1761. <https://doi.org/10.1021/acs.jproteome.2c00785>
- Huang, L., Zhang, H., Wu, P., Entwistle, S., Li, X., Yohe, T., Yi, H., Yang, Z., & Yin, Y. (2018). dbCAN-seq: A database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Research*, 46, D516–D521. <https://doi.org/10.1093/nar/gkx894>
- Hubler, S. L., Kumar, P., Mehta, S., Easterly, C., Johnson, J. E., Jagtap, P. D., & Griffin, T. J. (2020). Challenges in peptide-spectrum matching: A robust and reproducible statistical framework for removing low-accuracy, high-scoring hits. *Journal of Proteome Research*, 19(1), 161–173. <https://doi.org/10.1021/acs.jproteome.9b00478>
- Jagtap, P. D., Blakely, A., Murray, K., Stewart, S., Kooren, J., Johnson, J. E., Rhodus, N. L., Rudney, J., & Griffin, T. J. (2015). Metaproteomic analysis using the galaxy framework. *Proteomics*, 15(20), 3553–3565. <https://doi.org/10.1002/pmic.201500074>
- James, K., & Muñoz-Muñoz, J. (2022). Computational network inference for bacterial interactomics (J. A. Gilbert, Ed.) [Publisher: American Society for Microbiology]. *mSystems*, 7(2). <https://doi.org/10.1128/msystems.01456-21>
- Javier Ruiz-Ojeda, F., Plaza-Diaz, J., Jose Saez-Lara, M., & Gil, A. (2019). Effects of sweeteners on the gut microbiota: A review of experimental studies and clinical trials. *Advances in Nutrition*, 10(1), S31–S48. <https://doi.org/10.1093/advances/nmy037>
- Jesse, K., & Laurens van der, M. (2023). Rtsne: T-distributed stochastic neighbor embedding using a barnes-hut implementation. <https://doi.org/10.32614/CRAN.package.Rtsne>
- Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1), 118–127. <https://doi.org/10.1093/biostatistics/kxj037>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kalantar-Zadeh, K., Berean, K. J., Burgell, R. E., Muir, J. G., & Gibson, P. R. (2019). Intestinal gases: Influence on gut disorders and the role of dietary manipulations. *Nature Reviews Gastroenterology & Hepatology*, 16(12), 733–747. <https://doi.org/10.1038/s41575-019-0193-z>
- Kennedy, M. S., & Chang, E. B. The microbiome: Composition and locations. In: *Progress in molecular biology and translational science*. Vol. 176. Elsevier, 2020, pp. 1–42. ISBN: 978-0-12-820797-0. <https://doi.org/10.1016/bs.pmbts.2020.08.013>
- Kim, J., & Lee, H. K. (2022). Potential role of the gut microbiome in colorectal cancer progression. *Frontiers in Immunology*, 12, 807648. <https://doi.org/10.3389/fimmu.2021.807648>
- Kleiner, M. (2019). Metaproteomics: Much more than measuring gene expression in microbial communities. *mSystems*, 4(3), e00115–19. <https://doi.org/10.1128/mSystems.00115-19>
- Kleiner, M., Dong, X., Hinzke, T., Wippler, J., Thorson, E., Mayer, B., & Strous, M. (2018). Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. *Proceedings of the National Academy of Sciences*, 115(24). <https://doi.org/10.1073/pnas.1722325115>
- Kleiner, M., Thorson, E., Sharp, C. E., Dong, X., Liu, D., Li, C., & Strous, M. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. *Nature Communications*, 8(1), 1558. <https://doi.org/10.1038/s41467-017-01544-x>

REFERENCES

- Kolmeder, C. A., Ritari, J., Verdam, F. J., Muth, T., Keskitalo, S., Varjosalo, M., Fuentes, S., Greve, J. W., Buurman, W. A., Reichl, U., Rapp, E., Martens, L., Palva, A., Salonen, A., Rensen, S. S., & De Vos, W. M. (2015). Colonic metaproteomic signatures of active bacteria and the host in obesity. *Proteomics*, *15*(20), 3544–3552. <https://doi.org/10.1002/pmic.201500049>
- Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, *14*(5), 513+. <https://doi.org/10.1038/NMETH.4256>
- Kostic, A., Chun, E., Robertson, L., Glickman, J., Gallini, C., Michaud, M., Clancy, T., Chung, D., Lochhead, P., Hold, G., El-Omar, E., Brenner, D., Fuchs, C., Meyerson, M., & Garrett, W. (2013). *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host & Microbe*, *14*(2), 207–215. <https://doi.org/10.1016/j.chom.2013.07.007>
- Kustatscher, G., Collins, T., Gingras, A.-C., Guo, T., Hermjakob, H., Ideker, T., Lilley, K. S., Lundberg, E., Marcotte, E. M., Ralser, M., & Rappsilber, J. (2022). Understudied proteins: Opportunities and challenges for functional proteomics. *Nature Methods*, *19*(7), 774–779. <https://doi.org/10.1038/s41592-022-01454-x>
- Kustatscher, G., Grabowski, P., Schrader, T. A., Passmore, J. B., Schrader, M., & Rappsilber, J. (2019). Co-regulation map of the human proteome enables identification of protein functions. *Nature Biotechnology*, *37*(11), 1361–1371. <https://doi.org/10.1038/s41587-019-0298-5>
- Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., & Aebersold, R. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, *7*(5), 655–667. <https://doi.org/10.1002/pmic.200600625>
- Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., & Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. *Nature Methods*, *5*(10), 873–875. <https://doi.org/10.1038/NMETH.1254>
- Laska, M. N., Murray, D. M., Lytle, L. A., & Harnack, L. J. (2012). Longitudinal associations between key dietary behaviors and weight gain over time: Transitions through the adolescent years. *Obesity*, *20*(1), 118–125. <https://doi.org/10.1038/oby.2011.179>
- Lederberg, J., & McCray, A. T. (2001). 'ome sweet 'omics— a genealogical treasury of words. *The scientist*, *15*(7), 8–8.
- Lenhart, A., & Chey, W. D. (2017). A systematic review of the effects of polyols on gastrointestinal health and irritable bowel syndrome. *Advances in Nutrition*, *8*(4), 587–596. <https://doi.org/10.3945/an.117.015560>
- Leo, B., Adele, C., Andy, L., & Matthew, W. (2024). randomForest: Breiman and cutlers random forests for classification and regression. <https://doi.org/10.32614/CRAN.package.randomForest>
- Levitan, O., Ma, L., Giovannelli, D., Burleson, D. B., McCaffrey, P., Vala, A., & Johnson, D. A. (2023). The gut microbiome—does stool represent right? *Heliyon*, *9*(3), e13602. <https://doi.org/10.1016/j.heliyon.2023.e13602>
- Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., Nielsen, T., Juncker, A. S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., Liang, S., ... Consortium, M. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, *32*(8), 834–841. <https://doi.org/10.1038/nbt.2942>
- Li, L., Abou-Samra, E., Ning, Z., Zhang, X., Mayne, J., Wang, J., Cheng, K., Walker, K., Stintzi, A., & Figeys, D. (2019). An in vitro model maintaining taxon-specific functional activities of the gut microbiome. *Nature Communications*, *10*. <https://doi.org/10.1038/s41467-019-12087-8>
- Li, L., Ning, Z., Cheng, K., Zhang, X., Simopoulos, C. M. A., & Figeys, D. (2022). iMetaLab suite: A one-stop toolset for metaproteomics. *iMeta*, *1*(2). <https://doi.org/10.1002/imt.2.25>
- Li, L., Ning, Z., Zhang, X., Mayne, J., Cheng, K., Stintzi, A., & Figeys, D. (2020). RapidAIM: A culture- and metaproteomics-based rapid assay of individual microbiome responses to drugs. *Microbiome*, *8*(1), 33. <https://doi.org/10.1186/s40168-020-00806-z>
- Li, L., Simopoulos, C. M., Mayne, J., Ning, Z., Zhang, X., Hamada, M., Butcher, J., Serrana, J. M., Wang, L., Cheng, K., Qin, H., Walker, K., Zhang, X., Stintzi, A., & Figeys, D. (2025, February 14). Systematic metaproteomics mapping reveals functional and ecological landscapes of human gut microbiota responses to therapeutic drugs. <https://doi.org/10.1101/2025.02.13.637346>

REFERENCES

- Li, L., Wang, T., Ning, Z., Zhang, X., Butcher, J., Serrana, J. M. M., Simopoulos, C. M. A., Mayne, J., Stintzi, A., Mack, D. R. R., Liu, Y.-Y., & Figeys, D. (2023). Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics. *Nature Communications*, *14*(1). <https://doi.org/10.1038/s41467-023-39149-2>
- Liauchonak, I., Qorri, B., Dawoud, F., Riat, Y., & Szewczuk, M. R. (2019). Non-nutritive sweeteners and their implications on the development of metabolic syndrome. *Nutrients*, *11*(3). <https://doi.org/10.3390/nu11030644>
- Livesey, G. (2003). Health potential of polyols as sugar replacers, with emphasis on low glycaemic properties. *Nutrition Research Reviews*, *16*(2), 163–191. <https://doi.org/10.1079/NRR200371>
- Lloyd-Price, J., Abu-Ali, G., & Huttenhower, C. (2016). The healthy human microbiome. *Genome Medicine*, *8*(1), 51. <https://doi.org/10.1186/s13073-016-0307-y>
- Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., . . . Huttenhower, C. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, *569*(7758), 655–662. <https://doi.org/10.1038/s41586-019-1237-9>
- Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., Brady, A., Creasy, H. H., McCracken, C., Giglio, M. G., McDonald, D., Franzosa, E. A., Knight, R., White, O., & Huttenhower, C. (2017). Strains, functions and dynamics in the expanded human microbiome project. *Nature*, *550*(7674), 61–66. <https://doi.org/10.1038/nature23889>
- Lobb, B., Kurtz, D. A., Moreno-Hagelsieb, G., & Doxey, A. C. (2015). Remote homology and the functions of metagenomic dark matter. *Frontiers in Genetics*, *6*. <https://doi.org/10.3389/fgene.2015.00234>
- Lohmann, P., Schäpe, S. S., Haange, S.-B., Oliphant, K., Allen-Vercocoe, E., Jehmlich, N., & Von Bergen, M. (2020). Function is what counts: How microbial community complexity affects species, proteome and pathway coverage in metaproteomics. *Expert Review of Proteomics*, *17*(2), 163–173. <https://doi.org/10.1080/14789450.2020.1738931>
- Long, S., Yang, Y., Shen, C., Wang, Y., Deng, A., Qin, Q., & Qiao, L. (2020). Metaproteomics characterizes human gut microbiome function in colorectal cancer. *npj Biofilms and Microbiomes*, *6*(1), 14. <https://doi.org/10.1038/s41522-020-0123-4>
- Lopetuso, L. R., Scaldaferri, F., Petito, V., & Gasbarrini, A. (2013). Commensal clostridia: Leading players in the maintenance of gut homeostasis. *Gut Pathogens*, *5*. <https://doi.org/10.1186/1757-4749-5-23>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). <https://doi.org/10.1186/s13059-014-0550-8>
- Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, *489*(7415), 220–230. <https://doi.org/10.1038/nature11550>
- Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., & Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: A tutorial. *Molecular Systems Biology*, *14*(8). <https://doi.org/10.15252/msb.20178126>
- Lyn, O.-N. (2011). *Alternative sweeteners* (4th). CRC Press.
- Lynch, S. V., & Pedersen, O. (2016). The human intestinal microbiome in health and disease (E. G. Phimister, Ed.). *New England Journal of Medicine*, *375*(24), 2369–2379. <https://doi.org/10.1056/NEJMr1600266>
- Ma, B., Kankan, Z., Xiaofei, L., Weiqin, S., Zhongmin, D., Jack A., G., Philip, C. B., Karoline, F., & Jianming, X. (2018). Genetic correlation network prediction of forest soil microbial functional organization. *The ISME Journal*, (12), 2492–2505. <https://doi.org/https://doi.org/10.1038/s41396-018-0232-8>
- Macek, B., Forchhammer, K., Hardouin, J., Weber-Ban, E., Grangeasse, C., & Mijakovic, I. (2019). Protein post-translational modifications in bacteria. *Nature Reviews Microbiology*, *17*(11), 651–664. <https://doi.org/10.1038/s41579-019-0243-0>
- Maghini, D. G., Oduaran, O. H., Olubayo, L. A. I., Cook, J. A., Smyth, N., Mathema, T., Belger, C. W., Agongo, G., Boua, P. R., Choma, S. S. R., Gómez-Olivé, F. X., Kisiangani, I., Mashaba, G. R., Micklesfield, L., Mohamed, S. F., Nonterah, E. A., Norris, S., Sorgho, H., Tollman, S., . . . Hazel-

REFERENCES

- hurst, S. (2025). Expanding the human gut microbiome atlas of africa. *Nature*, 638(8051), 718–728. <https://doi.org/10.1038/s41586-024-08485-8>
- Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Raught, B., Schmitt, R., Werner, T., Kuster, B., & Aebersold, R. (2007). Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology*, 25(1), 125–131. <https://doi.org/10.1038/nbt1275>
- Marchesi, J. R., & Ravel, J. (2015). The vocabulary of microbiome research: A proposal. *Microbiome*, 3(1), 31, s40168–015–0094–5. <https://doi.org/10.1186/s40168-015-0094-5>
- Mars, R. A., Yang, Y., Ward, T., Houtti, M., Priya, S., Lekatz, H. R., Tang, X., Sun, Z., Kalari, K. R., Korem, T., Bhattarai, Y., Zheng, T., Bar, N., Frost, G., Johnson, A. J., Van Treuren, W., Han, S., Ordog, T., Grover, M., . . . Kashyap, P. C. (2020). Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell*, 182(6), 1460–1473.e17. <https://doi.org/10.1016/j.cell.2020.08.007>
- Mateus, A., Hevler, J., Bobonis, J., Kurzawa, N., Shah, M., Mitosch, K., Goemans, C. V., Helm, D., Stein, F., Typas, A., & Savitski, M. M. (2020). The functional proteome landscape of escherichia coli. *Nature*, 588(7838), 473–478. <https://doi.org/10.1038/s41586-020-3002-5>
- McAlister, G. C., Phanstiel, D., Wenger, C. D., Lee, M. V., & Coon, J. J. (2010). Analysis of tandem mass spectra by FTMS for improved large-scale proteomics with superior protein quantification. *Analytical Chemistry*, 82(1), 316–322. <https://doi.org/10.1021/ac902005s>
- McCallum, G., & Tropini, C. (2024). The gut microbiota and its biogeography. *Nature Reviews Microbiology*, 22(2), 105–118. <https://doi.org/10.1038/s41579-023-00969-0>
- Messner, C. B., Demichev, V., Muenzner, J., Aulakh, S. K., Barthel, N., Röhl, A., Herrera-Domínguez, L., Egger, A.-S., Kamrad, S., Hou, J., Tan, G., Lemke, O., Calvani, E., Szyrwiell, L., Mülleder, M., Lilley, K. S., Boone, C., Kustatscher, G., & Ralser, M. (2023). The proteomic landscape of genome-wide genetic perturbations. *Cell*, 186(9), 2018–2034.e21. <https://doi.org/10.1016/j.cell.2023.03.026>
- Mesuer, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., & Dawyndt, P. (2012). Unipept: Tryptic peptide-based biodiversity analysis of metaproteome samples. *Journal of Proteome Research*, 11(12), 5773–5780. <https://doi.org/10.1021/pr300576s>
- MetaHIT Consortium, Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., Nielsen, T., Juncker, A. S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., . . . Wang, J. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, 32(8), 834–841. <https://doi.org/10.1038/nbt.2942>
- Metcalf, A. M., Phillips, S. F., Zinsmeister, A. R., MacCarty, R. L., Beart, R. W., & Wolff, B. G. (1987). Simplified assessment of segmental colonic transit. *Gastroenterology*, 92(1), 40–47. [https://doi.org/10.1016/0016-5085\(87\)90837-7](https://doi.org/10.1016/0016-5085(87)90837-7)
- Millikin, R. J., Solntsev, S. K., Shortreed, M. R., & Smith, L. M. (2018). Ultrafast peptide label-free quantification with FlashLFQ. *Journal of Proteome Research*, 17(1), 386–391. <https://doi.org/10.1021/acs.jproteome.7b00608>
- Milo, R. (2013). What is the total number of protein molecules per cell volume? a call to rethink some published values. *BioEssays*, 35(12), 1050–1055. <https://doi.org/10.1002/bies.201300066>
- Mishra, S. P., Wang, B., Jain, S., Ding, J., Rejeski, J., Furdul, C. M., Kitzman, D. W., Taraphder, S., Brechot, C., Kumar, A., & Yadav, H. (2023). A mechanism by which gut microbiota elevates permeability and inflammation in obese/diabetic mice and human gut. *Gut*, 72(10), 1848–1865. <https://doi.org/10.1136/gutjnl-2022-327365>
- Morais, L. H., Schreiber, H. L., & Mazmanian, S. K. (2021). The gut microbiota–brain axis in behaviour and brain disorders. *Nature Reviews Microbiology*, 19(4), 241–255. <https://doi.org/10.1038/s41579-020-00460-0>
- Morton, J. T., Jin, D.-M., Mills, R. H., Shao, Y., Rahman, G., McDonald, D., Zhu, Q., Balaban, M., Jiang, Y., Cantrell, K., Gonzalez, A., Carmel, J., Frankiensztajn, L. M., Martin-Brevet, S., Berding, K., Needham, B. D., Zurita, M. F., David, M., Averina, O. V., . . . Taroncher-Oldenburg, G. (2023). Multi-level analysis of the gut–brain axis shows autism spectrum disorder-associated molecular and microbial profiles. *Nature Neuroscience*, 26(7), 1208–1217. <https://doi.org/10.1038/s41593-023-01361-0>

REFERENCES

- Mullard, A. (2023). FDA approves second microbiome-based *C. difficile* therapy. *Nature Reviews Drug Discovery*, 22, 436. <https://doi.org/10.1038/d41573-023-00081-1>
- Muth, T., Kolmeder, C. A., Salojärvi, J., Keskitalo, S., Varjosalo, M., Verdam, F. J., Rensen, S. S., Reichl, U., de Vos, W. M., Rapp, E., & Martens, L. (2015). Navigating through metaproteomics data: A logbook of database searching. *Proteomics*, 15(20), 3439–53. <https://doi.org/10.1002/pmic.201400560>
- Muth, T., Kohrs, F., Heyer, R., Benndorf, D., Rapp, E., Reichl, U., Martens, L., & Renard, B. Y. (2018). MPA portable: A stand-alone software package for analyzing metaproteome samples on the go. *Analytical Chemistry*, 90(1), 685–689. <https://doi.org/10.1021/acs.analchem.7b03544>
- Nguyen, J., Lara-Gutiérrez, J., & Stocker, R. (2021). Environmental fluctuations and their effects on microbial communities, populations and individuals. *FEMS Microbiology Reviews*, 45(4), fuaa068. <https://doi.org/10.1093/femsre/fuaa068>
- NIST, N. (2022). NIST libraries of peptide tandem mass spectra. <https://doi.org/10.18434/T4ZK5S>
- Nitzan, Y., Deutsch, E., & Pechatnikov, I. (2002). Diffusion of β -lactam antibiotics through oligomeric or monomeric porin channels of some gram-negative bacteria. *Current Microbiology*, 45(6), 446–455. <https://doi.org/10.1007/s00284-002-3778-6>
- Noble, W. S. (2015). Mass spectrometrists should search only for peptides they care about. *Nature Methods*, 12(7), 605–608. <https://doi.org/10.1038/nmeth.3450>
- Pang, Z., Chong, J., Zhou, G., Morais, D. A. d. L., Chang, L., Barrette, M., Gauthier, C., Jacques, P.-E., Li, S., & Xia, J. (2021). MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research*, 49, W388–W396. <https://doi.org/10.1093/nar/gkab382>
- Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D., Prakash, A., Frericks-Zipper, A., Eisenacher, M., Walzer, M., Wang, S., Brazma, A., & Vizcaino, J. (2022). The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, 50, D543–D552. <https://doi.org/10.1093/nar/gkab1038>
- Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Perez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., Tiwary, S., Cox, J., Audain, E., Walzer, M., ... Vizcaino, J. A. (2019). The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Research*, 47, D442–D450. <https://doi.org/10.1093/nar/gky1106>
- Piquer-Esteban, S., Ruiz-Ruiz, S., Arnau, V., Diaz, W., & Moya, A. (2022). Exploring the universal healthy human gut microbiota around the world. *Computational and Structural Biotechnology Journal*, 20, 421–433. <https://doi.org/10.1016/j.csbj.2021.12.035>
- Plubell, D. L., Käll, L., Webb-Robertson, B.-J., Bramer, L. M., Ives, A., Kelleher, N. L., Smith, L. M., Montine, T. J., Wu, C. C., & MacCoss, M. J. (2022). Putting humpty dumpty back together again: What does protein quantification mean in bottom-up proteomics? *Journal of Proteome Research*, 21(4), 891–898. <https://doi.org/10.1021/acs.jproteome.1c00894>
- Pribyl, A. L., Hugenholtz, P., & Cooper, M. A. (2025). A decade of advances in human gut microbiome-derived biotherapeutics. *Nature Microbiology*, 10(2), 301–312. <https://doi.org/10.1038/s41564-024-01896-3>
- Proctor, L., LoTempio, J., Marquitz, A., Daschner, P., Xi, D., Flores, R., Brown, L., Ranallo, R., Maruvada, P., Regan, K., Lunsford, R. D., Reddy, M., Caler, L., & Ana, N. H. M. P. (2019). A review of 10 years of human microbiome research activities at the US national institutes of health, fiscal years 2007–2016. *Microbiome*, 7. <https://doi.org/10.1186/s40168-019-0620-y>
- Pushalkar, S., Hundeyin, M., Daley, D., Zambirinis, C. P., Kurz, E., Mishra, A., Mohan, N., Aykut, B., Usyk, M., Torres, L. E., Werba, G., Zhang, K., Guo, Y., Li, Q., Akkad, N., Lall, S., Wadowski, B., Gutierrez, J., Kochen Rossi, J. A., ... Miller, G. (2018). The pancreatic cancer microbiome promotes oncogenesis by induction of innate and adaptive immune suppression. *Cancer Discovery*, 8(4), 403–416. <https://doi.org/10.1158/2159-8290.CD-17-1134>
- Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., ... Wang, J. (2012). A

REFERENCES

- metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418), 55–60. <https://doi.org/10.1038/nature11450>
- Ram, R. J., VerBerkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., Shah, M., Hettich, R. L., & Banfield, J. F. (2005). Community proteomics of a natural microbial biofilm. *Science*, 308(5730), 1915–1920. <https://doi.org/10.1126/science.1109070>
- Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. S. (2014). Protein-protein interaction detection: Methods and analysis [Publisher: Wiley]. *International Journal of Proteomics*, 2014, 1–12. <https://doi.org/10.1155/2014/147648>
- Rechenberger, J., Samaras, P., Jarzab, A., Behr, J., Frejno, M., Djukovic, A., Sanz, J., González-Barberá, E. M., Salavert, M., López-Hontangas, J. L., Xavier, K. B., Debrauwer, L., Rolain, J.-M., Sanz, M., Garcia-Garcera, M., Wilhelm, M., Ubeda, C., & Kuster, B. (2019). Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant enterobacteriaceae. *Proteomes*, 7(1), 2. <https://doi.org/10.3390/proteomes7010002>
- Renwick, A. G. (1986). The metabolism of intense sweeteners. *Xenobiotica*, 16(10), 1057–1071. <https://doi.org/10.3109/00498258609038983>
- Richards, A. L., Eckhardt, M., & Krogan, N. J. (2021). Mass spectrometry-based protein–protein interaction networks for the study of human diseases [Publisher: Springer Science and Business Media LLC]. *Molecular Systems Biology*, 17(1). <https://doi.org/10.15252/msb.20188792>
- Riffle, M., May, D. H., Timmins-Schiffman, E., Mikan, M. P., Jaschob, D., Noble, W. S., & Nunn, B. L. (2017). MetaGOmics: A web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. *Proteomes*, 6(1), 2. <http://www.ncbi.nlm.nih.gov/pubmed/29280960>
- Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P., Sahin, I., Chandra, V., San Lucas, A., Scheet, P., Xu, H., Hanash, S. M., Feng, L., Burks, J. K., Do, K.-A., Peterson, C. B., Nejman, D., Tzeng, C.-W. D., ... McAllister, F. (2019). Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell*, 178(4), 795–806.e12. <https://doi.org/10.1016/j.cell.2019.07.008>
- Roberts, A., Renwick, A., Sims, J., & Snodin, D. (2000). Sucralose metabolism and pharmacokinetics in man. *Food and Chemical Toxicology*, 38, 31–41. [https://doi.org/10.1016/S0278-6915\(00\)00026-0](https://doi.org/10.1016/S0278-6915(00)00026-0)
- Robin, T., Bairoch, A., Mueller, M., Lisacek, F., & Lane, L. (2018). Large-scale reanalysis of publicly available HeLa cell proteomics data in the context of the human proteome project. *Journal of Proteome Research*, 17(12), 4160–4170. <https://doi.org/10.1021/acs.jproteome.8b00392>
- Rode, J., Brengešjö Johnson, L., König, J., Rangel, I., Engstrand, L., Repsilber, D., & Brummer, R. J. (2024). Fecal samples and rectal swabs adequately reflect the human colonic luminal microbiota. *Gut Microbes*, 16(1), 2416912. <https://doi.org/10.1080/19490976.2024.2416912>
- Salvato, F., Hettich, R. L., & Kleiner, M. (2021). Five key aspects of metaproteomics as a tool to understand functional interactions in host-associated microbiomes. *PLOS Pathogens*, 17(2), e1009245. <https://doi.org/10.1371/journal.ppat.1009245>
- Sanna, S., Van Zuydam, N. R., Mahajan, A., Kurilshikov, A., Vich Vila, A., Vösa, U., Mujagic, Z., Masclee, A. A. M., Jonkers, D. M. A. E., Oosting, M., Joosten, L. A. B., Netea, M. G., Franke, L., Zhernakova, A., Fu, J., Wijmenga, C., & McCarthy, M. I. (2019). Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics*, 51(4), 600–605. <https://doi.org/10.1038/s41588-019-0350-x>
- Sato, T., Kusuhabara, S., Yokoi, W., Ito, M., & Miyazaki, K. (2017). Prebiotic potential of l-sorbose and xylitol in promoting the growth and metabolic activity of specific butyrate-producing bacteria in human fecal culture. *FEMS Microbiology Ecology*, 93(1). <https://doi.org/10.1093/femsec/fiw227>
- Saulnier, D. M. A., Gibson, G. R., & Kolida, S. (2008). In vitro effects of selected synbiotics on the human faecal microbiota composition. *FEMS Microbiology Ecology*, 66(3), 516–527. <https://doi.org/10.1111/j.1574-6941.2008.00561.x>
- Schiebenhoefer, H., Schallert, K., Renard, B. Y., Trappe, K., Schmid, E., Benndorf, D., Riedel, K., Muth, T., & Fuchs, S. (2020). A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and prophane. *Nature Protocols*, 15(10), 3212–3239. <https://doi.org/10.1038/s41596-020-0368-7>

REFERENCES

- Schiebenhoefer, H., Van Den Bossche, T., Fuchs, S., Renard, B. Y., Muth, T., & Martens, L. (2019). Challenges and promise at the interface of metaproteomics and genomics: An overview of recent progress in metaproteogenomic data analysis. *Expert Review of Proteomics*, *16*(5), 375–390. <https://doi.org/10.1080/14789450.2019.1609944>
- Schiffman, S. S., & Nagle, H. T. (2019). Revisited: Assessing the in vivo data on low/no-calorie sweeteners and the gut microbiota. *Food and Chemical Toxicology*, *132*. <https://doi.org/10.1016/j.fct.2019.110692>
- Sender, R., Fuchs, S., & Milo, R. (2016a). Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans. *Cell*, *164*(3), 337–340. <https://doi.org/10.1016/j.cell.2016.01.013>
- Sender, R., Fuchs, S., & Milo, R. (2016b). Revised estimates for the number of human and bacteria cells in the body. *PLOS Biology*, *14*(8), e1002533. <https://doi.org/10.1371/journal.pbio.1002533>
- Serrano, J., Smith, K. R., Crouch, A. L., Sharma, V., Yi, F., Vargova, V., LaMoia, T. E., Dupont, L. M., Serna, V., Tang, F., Gomes-Dias, L., Blakeslee, J. J., Hatzakis, E., Peterson, S. N., Anderson, M., Pratley, R. E., & Kyriazis, G. A. (2021). High-dose saccharin supplementation does not induce gut microbiota changes or glucose intolerance in healthy humans and mice. *Microbiome*, *9*(1). <https://doi.org/10.1186/s40168-020-00976-w>
- Shi, M., Evans, C. A., McQuillan, J. L., Noirel, J., & Pandhal, J. (2024). LFQRatio: A normalization method to decipher quantitative proteome changes in microbial coculture systems. *Journal of Proteome Research*. <https://doi.org/10.1021/acs.jproteome.3c00714>
- Shiferaw, G. A., Vandermarliere, E., Hulstaert, N., Gabriels, R., Martens, L., & Volders, P.-J. (2020). COSS: A fast and user-friendly tool for spectral library searching. *Journal of Proteome Research*, *19*(7), 2786–2793. <https://doi.org/10.1021/acs.jproteome.9b00743>
- Simopoulos, C. M. A., Ning, Z., Zhang, X., Li, L., Walker, K., Lavallée-Adam, M., & Figeys, D. (2020). pep-Funk: A tool for peptide-centric functional analysis of metaproteomic human gut microbiome studies (J. Wren, Ed.). *Bioinformatics*, *36*(14), 4171–4179. <https://doi.org/10.1093/bioinformatics/btaa289>
- Suez, J., Korem, T., Zeevi, D., Zilberman-Schapira, G., Thaiss, C. A., Maza, O., Israeli, D., Zmora, N., Gilad, S., Weinberger, A., Kuperman, Y., Harmelin, A., Kolodkin-Gal, I., Shapiro, H., Halpern, Z., Segal, E., & Elinav, E. (2014). Artificial sweeteners induce glucose intolerance by altering the gut microbiota. *Nature*, *514*(7521), 181+. <https://doi.org/10.1038/nature13793>
- Sun, Z., Ning, Z., Cheng, K., Duan, H., Wu, Q., Mayne, J., & Figeys, D. (2023). MetaPep: A core peptide database for faster human gut metaproteomics database searches. *Computational and Structural Biotechnology Journal*, *21*, 4228–4237. <https://doi.org/10.1016/j.csbj.2023.08.025>
- Sun, Z., Ning, Z., & Figeys, D. (2024). The landscape and perspectives of the human gut metaproteomics. *Molecular & Cellular Proteomics*, *23*(5), 100763. <https://doi.org/10.1016/j.mcpro.2024.100763>
- Sundh, U. B., Beltoft, V., Bursch, W., Carere, A., & Engel, K.-H. (2010). SCIENTIFIC OPINION flavouring group evaluation 32 (FGE.32): Flavonoids (flavanones and dihydrochalcones) from chemical groups 25 and 30. *EFSA Journal*, *(8)*. <https://doi.org/10.2903/j.efsa.2010.1065>
- Sylvetsky, A. C., Jin, Y., Clark, E. J., Welsh, J. A., Rother, K. I., & Talegawkar, S. A. (2017). Consumption of low-calorie sweeteners among children and adults in the united states. *Journal of the Academy of Nutrition and Dietetics*, *117*(3), 441–448. <https://doi.org/10.1016/j.jand.2016.11.004>
- Sylvetsky, A. C., & Rother, K. I. (2016). Trends in the consumption of low-calorie sweeteners. *Physiology & Behavior*, *164*, 446–450. <https://doi.org/10.1016/j.physbeh.2016.03.030>
- Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., Jensen, L. J., & von Mering, C. (2017). The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible [Publisher: Oxford University Press (OUP)]. *Nucleic Acids Research*, *45*, D362–D368. <https://doi.org/10.1093/nar/gkw937>
- Tanca, A., Abbondio, M., Fiorito, G., Pira, G., Sau, R., Manca, A., Muroni, M. R., Porcu, A., Scanu, A. M., Cossu-Rocca, P., De Miglio, M. R., & Uzzau, S. (2022). Metaproteomic profile of the colonic luminal microbiota from patients with colon cancer. *Frontiers in Microbiology*, *13*, 869523. <https://doi.org/10.3389/fmicb.2022.869523>

REFERENCES

- Tennant, D. R. (2014). Potential intakes of total polyols based on UK usage survey data. *Food Additives & Contaminants Part A*, *31*(4), 574–586. <https://doi.org/10.1080/19440049.2014.886132>
- The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. <https://doi.org/10.1038/nature11234>
- Tilg, H., Adolph, T. E., Gerner, R. R., & Moschen, A. R. (2018). The intestinal microbiota in colorectal cancer. *Cancer Cell*, *33*(6), 954–964. <https://doi.org/10.1016/j.ccell.2018.03.004>
- Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, *11*(12), 2301–2319. <https://doi.org/10.1038/nprot.2016.136>
- US FDA, U. F. a. D. A. (2018). Additional information about high-intensity sweeteners permitted for use in food in the united states. <https://www.fda.gov/food/food-additives-petitions/additional-information-about-high-intensity-sweeteners-permitted-use-food-united-states>
- Vakil, N. (2018). Dietary fermentable oligosaccharides, disaccharides, monosaccharides, and polyols (FODMAPs) and gastrointestinal disease. *Nutrition in Clinical Practice*, *33*(4), 468–475. <https://doi.org/10.1002/ncp.10108>
- Van den Bossche, T., Kunath, B. J., Schallert, K., Schaepe, S. S., Abraham, P. E., Armengaud, J., Arntzen, M. O., Bassignani, A., Benndorf, D., Fuchs, S., Giannone, R. J., Griffin, T. J., Hagen, L. H., Halder, R., Henry, C., Hettich, R. L., Heyer, R., Jagtap, P., Jehmlich, N., ... Muth, T. (2021). Critical assessment of MetaProteome investigation (CAMPI): A multi-laboratory comparison of established workflows. *Nature Communications*, *12*(1). <https://doi.org/10.1038/s41467-021-27542-8>
- Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Židek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., ... Velankar, S. (2024). AlphaFold protein structure database in 2024: Providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, *52*, D368–D375. <https://doi.org/10.1093/nar/gkad1011>
- Vaudel, M., Burkhardt, J. M., Breiter, D., Zahedi, R. P., Sickmann, A., & Martens, L. (2012). A complex standard for protein identification, designed by evolution. *Journal of Proteome Research*, *11*(10), 5065–5071. <https://doi.org/10.1021/pr300055q>
- Verschaffelt, P., Tanca, A., Abbondio, M., Van Den Bossche, T., Moortele, T. V., Dawyndt, P., Martens, L., & Mesuere, B. (2023). Unipept desktop 2.0: Construction of targeted reference protein databases for metaproteogenomics analyses. *Journal of Proteome Research*, *22*(8), 2620–2628. <https://doi.org/10.1021/acs.jproteome.3c00091>
- Vicedomini, R., Bouly, J., Laine, E., Falciatore, A., & Carbone, A. (2022). Multiple profile models extract features from protein sequence data and resolve functional diversity of very different protein families (A. Ouangraoua, Ed.). *Molecular Biology and Evolution*, *39*(4), msac070. <https://doi.org/10.1093/molbev/msac070>
- Vogt, N. M., Kerby, R. L., Dill-McFarland, K. A., Harding, S. J., Merluzzi, A. P., Johnson, S. C., Carlsson, C. M., Asthana, S., Zetterberg, H., Blennow, K., Bendlin, B. B., & Rey, F. E. (2017). Gut microbiome alterations in alzheimer’s disease. *Scientific Reports*, *7*(1), 13537. <https://doi.org/10.1038/s41598-017-13601-y>
- Wang, H., Wan, L., Shi, J., Zhang, T., Zhu, H., Jiang, S., Meng, S., Wu, S., Sun, J., Chang, L., Zhang, L., Wan, K., Yang, J., Zhao, X., Liu, H., Zhang, Y., Dai, E., & Xu, P. (2021). Quantitative proteomics reveals that dormancy-related proteins mediate the attenuation in mycobacterium strains. *Virulence*, *12*(1), 2228–2246. <https://doi.org/10.1080/21505594.2021.1965703>
- Wang, J., Ma, Z., Carr, S. A., Mertins, P., Zhang, H., Zhang, Z., Chan, D. W., Ellis, M. J., Townsend, R. R., Smith, R. D., McDermott, J. E., Chen, X., Paulovich, A. G., Boja, E. S., Mesri, M., Kinsinger, C. R., Rodriguez, H., Rodland, K. D., Liebler, D. C., & Zhang, B. (2017). Proteome profiling outperforms transcriptome profiling for coexpression based gene function prediction. *Molecular & Cellular Proteomics*, *16*(1), 121–134. <https://doi.org/10.1074/mcp.M116.060301>
- Whipps, J., Lewis, K., & Cooke, R. (1988). *Fungi in biological control systems*. Manchester University Press Manchester, UK.

REFERENCES

- Whiteside, S. A., Razvi, H., Dave, S., Reid, G., & Burton, J. P. (2015). The microbiome of the urinary tract—a role beyond infection. *Nature Reviews Urology*, *12*(2), 81–90. <https://doi.org/10.1038/nrur.2014.361>
- Wiles, T. J., Schlomann, B. H., Wall, E. S., Betancourt, R., Parthasarathy, R., & Guillemin, K. (2020). Swimming motility of a gut bacterial symbiont promotes resistance to intestinal expulsion and enhances inflammation. *PLOS Biology*, *18*(3). <https://doi.org/10.1371/journal.pbio.3000661>
- Wilmes, P., & Bond, P. L. (2006). Metaproteomics: Studying functional gene expression in microbial ecosystems. *Trends in Microbiology*, *14*(2), 92–97. <https://doi.org/10.1016/j.tim.2005.12.006>
- Wilmes, P., & Bond, P. L. (2004). The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology*, *6*(9), 911–920. <https://doi.org/10.1111/j.1462-2920.2004.00687.x>
- Wolf, M., Schallert, K., Knipper, L., Sickmann, A., Sczyrba, A., Benndorf, D., & Heyer, R. (2023). Advances in the clinical use of metaproteomics. *Expert Review of Proteomics*, *20*(4), 71–86. <https://doi.org/10.1080/14789450.2023.2215440>
- Wong, J. M. W., & Jenkins, D. J. A. (2007). Carbohydrate digestibility and metabolic effects. *Journal of Nutrition*, *137*(11), 2539S–2546S. <https://doi.org/10.1093/jn/137.11.2539S>
- Wu, G., Xu, T., Zhao, N., Lam, Y. Y., Ding, X., Wei, D., Fan, J., Shi, Y., Li, X., Li, M., Ji, S., Wang, X., Fu, H., Zhang, F., Shi, Y., Zhang, C., Peng, Y., & Zhao, L. (2024). A core microbiome signature as an indicator of health. *Cell*, *187*(23), 6550–6565.e11. <https://doi.org/10.1016/j.cell.2024.09.019>
- Wu, G., Zhao, N., Zhang, C., Lam, Y. Y., & Zhao, L. (2021). Guild-based analysis for understanding gut microbiome in human health and diseases. *Genome Medicine*, *13*(1), 22. <https://doi.org/10.1186/s13073-021-00840-y>
- Xu, F., Li, D.-P., Huang, Z.-C., Lu, F.-L., Wang, L., Huang, Y.-L., Wang, R.-F., Liu, G.-X., Shang, M.-Y., & Cai, S.-Q. (2015). Exploring in vitro, in vivo metabolism of mogroside v and distribution of its metabolites in rats by HPLC-ESI-IT-TOF-MSn. *Journal of Pharmaceutical and Biomedical Analysis*, *115*, 418–430. <https://doi.org/10.1016/j.jpba.2015.07.024>
- Xu, J., Zhang, H., Zheng, J., Dovoedo, P., & Yin, Y. (2020). eCAMI: Simultaneous classification and motif identification for enzyme annotation (J. Xu, Ed.). *Bioinformatics*, *36*(7), 2068–2075. <https://doi.org/10.1093/bioinformatics/btz908>
- Yamamura, R., Inoue, K. Y., Nishino, K., & Yamasaki, S. (2023). Intestinal and fecal pH in human health. *Frontiers in Microbiomes*, *2*, 1192316. <https://doi.org/10.3389/frmbi.2023.1192316>
- Yang, J., Wei, H., Zhou, Y., Szeto, C.-H., Li, C., Lin, Y., Coker, O. O., Lau, H. C. H., Chan, A. W., Sung, J. J., & Yu, J. (2022). High-fat diet promotes colorectal tumorigenesis through modulating gut microbiota and metabolites. *Gastroenterology*, *162*(1), 135–149.e2. <https://doi.org/10.1053/j.gastro.2021.08.041>
- Yang, J., Gao, Z., Ren, X., Sheng, J., Xu, P., Chang, C., & Fu, Y. (2021a). DeepDigest: Prediction of protein proteolytic digestion with deep learning. *Analytical Chemistry*, *93*(15), 6094–6103. <https://doi.org/10.1021/acs.analchem.0c04704>
- Yang, Y., Lin, L., & Qiao, L. (2021b). Deep learning approaches for data-independent acquisition proteomics. *Expert Review of Proteomics*, *18*(12), 1031–1043. <https://doi.org/10.1080/14789450.2021.2020654>
- Yang, Y., Liu, X., Shen, C., Lin, Y., Yang, P., & Qiao, L. (2020). In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications*, *11*(1). <https://doi.org/10.1038/s41467-019-13866-z>
- Yatsunenkov, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., ... Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222–227. <https://doi.org/10.1038/nature11053>
- Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y., & Yin, Y. (2018a). dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, *46*, W95–W101. <https://doi.org/10.1093/nar/gky418>

REFERENCES

- Zhang, X., Li, Y., Shao, W., & Lam, H. (2011). Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics*, *11*(6), 1075–1085. <https://doi.org/10.1002/pmic.201000492>
- Zhang, X., Deeke, S. A., Ning, Z., Starr, A. E., Butcher, J., Li, J., Mayne, J., Cheng, K., Liao, B., Li, L., Singleton, R., Mack, D., Stintzi, A., & Figeys, D. (2018b). Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nature Communications*, *9*(1), 2873. <https://doi.org/10.1038/s41467-018-05357-4>
- Zhang, X., & Figeys, D. (2019). Perspective and guidelines for metaproteomics in microbiome studies. *Journal of Proteome Research*, *18*(6), 2370–2380. <https://doi.org/10.1021/acs.jproteome.9b00054>
- Zhang, X., Ning, Z., Mayne, J., Moore, J. I., Li, J., Butcher, J., Deeke, S. A., Chen, R., Chiang, C.-K., Wen, M., Mack, D., Stintzi, A., & Figeys, D. (2016). MetaPro-IQ: A universal metaproteomic approach to studying human and mouse gut microbiota. *Microbiome*, *4*(1), 31. <https://doi.org/10.1186/s40168-016-0176-z>
- Zhang, Z., Geng, J., Tang, X., Fan, H., Xu, J., Wen, X., Ma, Z. S., & Shi, P. (2014). Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *The ISME Journal*, *8*(4), 881–893. <https://doi.org/10.1038/ismej.2013.185>
- Zhao, J., Yang, Y., Xu, H., Zheng, J., Shen, C., Chen, T., Wang, T., Wang, B., Yi, J., Zhao, D., Wu, E., Qin, Q., Xia, L., & Qiao, L. (2023). Data-independent acquisition boosts quantitative metaproteomics for deep characterization of gut microbiota. *npj Biofilms and Microbiomes*, *9*(1), 4. <https://doi.org/10.1038/s41522-023-00373-9>
- Zheng, D., Liwinski, T., & Elinav, E. (2020). Interaction between microbiota and immunity in health and disease. *Cell Research*, *30*(6), 492–506. <https://doi.org/10.1038/s41422-020-0332-7>
- Zmora, N., Suez, J., & Elinav, E. (2019). You are what you eat: Diet, health and the gut microbiota. *Nature Reviews Gastroenterology & Hepatology*, *16*(1), 35–56. <https://doi.org/10.1038/s41575-018-0061-2>

CURRICULUM VITAE

PERSONAL INFORMATION

- **Name:** Zhongzhi Sun
- **Languages:** English (fluent), Chinese (native)

EDUCATION

- **Ph.D. in Biochemistry with Specialization in Bioinformatics**

Faculty of Medicine, University of Ottawa, Canada (2021–2025)

Research focus: Peptide-centric human gut metaproteomics analysis.

Supervisor: Dr. Daniel Figeys

- **M.Sc. in Microbiology**

State Key Laboratory of Microbial Technology, Shandong University, China (2018–2021)

Research focus: Phylogenetic distribution of polysaccharide-degrading enzymes in marine bacteria.

Supervisor: Dr. Binbin Xie

- **B.Sc. in Biotechnology**

School of Life Science, Shandong University, China (2014–2018)

SELECTED PUBLICATIONS

1. **MetaPep: A core peptide database for faster human gut metaproteomics database searches.** *Computational and Structural Biotechnology Journal*. 2023. **Zhongzhi Sun**, Zhibin Ning, Kai Cheng, Haonan Duan, Qing Wu, Janice Mayne, and Daniel Figeys. doi: <https://doi.org/10.1016/j.csbj.2023.08.025>
2. **Comprehensive assessment of functional effects of commonly used sweeteners on *ex vivo* human gut microbiomes.** *Microbiology Spectrum*. 2022. **Zhongzhi Sun**, Wenju Wang, Leyuan Li, Xu Zhang, Zhibin Ning, Janice Mayne, Krystal Walker, Alain Stintzi, and Daniel Figeys. doi: <https://doi.org/10.1128/spectrum.00412-22>
3. **Peptide abundance correlations in metaproteomics enhance taxonomic and functional analysis of the human gut microbiome.** *bioRxiv*. 2025. **Zhongzhi Sun**, Zhibin Ning, Qing Wu, Leyuan Li, Andrew C. Doxey, and Daniel Figeys. doi: <https://doi.org/10.1101/2025.03.13.642886>
4. **The landscape and perspectives of the human gut metaproteomics.** *Molecular & Cellular Proteomics*. 2024. **Zhongzhi Sun**, Zhibin Ning, and Daniel Figeys. doi: <https://doi.org/10.1016/j.mcpro.2024.100763>
5. **Assessing the dark field of metaproteome.** *Analytical Chemistry*. 2022. Haonan Duan, Kai Cheng, Zhibin Ning, Leyuan Li, Janice Mayne, **Zhongzhi Sun**, and Daniel Figeys. doi: <https://doi.org/10.1021/acs.analchem.2c02452>
6. **Phylogenetic distribution of polysaccharide-degrading enzymes in marine bacteria.** *Frontiers in Microbiology*. 2021. **Zhongzhi Sun**, Bowen Ji, Ning Zheng, Meng Wang, Ye Cao, Lu Wan, Yisong Li, Jincheng Rong, Hailun He, Xiulan Chen, Yuzhong Zhang and Binbin Xie. doi: <https://doi.org/10.3389/fmicb.2021.658620>

7. **Comparative genomics reveals broad genetic diversity, extensive recombination and nascent ecological adaptation in *Micrococcus luteus*.** *BMC Genomics*. 2021. Yisong Li, **Zhongzhi Sun**, Jincheng Rong, and Binbin Xie. doi: <https://doi.org/10.1186/s12864-021-07432-5>

CONFERENCES AND PRESENTATIONS

- Oral presentation, Advancing Human Gut Metaproteomics: From Protein-Level Profiles to Peptide-Centric Insights, 74th Annual Conference of the Canadian Society of Microbiologists, Montreal, Canada, 2025.
- Oral presentation, Constructing a Peptide Library for Faster Human Gut Metaproteomics Research, 5th International Metaproteomics Symposium, Avignon, France, 2023.
- Oral presentation, Comprehensive Assessment of Functional Effects of Commonly Used Sugar Substitute Sweeteners on *Ex Vivo* Human Gut Microbiome, Ottawa Institute of Systems Biology retreat, Cornwall, Canada, 2022.
- Poster presentation, Constructing a Core Peptide Database for Faster Human Gut Metaproteomics Research, 65th Annual Canadian Society for Molecular Biosciences, Ottawa, Canada, 2023.

HONOURS AND AWARDS

- International Doctoral Scholarship, University of Ottawa, 2021—2025
- Admission Scholarship, University of Ottawa, 2021—2025

PERMISSIONS AND COPYRIGHTS

Parts of this thesis include material that has been published or submitted for publication. Permission has been obtained from the respective publishers to reuse these materials where required.

Chapter 1: Zhongzhi Sun, Zhibin Ning, and Daniel Figeys. The landscape and perspectives of the human gut metaproteomics. 2024. *Molecular & Cellular Proteomics*. doi: <https://doi.org/10.1016/j.mcpro.2024.100763>.

This publication is licensed under: CC-BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

Chapter 2: Zhongzhi Sun, Wenju Wang, Leyuan Li, Xu Zhang, Zhibin Ning, Janice Mayne, Krystal Walker, Alain Stintzi, and Daniel Figeys. Comprehensive assessment of functional effects of commonly used sweeteners on *ex vivo* human gut microbiomes. 2022. *Microbiology Spectrum*. doi: <https://doi.org/10.1128/spectrum.00412-22>.

This publication is licensed under: CC-BY 4.0.

Chapter 3: Zhongzhi Sun, Zhibin Ning, Kai Cheng, Haonan Duan, Qing Wu, Janice Mayne, and Daniel Figeys. MetaPep: A core peptide database for faster human gut metaproteomics database searches. 2023. *Computational and Structural Biotechnology Journal*. doi: <https://doi.org/10.1016/j.csbj.2023.08.025>.

This publication is licensed under: CC BY-NC-ND 4.0 (<https://creativecommons.org/licenses/by-nc-nd/4.0/>).

Chapter 4: Zhongzhi Sun, Zhibin Ning, Qing Wu, Leyuan Li, Andrew C. Doxey, and Daniel Figeys. Peptide abundance correlations in metaproteomics enhance taxonomic and functional analysis of the human gut microbiome. 2025. *bioRxiv*. doi: <https://doi.org/10.1101/2025.03.13.642886>.

This publication is licensed under: CC BY-NC 4.0 (<https://creativecommons.org/licenses/by-nc/4.0/>).

All of the above publications are covered under licenses that permit the authors to:

- **Share** — copy and redistribute the material in any medium or format.
- **Adapt** — remix, transform, and build upon the material.

As a non-native English speaker, I sought to enhance the clarity and coherence of my writing with the assistance of a large language model. Throughout the process of composing this thesis, I used ChatGPT (developed by OpenAI) to obtain suggestions and support in refining the language. ChatGPT was employed solely to assist with language polishing and not for generating original content. I have carefully reviewed all wording and expressions in this thesis to ensure they accurately convey my intended meaning.