

**Systematic Evaluation of Database Builds in Metaproteomics
for the Advancement of Microbiome Research**

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Katherine Elizabeth Ostrouchov
December 2022

Copyright © 2022 by Katherine Elizabeth Ostrouchov
All rights reserved.

ACKNOWLEDGEMENTS

I would like to begin by expressing my sincere gratitude for my advisor, Dr. Robert L. Hettich. If it was not for him, I would not have been able to pursue research in medicine and land the job of my dreams. He took me in as an intern upon completion of my Bachelor's degree and taught me core principles in mass spectrometry and introduced me to microbiome research. His mentorship set me up for success as a graduate student, and none of this would have been possible without his help.

I would like to give a special thanks to Dr. Alfredo Blakeley-Ruiz and the soon-to-be Dr. Him Shrestha for their constant guidance and contributions made throughout the entirety of my graduate research in this lab. They helped me navigate many of the road bumps I encountered along the way when I needed it most and helped me be successful.

I would like to thank the many past and present members of the “Bob Mob” at the Oak Ridge National Lab for their constant support and mentorship over the past three years. Their friendship helped me keep a positive outlook in graduate school, and their advice helped me to become a better scientist.

I would like to personally thank Dr. Zachary Beamer. He taught me every wet-lab technique known to the world of biochemistry and introduced me to Dr. Hettich and the Genome Science and Technology program. Without his guidance and mentorship throughout my undergraduate and graduate coursework, I would not be the scientist I am today.

I would like to thank my committee members—Dr. Shawn Campagna and Dr. Jennifer Morrell-Falvey. Dr. Campagna, your organic chemistry class inspired me to pursue a Bachelor's in biochemistry and to enter the world of analytical chemistry. Dr. Morrell-Falvey, since I first met you at the Oak Ridge National Lab, your enthusiasm for research and contagious personality is something that I have always aspired to be as a scientist.

I would like to acknowledge the Genome Science and Technology graduate program offered by the University of Tennessee and the Oak Ridge National Lab for providing me with the opportunity to pursue my Master's degree.

I would also like to thank CADES at the Oak Ridge National Lab for providing me with the cloud resources to download and cluster thousands of proteomes vital to my thesis research.

Lastly, I would like to thank all my family and friends for their constant support throughout my life, and for pushing me to be all that I can be. Thank you to George Mathews V Powell for supporting me through all the ups and downs of graduate school and for pushing me to pursue my dreams above all else. Thank you to my mother, Susan Blackford, for helping me become the strong woman that I am today as a veteran and scientist. Thank you to my father, Dr. George Ostrouchov, for your unwavering support and mentorship throughout my life. I sincerely dedicate this thesis to you. Your tutoring in computational statistics molded me into the great computational biologist I am today. I follow in the footsteps of not only a brilliant scientist but one incredible father.

ABSTRACT

The microscopic life that inhabits a human shares a unique bond with its host. Microbes perform many functions that are vital to the survival of the human species and have long been shown to regulate the absorption of nutrients and to promote immune function. A lack of exposure to certain microbes early in life, excessive antibiotic usage, and improper diet can perturb human microbiomes and lead to disease. Since the emergence of omic sequencing technologies, it has now become possible to measure and monitor the genes and proteins made by these microorganisms to better understand how they contribute to host health or drive potential disease conditions. Early attempts at studying the genomes and proteomes of these environments have revealed that each person may house a unique community of microbial species, each of which contributes distinct and unique functions that separate healthy vs. disbiotic human gut microbiomes. A subfield of the omics known as metaproteomics is now being used to help characterize all proteins found in an environment, and it has been found to show tremendous potential for describing these microbial systems. However, a great deal remains to be done in this field to improve the accuracy and depth of information gained from these investigations. There is currently no accepted standard for determining which genomic databases are best suitable for searching protein sequencing data and it is difficult to know how/ if false positives are being incorporated into the results. The research for this thesis investigated how database curation affects the number of identified proteins and peptides identified from mice gut metaproteome spectra and explored how functional and taxonomic annotation varied according to the database used. The results indicated that translating a high-depth sequenced and assembled metagenome yielded the highest number of identifications while maintaining a low false discovery rate for mice fecal samples and that each database build identified unique and distinct functional and taxonomic information. The goal of this thesis is to better inform the field of metaproteomics and hopefully guide researchers towards a standard practice of using deep metagenome sequencing for database curation, resulting in more thorough coverage of microbiomes with greater confidence.

TABLE OF CONTENTS

Chapter 1 Introduction to the Microbiome and Metaproteomics.....	1
1.1 Defining the Human Microbiome.....	1
1.2 Composition of the Human Gut Microbiome	2
1.3 Techniques to Study Microbiomes	4
1.4 Challenges That Remain to Be Addressed in Metaproteomics	7
1.5 Scope of Thesis	11
Chapter 2 A Description of the Bioinformatics Toolbox for Metaproteome Research.....	13
2.1 Introduction to Metaproteomics Data Analysis	13
2.2 Quality Checking of Spectra.....	15
2.3 Database Construction	15
2.3.1 Matched-Metagenomes and Catalogs	16
2.3.2 Pseudo-Metagenomes	16
2.3.3 Clustering of Sequences.....	17
2.4 Database Searching	18
2.5 Data Processing.....	20
2.5.1 Filtering.....	20
2.5.2 Normalization	24
2.5.3 Imputation	24
2.5.4 Statistical Testing.....	27
2.6 Data Display.....	29
2.7 Conclusion	31
Chapter 3 Evaluating Measurement Depth, Accuracy, and Functional Assignment in Metaproteomics Using Matched-Metagenomes, Catalogs, and Pseudo-Metagenomes in Mice Gut Model	34
3.1 Introduction.....	35
3.2 Materials and Methods.....	38
3.2.1 Study Design.....	38
3.2.2 Database Construction	38

3.2.3 Search Parameters	41
3.2.4 Counting Total Inferred Proteins and Peptides	43
3.2.5 Assessing the Sequence Coverage	43
3.2.6 Checking Sequence Similarity with BLASTP	43
3.2.7 Assigning Function and Taxonomy to Identified Proteins with GhostKOALA.....	45
3.3 Results and Discussion	45
3.3.1 Investigating the Quality of Identifications Reveals the Matched-Metagenomes Significantly Outperformed All Database Builds	46
3.3.2 Inferring Function and Taxonomy Across Databases Reveals Significant Losses in Annotation and Drastic Changes in Abundance	57
3.4 Conclusions from Study and Closing Remarks	64
List of References	67
Appendix.....	80
Vita.....	88

LIST OF TABLES

Table 1 Total entrapment protein identifications across databases at FDR of 0.01 broken down by organism.....	55
Supplementary Table 1 Total entrapment protein identifications obtained for multi-step-search versus the matched-metagenomes.....	87

LIST OF FIGURES

Figure 1 General metaproteomics workflow	14
Figure 2 General database searching workflow in Proteome Discoverer	21
Figure 3 Filtering of protein expression data	23
Figure 4 Normalization of protein expression data	25
Figure 5 Imputation of protein expression data	28
Figure 6 KEGG color mapping of protein expression data	32
Figure 7 Study design	39
Figure 8 Steps in database construction	42
Figure 9 (A) Sizes of databases post clustering with cd-hit at ninety-eight percent sequence similarity and (B) description of databases	44
Figure 10 Total proteins and peptides identified at FDR 0.01 across databases and their corresponding margin of error from Proteome Discoverer	49
Figure 11 Protein sequence coverage (y axis) and protein length (x axis) obtained across databases for all measured identifications	52
Figure 12 Summed GhostKOALA KO (A/B), genus (C/D), and phylum (E/F) protein abundance values graphed in UpSetR plot and Euclidean distance heatmaps for high-fiber/saline (LIT) condition	60
Supplementary Figure 1 Obtaining thousands of RefSeq proteomes using a list of taxonomy or common names in bash	80
Supplementary Figure 2 Obtaining thousands of UniProt proteomes using taxonomy or common name	81
Supplementary Figure 3 Entrapment proteins/peptides identified for fiber-rich/saline (orange) & high-fat/saline (green) mice at FDR 0.01	82
Supplementary Figure 4 BLASTP results comparing microbial and polar bear entrapment protein sequences to all databases	83
Supplementary Figure 5 Summed GhostKoala KO (A/B), genus (C/D), and phylum (E/F) protein abundance values graphed in UpSetR plot and Euclidean distance heatmaps for high-fat/saline (GHP) condition	84

Supplemental Figure 6 Total protein and peptide identifications obtained for the multi-step- search versus the matched-metagenomes	86
---	----

CHAPTER 1

INTRODUCTION TO THE MICROBIOME AND METAPROTEOMICS

1.1 Defining the Human Microbiome

Eukaryotes house a dynamic ecosystem of microscopic life that performs vital functions which sustain the health of the host (Sullivan, 2019). This microbiota, collectively known as the microbiome, can be found inhabiting diverse surfaces both internal and external to the host. Each microbial member carries out distinct and unique metabolic functions which affect the growth of healthy eukaryotic life (Sullivan, 2019). Since the structure of the genetic code was uncovered through x-ray diffraction imaging conducted by Rosalind Franklin in 1953, it was believed that DNA was the driving force that influenced the development of different traits and behaviors in all organisms (Stasiak, 2003). However, over the past twenty years since the human microbiome project was formed in 2007, it was discovered that the microscopic life inhabiting humans may be influencing not only overall host health but also behaviors more than originally thought (NIH HMP Working Group, 2009; Sullivan, 2019). Within the human body, microbial genes are believed to contribute an additional one hundred times the amount of genetic information (Mueller, 2015). One excellent example that illustrates this concept was explained by Sullivan in 2019 which is that identical twins may have the same DNA, and, yet display different physical traits and behaviors.

All these phenomena are explained through what is known as epigenetics, which is the study of all processes that affect the expression of genes (Weinhold, 2006). Tissues from the animal kingdom arise from embryonic cells which later differentiate into many types such as epithelial, muscle, and neural tissue in response to chemical signals expressed in the environment which synchronously bind to DNA and influence the expression, suppression, creation, or deletion of specific genes (Simón, 2013). One of the first events that occur post-conception in eukaryotic organisms is the transmission of microorganisms from parents to progeny (Mueller, 2015). Some of the most critical colonizers for mammals are established through the direct contact of microbes colonizing the genital tract during birth (Mueller, 2015). Cesarean section and pregnancy-related antibiotic usage have been associated with an increase in disease if these

microbes are not allowed to colonize in early life (Mueller, 2015). Similarly, when rats are treated heavily with antibiotics and kept in a sanitized environment, these rats begin to exhibit disease (Bibiloni, 2012).

The reasoning behind why eukaryotes require microbiota can be explained by the nature of how eukaryotes were proposed to have emerged a billion years ago (Jékely, 2007). Research into prehistoric life in algae revealed that prokaryotes developed a gene that led to what is referred to as phagotrophy, the consumption of other life forms (Burki, 2014). This comes in contrast to what is known as autotrophy, an organism's ability to sustain its own health without engulfing another organism (Burki, 2014). Examples of autotrophs are plants that produce all the twenty-one amino acids necessary to sustain life by undergoing photosynthesis (Trovato, 2021). One disadvantage to being autotrophic is that photosynthesis is energetically taxing (Trovato, 2021). Adopting phagotrophy allowed these early prokaryotic organisms to grow and reproduce faster by means of stealing the nutrients necessary for their survival from other organisms (Trovato, 2021). This, however, also came at a cost. Eukaryotes are indebted to this cycle of relying on other life to sustain themselves. For example, mammals lack the genes necessary to produce nine amino acids critical to life, which are referred to as essential amino acids (Church, 2020). An amino acid necessary for protein biosynthesis to occur in humans is methionine (Trovato, 2021). Unfortunately, humans do not possess the genes necessary to absorb and metabolize methionine sufficiently without the help of microbial enzymes (Zuo, 2019). All these discoveries over the past decade have shifted the focus of the scientific community away from solely the host into the unknown world of the microbiome to better understand and treat disease.

1.2 Composition of the Human Gut Microbiome

The bond shared between microbiota and humans has become increasingly more apparent over the years as sequencing technology and database repositories have continued to grow. Studies ranging from allergy to cancer and even autism have identified key microorganisms responsible for modulating immune function in a variety

of different locations across the human body (Gilbert, 2018; Sullivan, 2019). According to estimates, humans house up to a thousand or more different bacterial species, which equates to roughly 1.3 fold more than the amount of cells found in humans (Gilbert, 2018). Given that each bacterium contains on average two thousand genes, it was estimated that roughly an additional two million genes may be contributing to the functions exhibited by the human body, which agrees with the size of assembled metagenomes obtained from the Human Microbiome Project (Gilbert, 2018; iHMP Research Network Consortium, 2014). For comparison, the human genome has been estimated to contain roughly twenty thousand genes, which is one hundred times less than those of microbes (Gilbert, 2018; iHMP Research Network Consortium, 2014). Given that microbial cells slightly outnumber human cells, it is reasonable to wonder who might be influencing the behaviors humans exhibit in their daily lives? New sequencing technology has afforded scientists the opportunity to investigate the various microbes that affect human behavior and how they might be the unsung warriors fending off pathogens that are encountered on a daily basis. Live microorganisms formulated as probiotics are now being administered as treatment options to patients suffering from all types of health problems (Gilbert, 2018).

Some of the first microbes discovered to colonize the infant gut in high abundance were the *Lactobacillus* and *Bifidobacteria* genera, which are obtained via contact with the vaginal wall and during breastfeeding (Milani, 2017). Infection resistance, improved digestion, and support for healthy immune function in early life have been demonstrated to be directly linked with high levels of these two genera (Milani, 2017). Conversely, high numbers of *Clostridium* and *Bacteroides* have been discovered in the intestines of newborns who undergo C-sections, receive antibiotics, or are fed formula, and these findings have been linked to an early onset of autoimmune disease, asthma, eczema, and allergies (Milani, 2017). *In-vitro Bifidobacterium breve* , *Bifidobacterium animalis*, and *Bifidobacterium longum* have all been found to stimulate the production of dendritic cells and activate T cells (Ruiz, 2017). *Lactobacillus rhamnosus* and *Lactobacillus helveticus* have been shown *in-vitro* to increase the production of interleukin-10, an anti-inflammatory cytokine, during clinical trials (Ding, 2017). *Lactobacillus rhamnosus* and *Lactobacillus plantarum* were shown to reduce cholesterol levels and suppress the production of pro-inflammatory cytokines in immune cell models (Yoon, 2013). Researchers have discovered

that the bacteria *Bifidobacterium spp.* and *Lactobacillus spp.* produce GABA, a major neurotransmitter inhibitor that is normally present at lower levels in individuals with autism spectrum disorder and has been demonstrated to lessen symptoms of anxiety and depression (Alharthi, 2022). There is also now strong evidence that has linked imbalances in gut microbes to the development of obesity (Sullivan, 2019; Aoun, 2020).

The significance of these findings has led to the administration of probiotics to help reduce the likelihood of infants and adults developing diseases (Milani, 2017). Gut microbe transplants from healthy subjects to diseased patients have also performed well for the treatment of gastrointestinal, neurological, and cardiovascular disease, but the mechanism by which this treatment works for some and not for others is not yet understood (Biazzo, 2022). Furthermore, there are currently no probiotics that have been approved by the FDA or EFSA in the United States due to insufficient clinical evidence suggesting that a lot of work is still ahead to understand these complex host-microbe interactions (Stavropoulou, 2020). *Lactobacillus* and *Bifidobacterium* represent some of the only genera found in probiotics today, and their discovery occurred over forty years ago (Zommiti, 2020). More gut microbiome studies are needed to identify and functionally characterize the other key microbes involved in human health.

1.3 Techniques for Studying the Microbiome

Characterizing complex communities of microorganisms, many of which may be unculturable, requires careful consideration prior to conducting an experiment. There is not a “one-size-fits-all” approach to studying these systems since each community may range in complexity, and extraction of DNA, RNA, or protein may present challenges depending on the environment these microorganisms are found. The construction of an unbiased and reproducible microbiome study remains to be one of the largest challenges (Allaband, 2019). Microbes are constantly sensing their environment and adapting to satisfy their metabolic needs. Because microbial colonization is dynamic and unique to each host, the collection of a representative microbiome sample across similar environments can be difficult. It is critical to perform sampling within the same location

and at multiple time points to control for changes that occur in an environment that might bias the communities obtained (Tang, 2020). Additionally, to classify functional differences between microbiomes, conditions such as the health, diet, temperature, pH, and cohabitation of a host should remain controlled in respect to treatments that might be tested (Human Microbiome Project Consortium, 2012; Dill-McFarland, 2019). Predicting causative agents in a perturbed environment is not possible if more than one variable is manipulated at the same time, as it can incorporate bias into the experiment. Additionally, another reason why the environment can have such a large effect on the characterization of these communities is because the species richness of microorganisms has been found to be host-specific (Gilbert, 2018). Numerous researchers have shown that the composition of microbial species which colonize eukaryotes drastically varies across individuals within the same species and that a driving factor that has been found to separate healthy from diseased is the function that is expressed within the host (Human Microbiome Project Consortium, 2012). Since the classification of microbiome functions are largely dependent on the gene products found in the direct environment and not the genes themselves, this explains why bias in sample collection can easily become a confounding factor in a study.

Because single-organism studies fail to address the interaction of other microbes within a host, the field of systems biology was developed to study all microbial members and their functions occurring in a host so that community perturbations could easily be identified. In systems biology, many robust and sensitive sequencing instruments have been developed to detect and characterize the molecular processes that are occurring in an environment. The main pillars of the central dogma have been integrated into systems biology to create specific fields referred to as “omics” which study the DNAs (metagenomics), RNAs (metatranscriptomics), proteins (metaproteomics), and metabolites (metabolomics) found in microbiome samples (Franklin, 2011).

In 1977, the Woese lab discovered a gene that was conserved across many generations in the same species of bacteria which could be used to catalog the lineage of organisms (Fox). This gene was known as *16S*, and it led to the development of the first microbiome RNA profiling technology known as amplicon sequencing (Woo, 2000). It was later discovered that a great deal

of bias existed for amplifying and sequencing only this gene across microbes, and thus whole genome and transcriptome sequencing of communities began to gain traction as the cost of sequencing became more affordable and automated (Heather, 2016). Efforts in metagenomics and metatranscriptomics to classify microbiomes have resulted in the assembly of numerous novel genomes, many of which were unculturable (Starr, 2018). These efforts throughout the years have increased awareness of the microbiota and its functions in humans and shifted the focus of the scientific community away from single organisms into systems (Starr, 2018; Ezzamouri, 2021). Although metagenomics, metatranscriptomics, and amplicon sequencing are powerful tools for profiling the genomes of microorganisms in a community, they do not interrogate the final functional products (i.e. proteins and metabolites) that best define phenotype, thus making it difficult to completely characterize the activities of organisms in a community (Salvato, 2021; Starr, 2018).

In 2004, an effort to characterize function within a microbial community was made for the first time which combined liquid chromatography in-line with tandem mass spectrometry (LC-MS/MS) to identify proteins extracted from a 2D-PAGE gel (Wilmes & Bond, 2004). Although this work only identified three proteins (due to lack of detailed metagenome information), this led to the explosion of a new field known as metaproteomics. About a year after this discovery, the Banfield and Hettich labs translated a metagenomic database and used it with a metaproteome approach to identify several thousand proteins from a microbial sample (Ram, 2005). The field of metaproteomics began moving away from 2D-PAGE in 2008 when the first nano-flow liquid chromatography device combined with a mass spectrometer was used by the Banfield and Hettich labs to separate peptides from proteins extracted from wastewater allowing metaproteome sequencing to become an automated process. Nano-flow LC-MS/MS has been revolutionary to provide the life sciences community with tools capable of determining the potential host or microbial genes responsible for the phenotypes observed in the environment (Wilmes, 2015).

Early shotgun whole metagenome sequencing efforts initially led to poor quality genomic assemblies, but as sequencing technology has advanced over the years it has made possible the annotation of millions of high-quality genes which have allowed for the translation of these genes into predicted protein sequences using various bioinformatic pipelines (Eiju, 2020). The Kyoto Encyclopedia of Genes and Genomes (KEGG) is an online database repository that was created in 1995 originally to serve as an international site for the pooling of genomic sequences to produce interactive maps that model gene expression pathways (Kanehisa, 2016). Initially, the annotation of proteins was confined to complete genomes, but as sequencing technology allowed for deeper measurements, this restriction was removed (Kanehisa, 2016). Over the years it has become an invaluable tool in metaproteomics now capable of annotating proteins by their corresponding enzyme names and mapping all pathways associated with these enzymes (Kanehisa, 2017). Another widely used international database created for the annotation of gene function was Gene Ontology (GO) in 1998 (Eiju, 2020). This database has been critical in characterizing the activity, location of gene products within the cell, and cellular functions exhibited by cells with a certain gene (Eiju, 2020). These large databases together have afforded the metaproteomics community the ability to identify biochemical pathways that might be perturbed in the environment, and they have now become paramount in the discovery process of identifying disease conditions (Kanehisa, 2017; Gene Ontology Consortium, 2019).

1.4 Challenges That Remain to Be Addressed in Metaproteomics

Over the past twenty years, metaproteomics research has advanced significantly, yet there are still numerous challenges that must be overcome to address inaccuracies in measurement. Accuracy in measurement, defined as arriving at a correct value, can be defined in two ways. One is assessing the number of true positives to determine if all the correct answers were obtained. The second is assessing the number of false positives, the incorrect answers, obtained out of the total which is known as the false discovery rate (FDR). Several methods have been developed to correct these types of errors in metaproteomics experiments, but a great deal of work remains to address some unknowns in the field related to this accuracy.

To begin, proteins have distinct chemical properties, and as a result, there is not a single extraction technique or sequencing method that has been found to capture them all (Issa Isaac, 2019). Second, the intensities measured for highly abundant proteins can hide the signal of low abundant proteins and cause them to be missed in the spectra (Issa Isaac, 2019). As a result, many microbial proteins are lost. Efforts to improve this depth in coverage of low abundant proteins have involved several filtration techniques that reduce the number of highly abundant proteins typically belonging to the host (Wilmes, 2015). This host protein depletion technique was applied to a metaproteome study of Crohn's disease subjects, and it resulted in the same community composition as a previous metagenome analysis (Manichanh, 2005; Wilmes, 2015). Similarly, another beneficial technique to circumvent some of the measurement issues of proteins mentioned above involves using a protease to break down proteins into peptides which has allowed for much easier and more thorough measurement of the proteins.

Numerous challenges also exist in the realm of protein and peptide identification using metaproteome spectra obtained from LC-MS/MS sequencing. To identify proteins through LC-MS/MS, mass spectra consisting of accurate measurement of the intact peptide masses (MS1) along with an examination of their collision-induced fragmentation products (MS2) must be matched to theoretical spectra representing potential protein sequences. Determining the right protein sequence database to utilize for matching to an unknown sample can seem like a daunting task. However, many methods have been developed to help with this process. Thanks to metagenomics, amplicon sequencing, and large database repositories like KEGG, databases can now be constructed for the protein identification process to take place.

Although characterizing all microbial proteins in a sample is the aim of metaproteomics, searching numerous distinct databases runs the risk of obtaining high false discovery rates (FDRs) because no current method is available for validation (Starr, 2018). The reason for this high FDR is due to many organisms in the environment sharing some level of similarity in their amino acid sequences, which is known as redundancy. This redundancy heavily influences the proteins that are identified in a study

(Starr, 2018). For example, if a sample contained a protein sequence that was ninety percent similar to another protein from a different organism found in a database, the wrong species and wrong protein might be identified if the real protein of interest was left out of the database. This is an unfortunate scenario, but it represents a major challenge in the field that has yet to be corrected. Not only does sequence redundancy present a challenge in the field, but inadequate metagenome sequencing can lead to improper protein annotation (Starr, 2018). From a simulated 20-member bacterial community containing a range of up to 10^7 genomes per μL , Pereira-Marques and colleagues in 2019 found that sequencing performed at less than ten million base pairs can lead to improper gene annotation at as low a concentration of 10^4 copies of DNA per μL .

Due to these shortcomings in metaproteomics due to false positive discovery, it is critical to create databases that accurately depict microbial composition expected for unknown peptide mixtures to minimize the risk of obtaining false positives. Although sample mixtures may be unknown, certain methods can be applied to obtain databases that represent the best prediction of what proteins might be in a sample. They consist of four types, and each has varying levels of confidence for the identifications that are obtained.

The first choice that would be ideal for protein discovery would be to utilize what is called a matched-metagenome (Tanca, 2013; Rechenberger, 2019; Van Den Bossche, 2021). This database is generated by performing direct metagenome sequencing on the same samples that will be interrogated by metaproteome measurements. The depth of the assembled metagenome is largely dependent on the concentration of DNA (Pereira-Marques, 2019). This metagenome data can be assembled so that genes can be annotated and protein assignments can be made (Timmins-Schiffman, 2017). Although this method would appear to be the best suited since it represents exactly what was measured in the samples previously, it can take significant time and effort to assemble after sequencing. It should also be noted that poor annotation, translation, and assembly of those metagenome sequences can lead to erroneous gene calls which then propagate into the translated protein databases. Time constraints and difficulty in assembly have typically driven away many folks in the metaproteomics community, but it has yet to be determined whether this database outperforms all others for gut microbiota identification.

Since metagenome sequencing and annotation can be quite time consuming, many groups have teamed up to sequence hosts or environments and deposited this data in shared repositories that represent a specific environment (Lesker, 2020; Zhu, 2021). This type of database is called a catalog. It is constructed from exhaustive metagenome sequencing of a particular host or environment which is then translated into protein sequences for use as a search database. The benefit to utilizing this type of database is that it comes close to representing the microbes that have been found to associate with a particular species or environment, and it represents a very deep measurement which might exceed that of a matched-metagenome. A drawback to utilizing this method is that it does not represent metagenome evidence directly obtained from the sample, and that sequencing and annotation strategies might differ across the samples represented within the catalog. It is, however, easy to obtain, which makes this strategy quite appealing.

Unfortunately, not all microbiomes have a catalog that has been dedicated to its research. In these cases, a pseudo-metagenome may be utilized. This database is constructed by obtaining protein sequences from known microbial isolates that are available from online repositories using some level of filtering criteria. The criteria for constructing this database can vary. Some groups have utilized only microbes pertaining to certain hosts or environments. Others have used amplicon and metagenome information to select protein sequences from annotated species of interest. The benefit of utilizing this method is that it is easy to obtain quickly. One major drawback, however, is that reference databases may contain bias towards a certain community of organisms, namely human gut colonizers, and they will miss or incorrectly assign proteins and taxonomy to microbes outside this host environment (Nath, 2021). Another problem using online repositories is that many databases can have incorrect annotation due to poor metagenome sequencing (Starr, 2018).

A fourth database type involves utilizing one of the above databases constructed by removal of members from the original database based upon initial search identifications called a multi-step-search. There has been a great deal of controversy in the field regarding this database type. Although it has been found to identify more

proteins and peptides, this technique appears to inflate the false discovery rate (Vaudel, 2012).

Previous groups have found that combining pseudo-metagenomes and matched-metagenomes can increase the number of identifications (Xaio, 2018). However, the accuracy of that assignment has not been systematically studied. Furthermore, a high-depth assembled metagenome has not been compared to all these database types to address which method might be appropriate for mice microbiomes.

A standard was developed in proteomics for single organisms to calculate the FDR of identifying proteins involved utilizing the reversed amino acid sequences of a database to create what is known as a decoy database (Vaudel, 2012). This decoy strategy has been applied in metaproteomics during the identification process to calculate an FDR where the likelihood of matching to a reverse sequence is less than a target sequence (Vaudel, 2012). In principle, this decoy database has been shown to perform well in proteomics for calculating a FDR for single organisms, but it fails to account for the false positives that might be occurring due to errors in database construction in metaproteomics (Starr, 2018). In addition to the decoy database, several protein inference strategies have been applied to classify a protein group with the amount of unique peptides measured that belong to a single organism (Li, 2001). This technique helps to down-select peptides that may have obtained hits to multiple organisms or multiple proteins and increases the confidence in metaproteome identification.

1.5 Scope of Thesis

The research for this thesis sought to advance the field of metaproteomics and human medicine by systematically comparing the performance of different database builds to identify a standard practice that could yield more accurate protein and peptide identifications for a mouse gut model system. Mouse represents one of the most common model organisms utilized to study human disease (Hugenholtz, 2018). The field of metaproteomics can be improved if a best practice for generating protein sequencing databases to match observed spectra can be identified. The performance and accuracy of databases curated from deep metagenome sequencing followed by assembly have not been compared to all the databases that are currently employed in the field to identify proteins from microbiome samples. In this thesis, it was found that databases curated

from the translation of deep metagenome sequencing followed by assembly led to the highest number of protein and peptide assignments with comparable accuracy to other database types, and that utilizing different protein sequencing databases drastically affects the inferred function and taxonomy identified in a study.

The introductory chapter is the first of three chapters that make up this thesis. Chapter 2 introduces the basics for bioinformatic analysis in metaproteomics. Chapter 3 compares the performance of many database builds on previously obtained experimentally derived spectra from mice fecal samples (Blakeley-Ruiz, ISME Communications, 2022) and reports the total proteins and peptides identified, the accuracy in those assignments, and the function and taxonomic differences observed from each database.

CHAPTER 2

A DESCRIPTION OF THE BIOINFORMATICS TOOLBOX FOR METAPROTEOME RESEARCH

2.1 Introduction to Metaproteomics Data Analysis

Metaproteomics relies on pre-existing DNA and RNA sequencing information as the lookup table for assisting in assigning protein identities. Even though reference databases currently contain sequencing data for millions of organisms, querying all these databases will result in false identifications because protein sequences are shared across species and databases can have improper annotation (Starr, 2018). To mitigate this problem, direct metagenome or amplicon sequencing of an unknown environment can be used to construct a metaproteome database that accurately depicts what is expected to be in the sample. Once a database has been identified, proteins can be extracted from microbiome samples and run through LC-MS/MS instruments. After LC-MS/MS sequencing has taken place and the data has been collected, there are five main steps involved in processing the metaproteome sequencing data. These steps include the quality checking of mass spectra, database construction, database searching, data processing, and data display. Quality checking of sequencing data involves inspecting the chromatogram for errors that might warrant another extraction and sequencing run to be performed. Database construction involves utilizing metagenome and amplicon evidence from the sample to construct protein sequence databases for matching with observed spectra. Database searching involves the searching of a metaproteome database to identify and quantify protein and peptide groups. Data processing involves the filtering, normalization, and imputation of that data followed by statistical testing. Data display involves generating graphics that explain the function and taxonomic classification of the proteins and peptides in biological systems. Figure 1 below outlines this process.

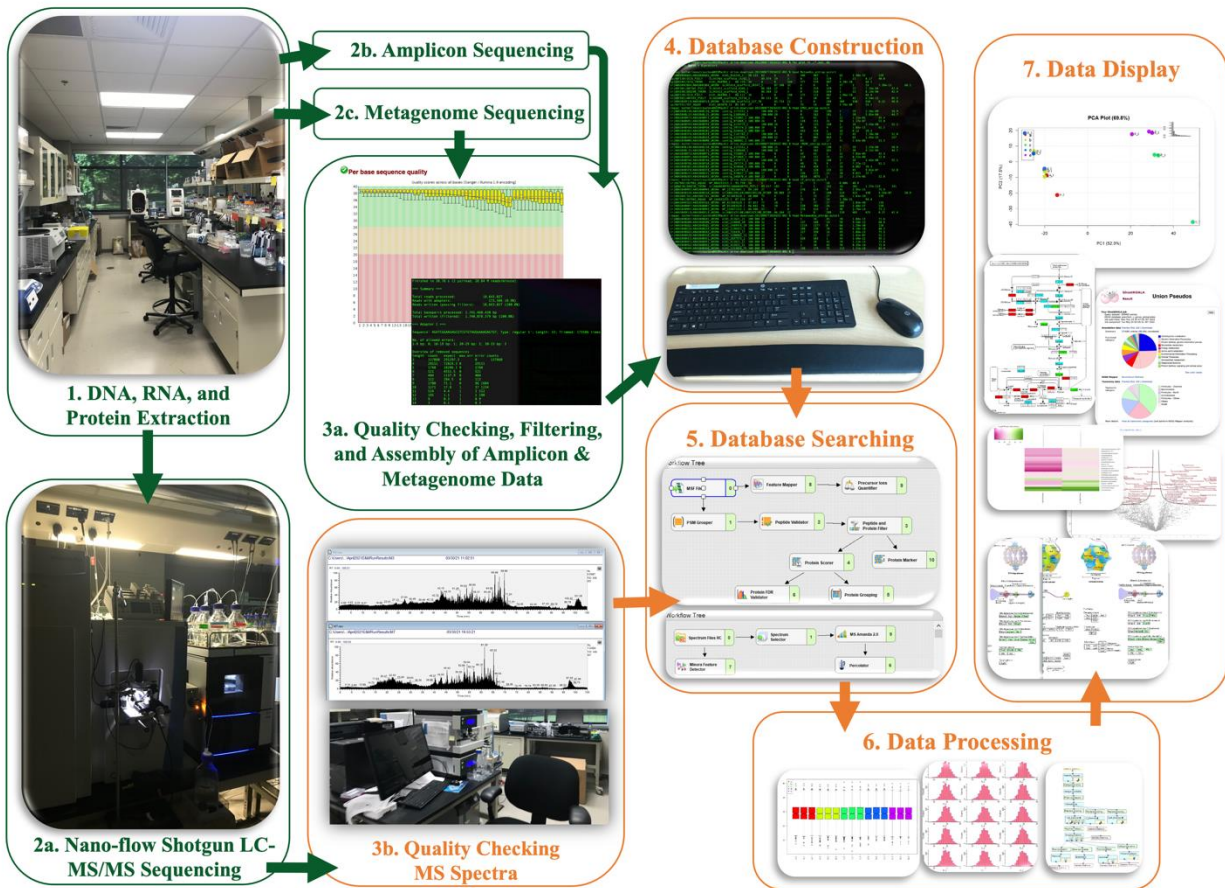


Figure 1 General metaproteomics workflow. This image depicts a typical metaproteomics workflow that is used to identify proteins from microbiome samples. The bioinformatics analyses discussed in this chapter will cover the steps in orange.

2.2 Quality Checking of Mass Spectra

While samples are being run through tandem mass spectrometry instruments, it is important to check the resulting data that is output from the instruments prior to processing the data to ensure that samples are being sequenced properly. Early identification of problems can prevent larger ones from arising if something were to go wrong. This process involves checking if the peptide sequencing data appears to have good coverage, is free of artifacts or noise, and if the coverage intensities match across replicates. The output spectral files can be visualized in software such as Xcalibur (Orsburn, 2021) and MaxQuant (Tyanova, 2015) that offer compatibility with a variety of output files types including .RAW. High coverage is obtained when the chromatogram contains many peaks that fill the entire width of the spectrum. Low coverage may appear to have separation at multiple places between peaks in the chromatogram. Low coverage may be a sign that an insufficient quantity of proteolytic peptides was loaded on the instrument. Artifacts are outliers within a chromatogram that may indicate contamination occurred during the run or during sample preparation. When replicates show different chromatograms, this can indicate that something ran differently and will need to be investigated. The MS instruments must be calibrated regularly to ensure that the peptide measurements are accurate. Typically, artifacts will appear across multiple samples in the same location if contamination occurred. High noise would also be an indicator of instrumental malfunction or poor sample cleanup. If acetonitrile was not properly removed during C-18 spin column cleanup, the peptide sample would be lost and thus the resulting mass spectra would be characterized by basically noise. High noise will preclude peptide identification in the sample and may lead to poor results. It is important to check for these problems as early in the process as possible to address why low protein identification and quantification may be observed, and it may even be necessary to ask for the samples to be run again.

2.3 Database Construction

In order to identify proteins from the resulting mass spectral data, a protein sequencing database must be constructed from the metagenomic information to match observed spectra to theoretical spectra that best represent the proteins that are expected to be found in the sample.

Since the environment is largely unknown, metagenomic evidence should be utilized to construct databases that accurately represent what might be in the samples. There are three main database types currently accepted in the field that will be discussed here which are the construction of matched-metagenomes, catalogs, and pseudo-metagenomes. Linux experience is recommended for constructing these databases because none of these methods have been made available through a graphical interface, and they all may require large amounts of storage and RAM to complete.

2.3.1 Matched-Metagenomes and Catalogs

The Human Microbiome Project initiated the creation of large sequencing repositories which have allowed for the discovery of many new microorganisms (Human Microbiome Project Consortium, 2012). These databases have grown significantly over the years, and they have greatly enhanced the genomic and proteomic coverage of the microbial members that inhabit this environment. However, because a lot of these efforts began by characterizing humans, a great deal of bias exists in these databases toward identifying only colonizers of this host system. When studying other microbiomes, reference databases are often missing. To fill this gap and eliminate biases in microbiome studies, many groups have generated repositories specific to a host system and performed deep metagenome sequencing of these environments (Zhu, 2021; Lesker, 2020). This sequencing information is recommended to be assembled into genes, but raw contigs may be utilized to create protein sequence databases. The translation of metagenome and catalog contigs into protein sequences can be accomplished from the command-line utilizing Prodigal (Hyatt, 2010) with the following script: “`prodigal -i dbcontigs.fasta -o db.gene.coords.gbk -a dbproteins.fasta`”.

2.3.2 Pseudo-Metagenomes

When metagenomic information specific to a system may not be possible to obtain, many labs have resulted to filtering reference repositories for specific organisms of interest. Some labs have relied on amplicon sequencing or metagenome sequencing

from similar studies to obtain a large reference database (Starr, 2018). The operational taxonomic units (OTUs), common names, and genome accessions that are identified for metagenome or amplicon datasets can be used as search criteria to obtain desired protein sequences. This can be achieved using two of the largest international protein sequence repositories, National Center for Biotechnology Information (NCBI) (Sayers, 2021) and UniProt (UniProt Consortium, 2021). It requires the use of the NCBI Datasets package (Sayers, 2021) and a custom Perl script for UniProt access (Nightingale, 2017). From the command-line, a list of OTUs or common names can be placed into a text file which then is searched against NCBI or UniProt to obtain and deposit metadata. Supplementary Figure 5 and Supplementary Figure 6 outline how to successfully download millions of protein sequences and annotate them with their corresponding organism identifier from RefSeq (a proteome resource from NCBI with quality-checked sequence annotations) and UniProt.

2.3.3 Clustering of Sequences

Prior to searching a database, it is important to remove sequences that may be redundant across the organisms which are found in a database. It has been suggested that proteins sharing common ancestors (orthologs) and proteins that developed from duplication events (paralogs) should be separated and clustered in such a way that the complexity of proteomes within a large database is roughly equivalent to avoid biases in the identification process (Zaslavsky, 2016). An accepted level of clustering is believed to lie somewhere above eighty-percent sequence similarity (Zaslavsky, 2016). Assigning a universal name to proteins that have a certain sequence similarity can allow for the identification of a protein sequence that might belong to multiple species (Xiong, 2015). This can be accomplished utilizing a command-line tool known as CD-HIT (Fu, 2012). The clustering of sequences at ninety-eight percent similarity can be achieved by running the following script: “cd-hit -i ./mydb_b4cdhit.fasta -o ./mydb_C98 -c 0.98 -n 5 -d 0 -M 100000 -T 0”. This script generates a report which can be searched to identify the sequences that overlapped at a certain threshold. In this way, no duplicate proteins exist in the database that will be searched against the spectra, and no taxonomic information regarding those sequences are lost during the identification process.

2.4 Database Searching

Once the spectra have been inspected and the databases have been curated, the raw mass spectrometry files can then be searched to identify and quantify proteins from the constructed protein fasta databases containing translated metagenomic information. Many programs exist for this specific task, and each have limitations for the types and quality of the searches they perform. Proteome Discoverer software is primarily utilized for untargeted, label-free data but requires a paid subscription (Orsburn, 2021). It has recently been updated to include SILAC and TNT labeling (Orsburn, 2021). MaxQuant is a comparable open-source version that is also capable of performing untargeted, label-free searches (Tyanova, 2016). Proteome Discoverer and MaxQuant are renowned for their application in detecting proteins with low concentrations (Palomba, 2021). Palomba and colleagues recently found that Proteome Discoverer still slightly outperforms MaxQuant with its MS Amanda 2.0 database search algorithm and its Percolator FDR control algorithm (Palomba, 2021). When searching for post-translational modifications (PTMs), however, another program is needed if multiple post-translational modifications must be identified. Peaks (Zhang, 2012) is a program that is used for identifying up to five hundred PTMs, while Proteome Discoverer is only capable of identifying a select few. In some cases, searching de-novo spectra in both Proteome Discoverer and Peaks is helpful to provide wider coverage. Many modified proteins can be missed with the standard database searching approach. A fourth program that is often used is called Skyline (MacLean, 2010) which searches targeted or labeled proteome spectra. This program is freely available on the web. Recently, a fifth software called Peptide Shaker by CompOmics has been developed that has integrated labeled, label-free, and PTM searching and has achieved favorable success comparable to these other programs mentioned (Vaudel, 2015). A downside to utilizing open-source software as a search engine would be that bugs readily occur and will need to be fixed from time to time. When utilizing subscription-based software, it is important to stay informed about the updates that regularly occur with these programs to ensure that they perform to an acceptable standard for specific applications.

One of the most heavily utilized databases across the proteomics community today for protein and peptide identification is the Universal Protein Resource (UniProt) (UniProt Consortium, 2021). Within its integrated database, UniProt provides one of the largest databases for protein and peptide identification to date. It provides the user with access to BLAST protein alignment searches that identify proteins with ranked similarity scoring. It also allows for the identification of associated DNA and RNA molecules using protein accessions as input through its “retrieve ID/mapping” function. The limitation of this tool is that it does not provide direct resources for visualizing biological systems from multiple proteins or peptides.

This thesis primarily utilized the Proteome Discoverer software for protein and peptide identification and quantification (Orsburn, 2021). The general setup for a Proteome Discoverer workflow includes uploading UniProt or other outside proteome databases, a contaminants database, and .RAW spectral files followed by setting MS1, MS2, & FDR tolerance; adding the MS Amanda 2.0 and Percolator algorithms, adjusting missed cleavages; adjusting static and dynamic modifications; and selecting a protease. A commonly used contaminants databases can be found at the MaxQuant website (Beer, 2017). The addition of a contaminants database to the search allows for contamination to be identified and filtered out in a later step.

Accurate m/z measurements of the parent peptides is achieved in the MS1 segment of the mass spectrometric experiment, and the related fragmentation spectra (achieved by collision-induced dissociation) are obtained in the MS2 segment of the experiment. The MS1 mass accuracy is generally ± 10 parts-per-million (ppm), which provides solid measurements for both parent peptide and fragmentation information at an acceptable false discovery rate. The false discovery rate can be set to a certain threshold (usually 0.01) to obtain only identifications that are less than a specified confidence level. The newest MS Amanda 2.0 and Percolator algorithms are recommended to achieve the highest coverage of samples (Palomba, 2021). Due to the unknown components in a microbiome sample, complete digestion of peptides with a protease may not always be successful and missed cleavage events can occur. Missed cleavages by a protease typically start at a factor of two but can be widened for a larger search, up to four in PD workflow. A dynamic modification of oxidized methionine should always be searched for, as protein sample preparation readily exposes the samples to oxygen. Secondly, a static

modification of carbamidomethylation should be added to account for cysteine labeling to prevent disulfide bond reformation. The most common protease utilized for most digestions would be trypsin since it cuts peptides at Arginine and Lysine residues and typically generated peptides containing between four to eleven amino acids. This protease is set as the default in the program, but other types of proteases can be selected. The database searching workflow using Proteome Discoverer is shown in Figure 2 on the next page.

2.5 Data Processing

Data processing involves the filtering, normalization, imputation, and statistical testing of obtained protein and peptide identification data. A filtering process is performed on the obtained results to adjust factors that are of interest in particular studies such as the FDR, obtaining only master proteins, or removing missing values. Normalization in metaproteomics involves correcting variance across replicates that may have occurred before or during the sequencing process. Metaproteomics data often contains missing measurements which are typically imputed to account for a limit of detection biases. Lastly, statistical testing is performed to determine groupings in a study and to identify factors that separate the samples.

2.5.1 Filtering

The first step in data processing is to export a result file from Proteome Discoverer in text tab delimited and open it in Excel. Processing in Python or R may be required later, so renaming the columns with underscores to omit blanks in the titles is preferred. To perform filtering, some or all of this process can be done in Excel, Python, or R. A Markdown file can be helpful to keep track of processing steps, and this avoids modifying the original file and causing permanent damage that may result if Excel is utilized. Although processing in R or Python may take longer to deal with different file types as opposed to processing in Excel, it ensures that mistakes are identified and corrected along the way.

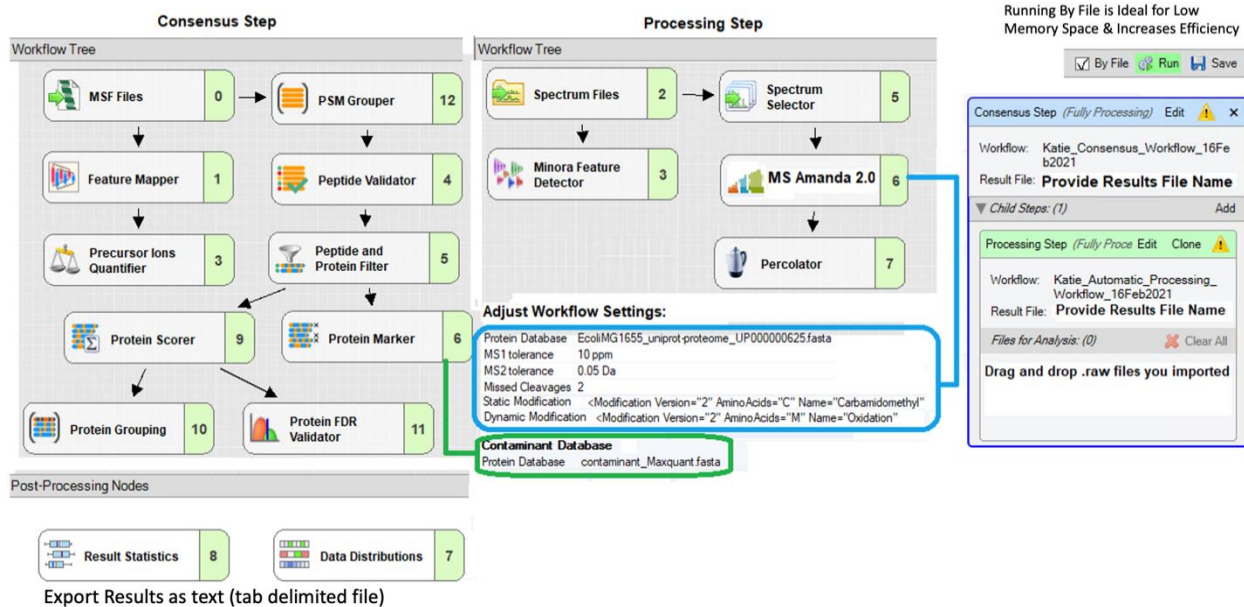


Figure 2 General database searching workflow in Proteome Discoverer. The following nodes can be selected in the Proteome Discoverer software for searching obtained sequencing information. In the Protein Marker node, a contaminants database can be added. In the MS Amanda 2.0 search algorithm, a database can be added with a specified level of tolerance along with modifications. Spectral files can be downloaded and dropped into the workflow. The “By File” setting allows the process to run as a batch. The standard protease used during the extraction process is set to Trypsin by default, but this can be changed.

Before filtering is performed, the sample types must be identified and grouped accordingly. If everything is the same, then it can all be filtered together as one file. However, if more than one sample type is present, the protein abundance values will need to be separated into individual files for filtering. Selectively filtering out rows will prove challenging if multiple rows require omission so this separation should occur prior to this process.

Filtering Proteome Discoverer output involves removing high q-values, low confidence identifications, proteins that have not been experimentally confirmed, contamination, low protein sequence coverage, and rows with excessive missing values. Q-values refer to the lowest FDR value obtained for a particular identification. Q-values of greater than 0.05 are the standard value that must be maintained in all protein quantifications. Anything greater than this value should be filtered out. Low confidence identifications refer to any protein quantification that exceeded the FDR of 0.05. “IsMasterProtein” signifies that a quantified protein in Proteome Discoverer has strong experimental evidence (Orsburn, 2021). Protein referred to as “NotMasterProtein” or “PredictedMasterProtein” should be filtered out because there is not experimental evidence to support that it qualifies as a protein. In Proteome Discoverer, contamination identified by the contaminants database are labeled as “TRUE” under the Contaminants column. It is recommended to first check whether the proteins labeled as contaminants fit this criterion prior to their removal. Additionally, if a contaminants database is not specified in the program and one is added separately, grep on the command-line can be used to remove contaminants from output text files. Low protein sequence coverage may also be omitted in this filtering step. Missing values occur in proteome datasets due to an instrument missing a measurement or due to a peptide falling outside the minimum or maximum limit of detection. Certain experiments might recommend a specific threshold for filtering this data. Figure 3 on the next page outlines R code that can be utilized for the filtering of obtained Proteome Discoverer output.

Read Excel file & Save Usable Columns as Variable

```
PDOut <- read_excel("ProteomeDiscovererOutput.xlsx", col_names = TRUE)
```

```
A <- PDOut[ c(1:19) ]
```

Count Total Protein IDs By Excluding Rows Missing All Measurements for a Protein ID

```
A <- WcoagA[rowSums(is.na(WcoagA[17:19])) >= 2, ]
```

```
cat("Dimensions of A dataset before & after:", dim(WcoagA), dim(A), "")
```

Output: Dimensions of A dataset before & after: 2381 19 2017 19

Remove Bad q-values that exceed 0.05 & Count Them

```
cat("Q vals in A greater than 0.05, less than 0.05:", length(which(A  
Exp_qvalue_MS_Amanda2 > 0.05)), length(which(AExp_qvalue_MS_Amanda2 <= 0.05)), "")
```

```
Aq <- A[A$Exp_qvalue_MS_Amanda2 <= 0.05,]
```

Output: Q vals in A greater than 0.05, less than 0.05: 322 1695

Remove Contaminants & Count Them

```
cat("A Contaminants, A w/o contaminants", length(which(AContaminant == "TRUE ")), length(which(AContaminant == "FALSE")), "")
```

```
Aqc <- Aq[Aq$Contaminant == "FALSE",]
```

Output: A Contaminants, A w/o contaminants 71 1946

Remove if False Discovery Rate > 0.05 & Count Them

```
cat("A FDR >0.05, A <= 0.05 FDR", length(which(A  
Protein_FDR_Confidence_MS_Amanda2 == "Low ")), length(which(AProtein_FDR_Confidence_MS_Amanda2 != "Low")), "")
```

```
Aqcf <- Aqc[Aqc$Protein_FDR_Confidence_MS_Amanda2 != "Low",]
```

Output: A FDR >0.05, A <= 0.05 FDR 323 1694

Keep "Is Master Protein" Only & Count Them

```
cat("A Predicted/Not Master Proteins, A Master Proteins", length(which(AMaster! == "IsMasterProtein ")), length(which(AMaster ==  
"IsMasterProtein")), "")
```

```
Aqcfm <- Aqcf[Aqcf$Master == "IsMasterProtein",]
```

Output: A Predicted/Not Master Proteins, A Master Proteins 29 1988

Keep if Protein Sequence Coverage >= 50% & Count Them

```
cat("A prots < 50% sequence coverage, A prots >= 50 perc sequence coverage", length(which(A  
Coverage_Percent < 50)), length(which(ACoverage_Percent >= 50)), "")
```

```
Aqcfmpr <- Aqcfm[Aqcfm$Coverage_Percent >= 50,]
```

Output: A prots < 50% sequence coverage, A prots >= 50% sequence coverage 1880 137

Separate Rows Missing > 25% of Data, Save Rows Missing >25% of Data for Later

```
Aqcfmpr <- Aqcfmpr[rowSums(is.na(Aqcfmpr[17:19])) < 2, ]
```

```
A_missvals <- Aqcfmpr[rowSums(is.na(Aqcfmpr[17:19])) > 2, ]
```

Merge Datasets and Write .txt File for InfernoRDN

```
mby3 = intersect(colnames(Aqcfm), colnames(Bqcfm)) #get like column names
```

```
Mqcfmpr <- merge(merge(merge(Aqcfmpr, Bqcfmpr, by = mby3, all= TRUE), Cqcfmpr, by = mby3, all= TRUE), Dqcfmpr, by = mby3, all=  
TRUE), Eqcfmpr, by = mby3, all= TRUE) #perform multiple database merges using like column names
```

```
write.table(Mqcfmpr, "prelogvals.txt", eol="\n", sep="\t", col.names = TRUE) #write table
```

Figure 3 Filtering of protein expression data. This figure display R code that can be utilized to filter Proteome Discoverer identified proteins output.

2.5.2 Normalization

After filtering, all the files will need to be merged as a consensus to allow for normalization across all the samples. The last lines in Figure 3 on the previous page outlines how this can be accomplished. Normalization is performed across all the samples to account for sample preparation and instrument-related bias. If comparisons are to be made across the dataset, this bias needs to be accounted for prior to analysis.

Normalization can be performed in an automated open-source software developed by Pacific Northwest Laboratory called InfernoRDN (Polpitiya, 2008). The expression file with the UniProt Accession names in the “Unique Row ID” column and the abundance readings in the “Data Columns” tab are imported. Groupings are then defined, and Log 2 transformation of the abundance values is performed. Log 2 transformation is necessary to visualize comparisons between protein expression data because the changes are extremely variable. The abundance values should then be visualized in histograms to determine if the dataset is unimodal with a relatively normal distribution. If the dataset fits this criterion, then a loess normalization can be performed followed by a central tendency adjustment of the mean. If the dataset is not normally distributed, normalization methods such as EigenMS would be more applicable (Välikangas, 2018). The setup for this normalization in InfernoRDN can be visualized in Figure 4 on the next page.

2.5.3 Imputation

Imputation is a widely used tool in the metaproteomics community to deal with missing measurements that arise from sequencing. Values missing completely at random in metaproteomic datasets can either be due to low loaded sample concentration or samples that fall outside the limit of detection of the instrument (Hediyeh-zadeh, Webb, & Davis, 2020). There are two well-accepted methods currently for imputing missing values in proteome data that include k-nearest-neighbors (kNN) (Zhang, 2016) and random forest (RF) (Touw, 2013). The kNN method is a supervised regression machine learning technique that measures the Euclidean distance between variables in a matrix and assigns a value that is the closest neighbor. If there are more than one that tie, the

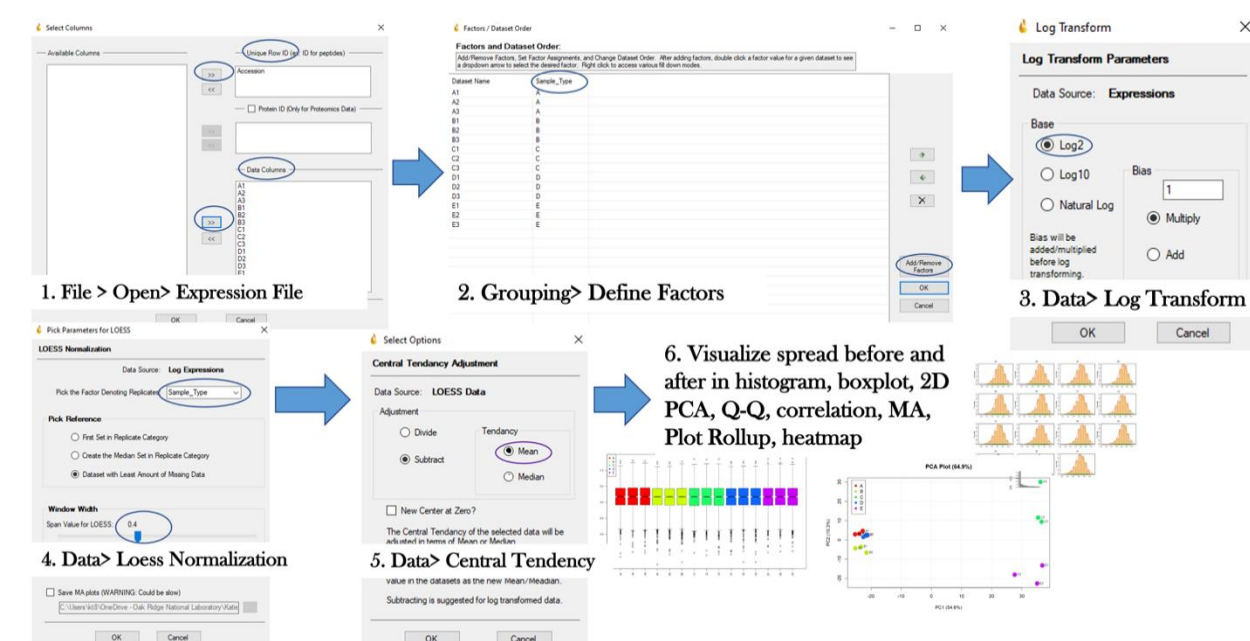


Figure 4 Normalization of protein expression data. This figure displays the typical workflow employed in metaproteomics for normalizing mass spectra data.

values are averaged to impute the missing value. RF is also a supervised method that takes multiple randomly selected samples of the dataset to build decision trees. These randomly selected samples are then continuously split with the addition of a random value until the entropy reaches zero. The final values reached are called the prediction. The predictions at the end of each tree are averaged to result in the imputation of a missing value.

For smaller datasets, kNN is typically used because the algorithm requires a lot of memory space to compute. It is the most common imputation method utilized because it has been implemented in automated software such as Perseus and InfernoRDN. It avoids exceeding the limit of detection by only imputing averages of the training data. A downside to this imputation method is that outliers can be amplified with this method and overfit the dataset. RF is recommended for larger datasets with greater missing values, has a faster processing speed, and reduces the effect of outliers within a dataset during imputation leading to less overfitting. These methods can be applied to a dataset in R using the MissForest package (Stekhoven, 2012) or the SciKits learn Python package (Pedregosa, 2011).

The most widely accepted method for imputation using kNN or RF is to impute missing values with a downshift from the mean (Hediyeh-zadeh, Webb, & Davis, 2020). This is because the probability of missing values falling into the category of low sample concentration is high, and they should be treated as values that fall slightly lower than other replicates per condition. To accomplish this, imputation is typically performed using a 1.8 to 2.8 downshifted variation from the mean using the kNN algorithm in Perseus, “replace missing values from a normal distribution”, on the samples that have less than a third of data missing (Kawashima, 2022; Woo, 2021). Replacing values 1.8 to 2.8 standard deviations from the mean allows for values to be imputed that fall roughly halfway between zero and other real measurements. This method of imputation follows the assumption that most missing values are missing at random are due to low protein concentration (Hediyeh-zadeh, Webb, & Davis, 2020). By confining the data within the limits of detection, it ensures that zero values do not lead to exaggerated differences in expression during analysis. A comparable open-source option for this imputation is

available on GitHub as msImpute developed by Hedyeh-zadeh, Webb, and Davis in 2020.

The imputation analysis in Perseus is performed by importing an expression file in text tab-delimited format with the accession names and normalized expression values. Missing values should be left blank for import as it does not recognize “NA” values, and this can quickly be done through replacement in Excel. The second step is to name groupings by sample type. Next, selecting the impute tab followed by “replace missing values from a normal distribution” should be selected. A downshift from the mean of 1.8 to 2.8 can be used followed by selecting either the total matrix or specific columns. The distribution of the data should then be visualized in a histogram to inspect normality. Finally, new groupings can be set to perform volcano plot visualization and a T-test among columns of interest. Figure 5 on the next page outlines this process.

2.5.4 Statistical Testing

Knowing when to utilize a specific statistical test is important in data analysis, and it depends on the hypothesis being tested. In order to claim that statistical differences exist between samples, a T-test or ANOVA analysis should be used. A T-test is used to compare the averages for only two conditions, and ANOVA is used to compute the averages between three or more conditions while retaining a level of significance. It is possible to compare multiple conditions using multiple T-tests, but the problem with this approach is that each individual test incorporates some amount of Type-1 error. Type-1 error occurs when the statistical difference is mistakenly rejected. When performing ANOVA, Type-1 error is only incorporated once. This is desirable for multiple group comparisons because it retains the desired level of significance. One-way ANOVA has one independent variable such as abundance, with the accession being the dependent variable. Perseus and InfernoRDN have automated workflows for performing one-way ANOVA or T-tests.

Before a T-test or ANOVA can be performed, the groupings of the samples need to be defined for the comparisons. This is done by generating Euclidian distance, Pearson’s, and Spearman’s correlation plots, along with Principal Component Analysis (PCA) and Partial Least-Squares Discriminant Analysis (PSL-DA) plots that can be generated in Perseus, InfernoRDN, R, Python, and multiple other open-source software. Euclidean distance correlation plots

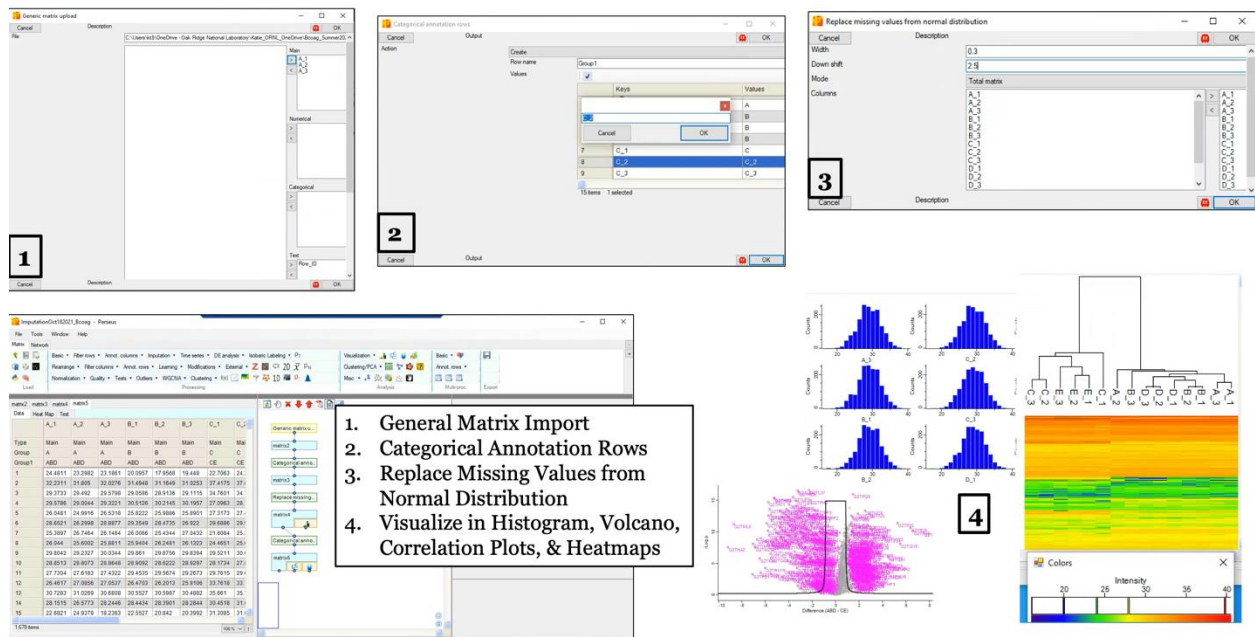


Figure 5 Imputation of protein expression data. This figure displays the basic Perseus Imputation workflow and data visualization options.

measure the Pythagorean distance between datapoints in a matrix that are used to explain the differences observed between samples. Pearson's coefficient is a measure of the similarity of the cosine across angles between datapoints and uses linear-regression. Spearman's coefficient measures the non-linear similarity across samples utilizing rank-ordered values. For datasets that have linear relationships, a Euclidean or Pearson's correlation plot is appropriate. For datasets that have more non-linear relationship, Spearman's correlation can be used to determine how similar two variables are. It is important to note that Spearman's cannot be used to provide statistical significance given that it uses a rank ordered metric. PCA plots are generated using an unsupervised method as well to visualize how like samples cluster. PSL-DA plots utilize a supervised method to show the differences between samples.

Once groupings have been identified, it is often useful to generate volcano plots between groups which utilize a t-test to determine the amount of protein expression that differs between groups. Correlation plots may also be used. A fold-change greater than two is generally considered acceptable for generating volcano plots. Finally, the volcano plot data should be exported into a text document for later visualization in Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), or other databases. In 2016, Gianetto and colleagues suggested that adjusting the fold change parameter in volcano plotting distorts the p-value. Even the smallest of statistically significant fold change should be considered when mapping. Adjustments to this fold change can be made if differences are not well defined.

2.6 Data Display

The final step in Proteome Data Analysis involves performing statistical testing to identify groupings of expression data and to use software to visualize how each sample is unique or similar. This is the most time intensive process in data analysis because it involves searching different databases to extract information relative to each identified protein. It is important to understand the bias that each database contains to search the appropriate databases of interest for more information on identified proteins. Searching the wrong databases can heavily bias results. Knowledge of chemical and biological processes that occur in nature are critical to extracting the appropriate proteins that are most relevant to a study.

The first step that must be done prior to searching expression data in a database is to convert proteins names to other identifiers so that they can be mapped. To do this a list of proteins can be uploaded, one per line, into the Uniprot Covert tool (Huang, 2011). The UniProtKBAC/ID to UniProtKB is then selected for conversion and the job is submitted. By clicking to add columns, other identifiers can be view and exported into .xlsx, .txt, or .fasta. These identifiers should then be merged with all the data in R or Python.

The first most widely used database for assigning protein function is the Gene Ontology (GO) database (Mi, 2019). This database contains information on metabolic pathways, localization, and cross-species relationships of proteins and genes. It is one of the most used sources for mapping the function of molecules in biological systems. The limitation of this tool, however, is that it remains heavily biased towards model organisms, and the classification of novel organisms is limited.

A second database that is often used for displaying functional information from protein expression data is the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2017). This database contains information focused on assigning molecular function to proteins. It can be used to generate models of biological systems that display information related to systems, genomes, chemicals, and health. This database is especially useful in functionally annotating proteomes due its use in reaction mapping. It outperforms GO by displaying the names of proteins and utilizes reaction arrows to designate how certain reactions among proteins and peptides proceed in nature. It also has a larger representation of non-model organisms, but a bias towards model-organisms still exists. The limitation to this tool is that it does not provide the localization of proteins within the cell, but other tools such as GO or SignalP (Almagro, 2019) can perform this function.

For color mapping reaction pathways in KEGG, the volcano data is assigned colors to the upregulated, downregulated, and equally expressed protein for mapping expression that fell above or below a specified fold change. Map comparisons are made across every relevant system, and the proteins from pathways of interest are extracted. An

example of input and output files from color mapping in KEGG is shown in Figure 6 on the next page. A literature search on identified proteins within these maps are necessary to confirm the pathways specific proteins are involved in because not all map associations may correspond to every function a protein may perform in a biological system.

Other databases such as Carbohydrate Active Enzymes (CAZy) (Lombard, 2013) or GO might be necessary to extract metabolic information that might be missing if only KEGG was used for functional annotation. Each database has a unique subset of organisms and metabolic processes that are represented. Mapping to multiple databases is often necessary to obtain a high coverage of the functional processes that may be occurring in a biological system.

GOMixer is an web-based platform for R that accepts data with text tab delimited format using the accession, description, KEGG identifiers, and abundance values for input (Darzi, 2015). It outputs a visualization of the over and under abundant proteins in metabolic maps (Darzi, 2015). A second mapping tool that requires a subscription is OmicsBox (Götz, 2008). It is one of the most widely used programs for visualization because it can perform annotation using .fasta protein sequence files through its integration of EggNOG Mapper (Dias, 2020). It is also capable of performing BLASTP and BLASTN which makes it well suited for multiple omics sequencing datasets (Dias et. al., 2020). Another relatively new mapping tool developed by Wang and Dassanayake in 2020 is the GO Markov-Clustering (GOMCL). It is also a GO-based Python toolkit that visualizes .obo expression files from BinGO, Gorilla, g:Profiler, or agriGO (Wang & Dassanayake, 2020). After clustering is performed, the GO terms can be visualized in heatmaps, networks, and hierarchical clustering plots (Wang & Dassanyake, 2020). GO-Causal Activity Modeling (GO-CAMs) is a fourth tool that is also widely used for lists of GO identifiers (Thomas, 2019). It generates maps of molecular function, metabolic function, and cellular localization (Thomas, 2019).

2.7 Conclusions

The goal of metaproteomic studies is to identify phenotypes expressed in the environment and determine their origin. This is accomplished by viewing from a systems perspective all proteins expressed in the environment and determining which organisms and proteins are the driving factors that separate two conditions. This chapter presented a systemic discussion of the

factors driving many bioinformatic analyses that can achieve this goal and highlighted several techniques that will be covered in the subsequent chapter.

CHAPTER 3

EVALUATING MEASUREMENT DEPTH, ACCURACY, AND FUNCTIONAL ASSIGNMENT IN METAPROTEOMICS USING MATCHED-METAGENOMES, CATALOGS, AND PSEUDO-METAGENOMES IN MICE GUT MODEL

Authors for future manuscript

Katherine E. Ostrouchov, J. Alfredo Blakeley-Ruiz, Him Shrestha, Robert L. Hettich

Katherine E. Ostrouchov contributions include all analyses and writing of the original manuscript.

J. Alfredo Blakeley-Ruiz contributions include project conception and construction of the Python script for counting Proteome Discoverer protein and peptide identifications and interpretation of figures.

Him Shrestha contributions include project conception and interpretation of figures.

Robert L. Hettich contributions include project conception, interpretation of figures, and edits to the original manuscript.

ABSTRACT

It is challenging to accurately characterize microbial systems in metaproteomics due to several factors such as the metagenome relevance, completeness, assembly at a species level, and sequence redundancy which all affect the proteins that are identified from these studies. To identify proteins from microbiome samples, a reference database must be made available to match theoretical protein spectra with observed spectra. The proteins that are identified are directly influenced by the database that is chosen for this identification. Many methods have been proposed for database construction, but none have addressed how the accuracy of utilizing a high-depth assembled metagenome might compare to reference databases for microbiomes. To help inform the field to move towards a standard practice that provides the highest identifications of proteins and peptides with the lowest false identification rate for microbiome investigations, this study compared the performance of the current database types. It was found that the

metagenome translated databases provided the highest number of protein and peptide identifications along with the highest amount of functional and taxonomic information with a false identification rate consistent with other database types.

3.1 Introduction

It is now well known that microbial life helps to support the growth of healthy eukaryotic life. They allow nutrients to be accessed and contribute to the prevention of disease.

Metaproteomics has emerged as one of the leading sequencing technologies capable of characterizing these systems. Metaproteomics, however, would not be possible without the efforts made in metagenomics, the study of all genes. Metagenomics has led to the assembly of millions of new genomes, some of which were unculturable in the past, many of which have been deposited in reference databases. The translation of these assembled genomes has led to the curation of many large databases of protein sequencing information which can now be used to identify proteins using mass spectrometry. Metagenome sequencing and annotation used to be a very expensive and time-consuming task, but over the years this technology has advanced significantly and led to more affordable services. Additionally, large reference databases now exist which can allow millions of proteins to be identified within samples.

With all this power it would seem like merging all these databases would lead to more identification in the field of metaproteomics. However, it has become increasingly more apparent that organisms share protein sequences, annotations in reference databases can be incorrect, and that matching proteins to their correct protein and organism from a mixture can present a multitude of challenges that negatively affect the accuracy and coverage obtained from these types of experiments.

Currently, there is no accepted method capable of determining the accuracy of protein assignments and their origin for complex environmental microbiomes when reference databases are utilized. Due to these shortcomings in metaproteomics, it is important to create databases that accurately depict members expected for unknown peptide mixtures to minimize the risk of obtaining false positives. Even if sample mixtures are unknown, specific techniques can be used

to create databases that most accurately reflect the proteins that might be present in a sample. There are four types, and they each have different levels of metagenome evidence.

The first are matched-metagenomes generated from performing metagenome sequencing on the samples tested which are then translated into protein sequences. Due to time constraints, difficulty in assembly, cost of sequencing, or difficulty in extracting sufficient DNA, this database may not always be accessible for studying certain systems. It is less clear whether sample-derived metagenome databases constructed from deep metagenome sequencing and assembly may lead to the best characterization of microbiomes regardless of type or complexity. The second database type is known as a catalog which represents exhaustive metagenome sequencing of a specific host environment that is typically deposited in online repositories for others in the microbiome community to obtain (Lesker, 2020; Zhu, 2021). These databases come close to representing a host environment and are quick to obtain. However, genomes have been found to be host-specific, and the assembly and sequencing technology can differ across studies which deposit this data leading to inconsistent or erroneous protein annotations. When neither of these databases are available, a third database type that is used is called a pseudo-metagenome. These databases use any relevant repository information that contains some level of experimental evidence to grab proteomes of interest. This is one of the fastest ways to generate a database. However, this database contains very little experimental evidence, and, as such, may lead to an unknown number of false positive protein assignments (Nath, 2021; Starr, 2018). Lastly, a fourth database that is utilized contains either one of the above databases, or a combination, and then performs what is known as a multi-step-search. These databases utilize identifications from a preliminary search to then remove sequences from the reference database that might be more relevant. The time in database curation can drastically vary depending on the types of databases that are used. Using a multi-step-search strategy has been shown to lead to more identifications, but some studies have shown that this method increases the FDR (Vaudel, 2012).

Xaio and colleagues in 2018 found that combining matched-metagenomes with pseudo-metagenomes increased protein and peptide identifications. For samples that come from the same location in the environment and have been extensively sequenced, catalogs have been shown to provide comparable coverage to that of a matched-metagenome (Kleiner, 2012). It was also found that omitting true sequences from a database that were present in a sample had a much larger effect than adding large pseudo-metagenomes to other databases (Tanca, 2016). Although each of these studies have highlighted that reference databases can provide a larger number of identifications, the accuracy of utilizing all database construction methods in metaproteomics for unknown microbial communities has not been compared (Starr, 2018). Furthermore, the performance of a high-depth assembled metagenome (i.e., greater than 18 million base pairs for fecal microbiomes) has not been tested against the other database types (Tanca, 2016). This study chose to construct the current databases available in metaproteomics while utilizing a deeper metagenome measurement and compared their accuracy in identification as well as the functional and taxonomic information gained utilizing six mice fecal sample datasets obtained by Blakely-Ruiz and colleagues in 2022 ([ISME Communications](#)). Two treatment groups were tested belonging to mice fed a high-fat or a fiber-rich diet. These two treatment groups were selected to compare the variance that might be observed across different habitats.

It was hypothesized that the high-depth matched-metagenomes would obtain the highest number of accurate identifications and deepen the functional and taxonomic information gained from these datasets. This study set out to evaluate and quantify if/how metaproteome databases constructed from metagenomes sequenced at 250 million base pairs per sample obtained more accurate protein and peptide identifications and increased the level of functional as well as taxonomic information gained compared to all database types. The results suggest that microbiome studies should first attempt to utilize matched-metagenomes if deep metagenome sequencing followed by assembly is possible since results are largely driven by the database selected and because reference databases lack cross-validation from other omics.

3.2 Materials and Methods

This study utilized previously obtained metagenome, amplicon, and metaproteome sequencing data from the Hettich lab which is currently in press for publication (Blakely-Ruiz, [ISME Communications](#), 2022). The details of the study design, database construction, search parameters, and bioinformatic analyses performed for each figure are discussed below.

3.2.1 Study Design

For this study, high-depth metagenome, amplicon, and metaproteome sequencing datasets were obtained from a previous study conducted in this group that compared the administration of high-fiber diet, high-fat diet, and morphine in mice (Blakely-Ruiz, [ISME Communications](#), 2022). The data collected from three of the mice fed the high-fat diet (samples L, I, and T) from the previous study, and three of the mice fed the fiber-rich diet (G, H, P) were utilized for this study. Figure 7 on the next page details this process. The 16S rRNA extracted from fecal material was sequenced at 2 by 250 nucleotides on an Illumina MiSeq with the V2 500-cycle MiSeq Reagent Kit which identified a total of 1,197 amplicon sequence variants (ASVs) using the Silva database. The DNA from surgically collected colonic fecal material was sequenced via Illumina NovaSeq at 250 million nucleotide base pairs per sample, assembled, and translated with prodigal to generate the matched-metagenomes for this study. The metagenome consisted of 460 assembled mice gut bacterial genomes, of which ninety-three percent of those genomes were not represented in RefSeq or GenBank. 2D-liquid chromatography tandem mass spectrometry was applied to sequence the peptides extracted from these same samples with four fractions, sequencing runs, per sample. A total of twenty-four datasets obtained from sequencing six mice fecal samples in the fiber-rich/saline and high-fat/saline group were utilized for the analysis in this study.

3.2.2 Database Construction

Databases were constructed to model all the previously described database choices in the field. As a starting point, the microbial proteins needed to be identified. To accomplish this, the

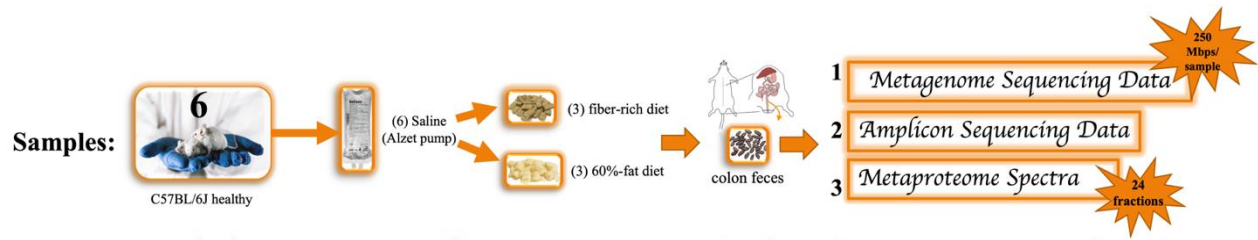


Figure 7 Study design. Sequencing data was obtained from a previous study conducted in the Hettich lab (Blakely-Ruiz, [ISME Communications](#), 2022). These diets were administered for a total of 2 weeks, and then mice were euthanized. Fecal samples were surgically extracted from the colon. DNA, 16S-rRNA, and peptides were extracted from these samples and run on an Illumina NovaSeq, Illumina MiSeq, and Orbitrap mass spectrometer.

matched-metagenomes previously constructed by Blakeley-Ruiz et. al. in 2022 were obtained for the microbes, and the databases with and without assembled genomes were utilized. A multi-step search database constructed previously by Blakeley-Ruiz et. al. in 2022 that reduced the proteins which were not identified from the matched-metagenome after performing an initial search in Crux/Tide with Sequest were also obtained. The most recent iMGMC and EMGC mice gut microbiome catalog contigs were translated with prodigal to generate protein databases (Lesker, 2020; Zhu, 2021). The EMGC catalog incorporated iMGMC contigs, but it did not contain metagenome sequencing information obtained from the Blakeley-Ruiz et. al. study in 2022 (ISME Communications).

For the pseudo-metagenomes, cloud computing resources and command-line scripting was used to download thousands of reference bacterial proteomes from UniProt and RefSeq associated with the intersection and union of operational taxonomic units (OTUs) identified from the amplicon and metagenome datasets. These lists confined the search space to only taxonomic members which were represented in both the amplicon and metagenome. The intersection of the two pseudo-metagenomes were meant to represent a reference database that had the highest level of confidence, while the union represented a broader subset meant to grab as many relevant OTUs indicated by both datasets. A total of 18 families, 23 genus, and 3 species were shared between the amplicon and metagenome which were utilized to obtain the Intersection Pseudo-metagenomes (IPseudos) from both RefSeq and UniProt. A total of 325 genus and 192 species were unique to both the amplicon and metagenome which were searched to obtain the Union Pseudo-metagenomes (UPseudos) from both RefSeq and UniProt.

To test whether nonsensical proteins could be identified in a metaproteome search, a custom entrapment database was generated that contained proteins that should not be found in the samples. This database consisted of the eukaryote, *Ursus maritimus*; two archaea, *Hyperthermus butylicus* and *Pyrodictium occultum*; and three bacteria, *Aquifex aeolicus* and *Thermotoga maritima* and *Pseudothermotoga lettingae*. All species lived and grew at temperatures and regions not suitable for growth in mice cages in Knoxville, Tennessee. Although past entrapment databases have focused on

eliminating all sequence redundancy with search databases in metaproteomics, this study took a different approach and chose to keep those sequences which were redundant because they represented real redundancy that might exist in current databases (Feng, 2017). Due to this redundancy and later discovered non-zero identification rate obtained from searching this database against the sample's spectra, FDR was not to be calculated using this database. Merely, the entrapment database was meant to serve as a test to see if false positives would occur, and the standard reverse decoy database was used to calculate FDR as is usually employed.

Lastly, because all the search databases needed to represent everything that should be present in a sample and not just the bacteria, the 2022 UniProt mouse proteome was added to each database that represented the host that should also be present in the samples. After the mouse database was added, all databases were clustered at ninety-eight percent sequence similarity with cd-hit to match that of the clustering in the previous matched-metagenome. This ensured that there was a two percent protein sequence difference between all proteins in a database. A report was generated from cd-hit which kept the names of proteins with similar sequences so that they could later be traced back to the many organisms that might match to a specific protein sequence. This helped to eliminate the bias of one peptide matching to multiple proteins from different organisms.

Lastly, the cRAP contaminants database was added to identify and remove contaminants. Figure 9 on the next page shows the resulting database sizes. For reference, one of the largest databases constructed in this study exceeded the size of all UniProt reference bacteria and was about a fifth the size of all RefSeq reference bacteria from January of 2022 (Figure 9). Due to a limit in available RAM on the virtual machine used in this study, the larger databases above 15 million sequences were not searched but remain assembled for future analysis. Details outlining this construction process can be visualize in Figure 8 on the next page.

3.2.3 Search Parameters

The searches were performed in Proteome Discoverer (PD) 2.5. Parameters involved Sequest HT search algorithm, Percolator for FDR calculation, 2 missed cleavages, C terminal

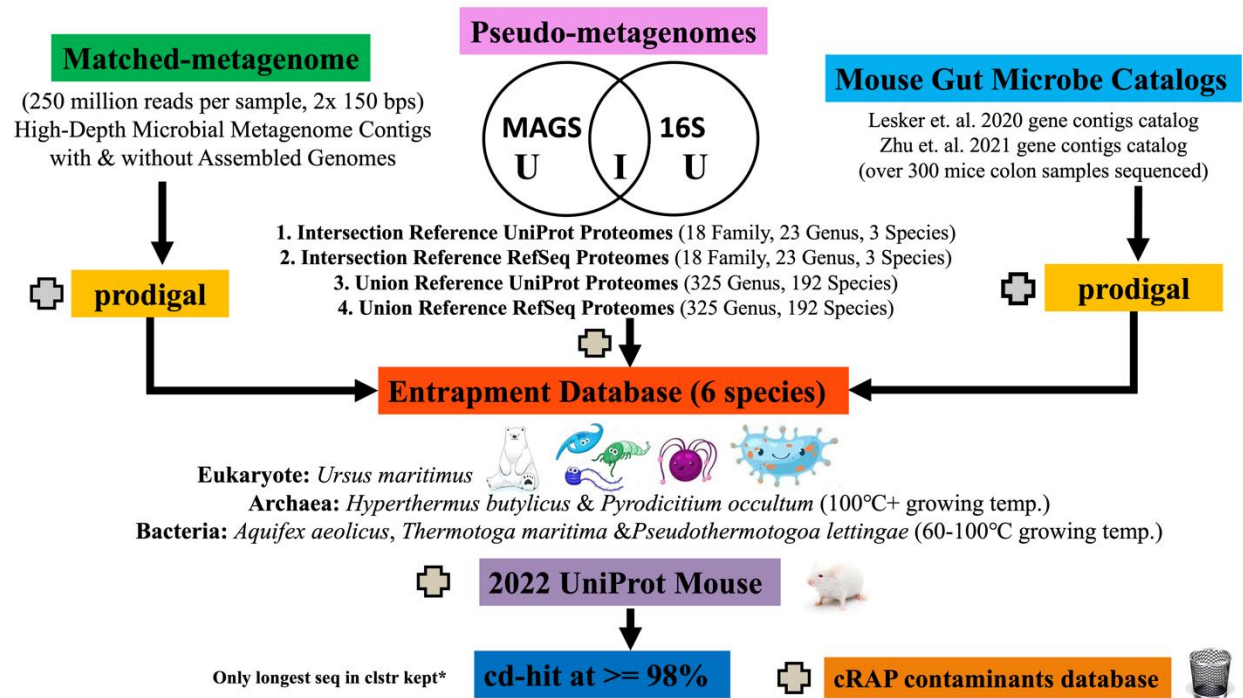


Figure 8 Steps in database construction. Pseudo-metagenomes were downloaded using custom UniProt API command-line Perl scripting & custom NCBI Datasets command-line Python scripting. Mouse catalog gene contigs were downloaded from previous studies and prodigal was used to predict protein sequences. All databases were clustered at ninety-eight percent sequence similarity using cd-hit to match that of the metagenome clustering.

carbamidomethylation of cysteine, N terminal oxidation modifications of methionine, protein reverse decoy, precursor window of 10ppm, minimum m/z of 300, maximum m/z of 1500, high FDR less than 0.01, medium FDR between 0.01 and 0.05, and the precursor ion quantification was set to unique.

3.2.4 Counting Total Inferred Proteins and Peptides

PD implements a Results Exporter node in the Consensus workflow which allows for text files of all the results to be saved at the end of a search. This feature was utilized to quickly obtain the results. Since hundreds of text files needed to be filtered, the command-line was utilized for the calculation of master proteins, proteins with at least one or two unique peptides, and unique peptides. The contaminants and entrapment proteins were filtered out prior to these calculations utilizing grep. The identification scores were uploaded into excel where the margin of error was calculated for each measurement. These scores were then plotted using ggplot2 with the geom_bar and geom_errorbar functions.

3.2.5 Assessing the Sequence Coverage

The contaminants and entrapment were filtered out from protein identification text files using grep. All protein sequences identified were plotted in RStudio using ggplot2 with the geom_jitter function.

3.2.6 Checking Sequence Similarity with BLASTP

The polar bear and mouse proteome were indexed on the command-line utilizing “makeblastdb -in 30Mar2022MouseUniProt.fasta -dbtype prot”. The search was performed with “blastp -db 30Mar2022MouseUniProt.fasta -query Ursus_maritimus.fasta -evaluate 1 -out PolarBearvsMouse_blastp.tab.out -outfmt 6 -num_threads 100”. The third column corresponding to the sequence similarity was filtered for each sequence similarity threshold, and the total proteins from the polar bear proteome were counted. The counts were then plotted in Excel. BLASTP searches were also performed using the entire entrapment database and searching it against each database construct. The results were filtered based on unique entrapment proteins

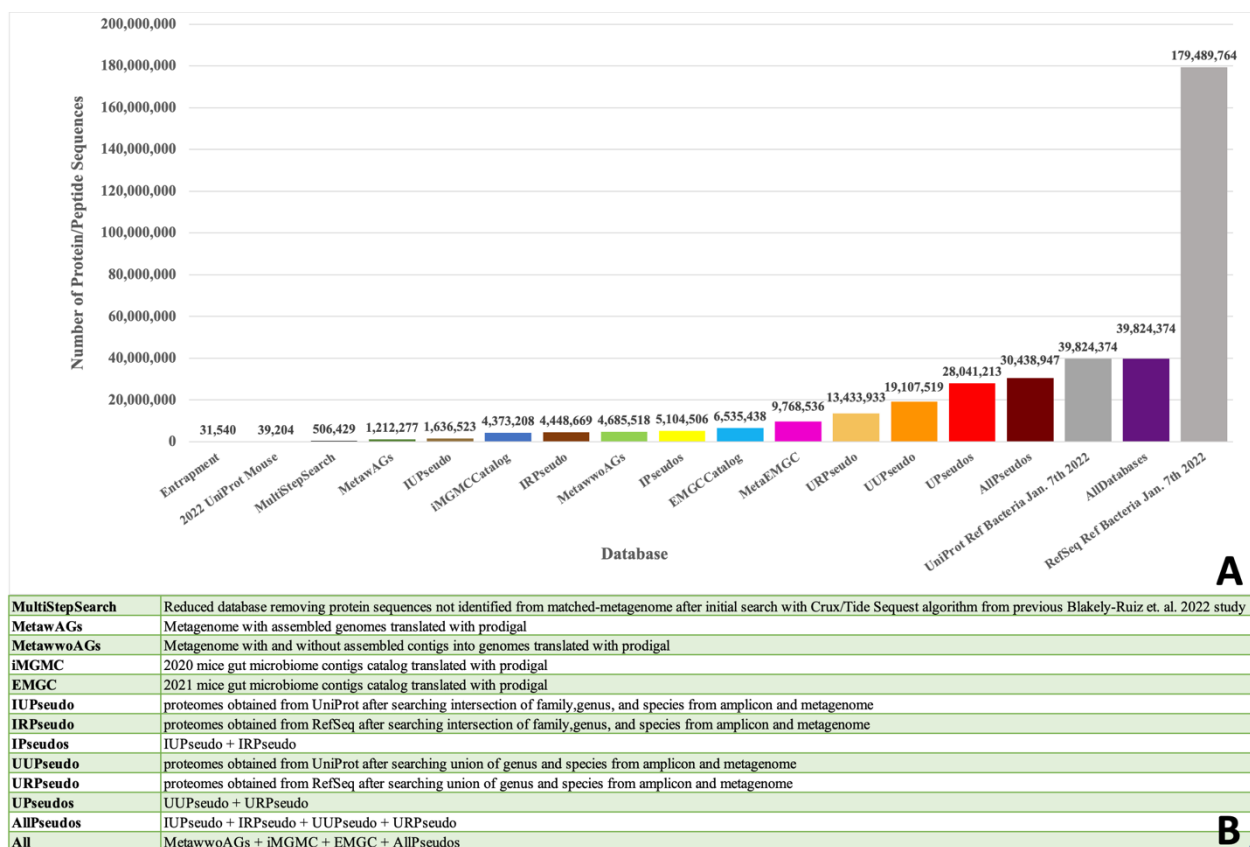


Figure 9 (A) Sizes of databases post clustering with cd-hit at ninety-eight percent sequence similarity and (B) description of databases. The databases sizes were calculated without the addition of the entrapment or the mouse proteomes. For comparison, the sizes of both RefSeq and UniProt reference bacteria (grey) from January of 2022 were included in this figure but were not utilized as a search space for the spectra.

with the highest sequence similarity to each database, and these unique sequence similarity scores were placed into text files for plotting in RStudio with `geom_boxplot`.

3.2.7 Assigning Function and Taxonomy to Identified Proteins with GhostKOALA

Entrapment and contaminants were removed utilizing `grep`, and all unique protein sequences identified by each database type were submitted in fasta file format through the GhostKOALA application. The resulting GhostKOALA KO, genus, and phylum were merged with their corresponding abundance value using `dplyr` function by their accession in RStudio. Raw protein abundance values were summed according to unique phylum, genus, or KO classification provided by GhostKOALA utilizing the `consolidate` feature in Excel. A binary file of these abundance values was created to annotate presence of an abundance reading equal to one and absence of an abundance reading equal to zero. This file was then utilized for UpSetR plotting. For the heatmaps, the raw protein expression values were converted to Log 2 in RStudio, zero was imputed for missing measurements, and the `heatmap.2` function was used with Euclidean distance for plotting.

3.3 Results and Discussion

To investigate the total number of proteins and peptides obtained across databases, several filtering strategies were applied that addressed the accuracy in the measurements obtained across databases. These filtering strategies involved counting the number of quality protein and peptide assignments, plotting the sequence coverage, and counting the amount of entrapment proteins identified. To compare the information gained between studies, identified protein sequences were then submitted through GhostKOALA to assign proteins to corresponding functional categories, genus, and phylum. These analyses are discussed in further detail below.

3.3.1 Investigating the Quality of Identifications Reveals the Matched-Metagenomes Significantly Outperformed All Database Builds

A peptide is a small molecule consisting of three or more amino acid monomers linked together in a linear fashion. The connection of several peptides make up what is referred to as a single functioning unit, a protein. To determine the sequence of proteins found in a sample, they must be broken down into smaller linear peptides which are then detectable with a high degree of specificity through nano-flow LC-MS/MS. Because many peptide components of the same protein in a mixture are typically present in a sample, it is possible to computationally assemble the sequences of proteins when each peptide that makes up a sequence has overlap with another in different positions.

Identifying a peptide from a mixture of organisms presents a great challenge in metaproteomics because in many but not all cases, the same peptide sequence can be produced by many different organisms which is referred to as sequence redundancy. In metaproteomics, this is a challenge that is hard to overcome. A strategy that was introduced to combat this problem is the clustering of sequences at certain similarity thresholds into “protein groups” and retaining the identities of the various organisms which have identical sequences (Li, 2001). This makes it possible to identify multiple peptides which can then be matched back to many organisms that might correspond to the same protein sequence. Despite removal of redundancy within a protein sequence database, the identification process can still result in matching a sequence to multiple peptides that might be from different organisms or proteins. This presents a challenge when trying to assign the peptide to the correct organism or protein. To determine the best-fit assignment for a peptide, several filtering strategies have been adopted in metaproteomics. These filtering strategies have been outlined below.

The Proteome Discover software implements a protein inference strategy referred to as master protein grouping (Orsburn, 2021). A master protein group is only assigned when it identifies a unique protein that has at least one or more unique peptides of which each contain the longest sequence of all observed peptide fragments belonging to that protein sequence (Orsburn, 2021). This represents one of the highest levels of confidence

that can be given to observed proteins in mass spectrometry. A second protein inference strategy that slightly deviates from the master protein grouping considers the fact that shorter peptide sequences may also be real observations, and it sets the threshold for an identified protein at an observation of one or more unique peptides irrespective of the sequence length. It is also typical for studies in metaproteomics to report the total number of unique peptides. The definition for a unique peptide is any peptide that matches only a single unique protein from a single organism (Zhao, 2010). This means that sequences that obtained matches to more than one organism or protein are not considered an identification for unique peptide matching and avoids those sequences where a matching protein or organism could not be determined.

Reporting all peptide identifications can be misleading when a spectral match corresponds to more than one protein or more than one organism. Proteins given the assignment as master or protein with unique peptides provides the highest confidence in a matching peptide spectral sequence given that it has only been predicted to correspond to a single protein from a single organism. The sequences which match to more than one protein or more than one organism are typically removed from the results as there is no current method that can determine which organism or protein a match belongs when multiple are found for one spectral match. Additionally, filtering for only the peptide sequences that were unique allows for a proper measurement of the richness of peptides found in the sample. Moving beyond a peptide richness assessment, the protein level assignments are required for characterizing these systems which is why the above mentioned protein grouping strategies are utilized in metaproteomics.

To assess the performance of the many protein database builds employed in this study, the aforementioned metrics for quantifying accurate peptide and protein identifications were utilized. The z-score was used to calculate an approximate ninety-five percent confidence margin of error across these identifications to show the amount of variability expected with similar studies utilizing the same number of replicates. The margin of error can be calculated as follows:

$$\text{Margin of Error} = \frac{Z \times \sigma}{\sqrt{n}}$$

where the z-score (Z) is multiplied by the standard deviation (σ) and then divided by the square root of the sample size (n). It provides an estimate for the precision within a measurement at a

certain level of confidence. A bar graph displaying the total identifications across these various protein inference groupings with their margin of error is shown in Figure 10 on the next page.

Referring to Figure 10 starting with master protein groups, the matched-metagenome databases (greens) each identified on average a range of 6,000-9,100 master proteins, and the catalogs (blues) identified a range of 5,700-8,700 across treatment groups. The combination of both the matched-metagenome and catalogs (pink) identified 5,800-8,800 master proteins. The pseudo-metagenomes (browns and yellows) all fell short of these values identifying on average a range of 2,700-4,500. Taking into account the margin of error, the master protein identifications for the pseudo-metagenomes were lower than all other database types. The catalogs appeared to achieve similar performance as the matched metagenomes falling within their margin of error. When comparing the two matched-metagenomes, it was found that both performed similarly. The master protein identifications herein provide evidence that the matched-metagenomes had many long unique peptide sequences which matched to a single protein from a single organism.

Continuing to proteins with at least one unique peptide within the same figure, the highest identifications averaged 5,300 for the matched-metagenomes, 4,100 for the catalogs, 3,200 for the combination of catalogs and matched-metagenomes, and 2,100 for the pseudo-metagenomes. The lows averaged 3,200 for the matched-metagenomes, 2,600 for the catalogs, 2,200 for the combination of catalogs and matched-metagenomes, and 1,100 for the pseudo-metagenomes. Again, it was observed that the pseudo-metagenomes and catalogs underperformed for proteins with at least one unique peptide measurement. The protein identifications here provide evidence that the matched-metagenomes had unique peptide sequences of all lengths which matched to a single protein from a single organism.

Applying a more stringent filter of proteins with at least two unique peptide measurements, the highs averaged from 400 in the pseudo-metagenomes to 1,800 in the matched-metagenomes. Here the matched metagenomes achieved the highest number of identifications while the catalogs came next, and the pseudo-metagenomes identified the least. When more than two peptide assignments were taken into consideration, the

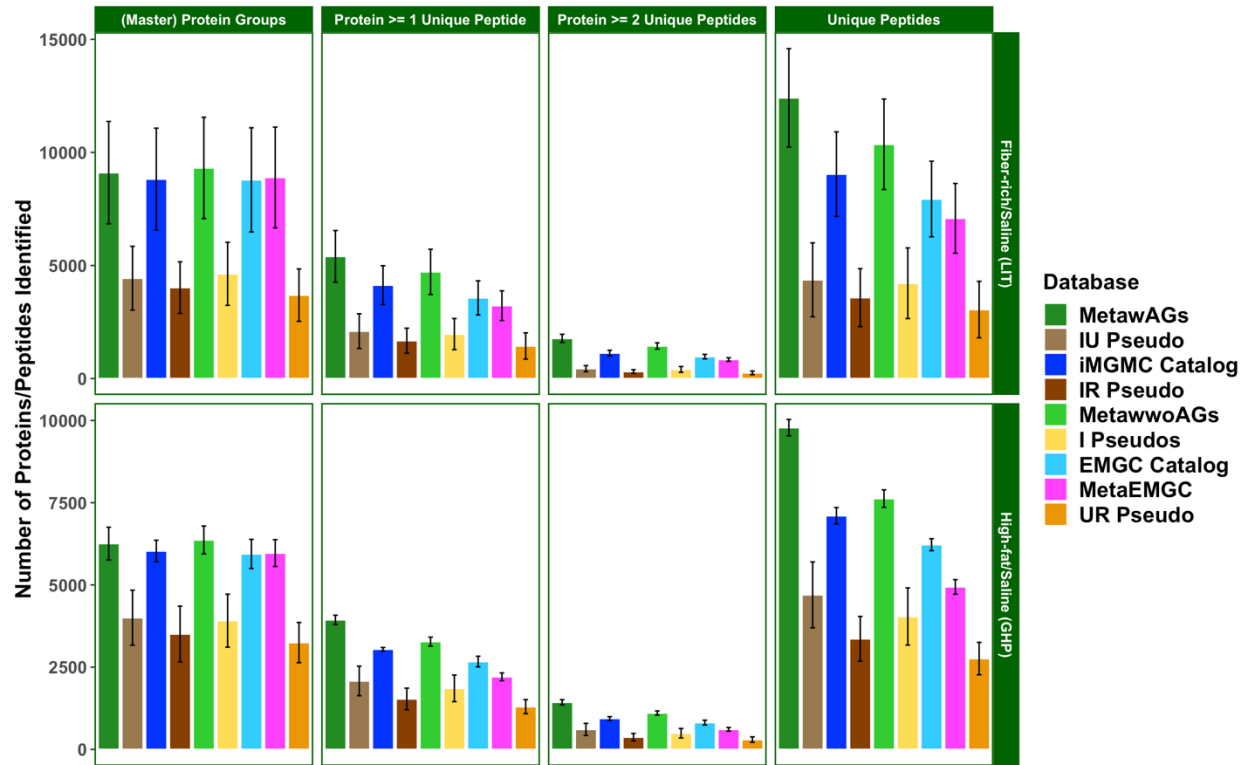


Figure 10 Total proteins and peptides identified at FDR 0.01 across databases and their corresponding margin of error from Proteome Discoverer. Pictured are matched-metagenomes (greens), pseudo-metagenomes (yellows and browns), and catalogs (blues). Databases are listed from left to right in increasing size with the MetawAGs (dark green) being the smallest, and UR Pseudo (orange) being the largest. The contaminants and entrapment proteins were removed utilizing grep on the command-line. A python script was utilized to calculate these counts directly from the PD output folders. Abbreviations: IU= Intersection UniProt Pseudo-metagenome, IR=Intersection RefSeq Pseudo-metagenome, IPseudos=Combined UniProt and RefSeq Intersection Pseudo-metagenomes, UR=Union RefSeq Pseudo-metagenome, MetawAGs=Matched-metagenome with assembled genomes, MetawwoAGs= Matched-metagenome with and without assembled genomes, MetaEMGC=MetawwoAGs+EMGC.

matched-metagenome without the additional contigs without a genome obtained higher numbers of identifications than the larger matched-metagenome. Furthermore, the protein identifications here provide evidence that the matched-metagenomes had many more unique peptide sequences of all lengths which matched to a single protein from a single organism than any other database.

Lastly, the average unique peptides drastically varied across all databases ranging from a low of 2,700 in the pseudo-metagenomes to high of 12,200 in the matched-metagenome with assembled genomes. Here it was also evident that the matched-metagenomes obtained significantly more identifications than all database types in the high-fat/saline treatment group. In the fiber-rich/saline group, the catalogs achieved a comparable number of unique peptides within its margin of error to that of the matched-metagenomes. For both treatment groups, the matched metagenomes obtained two to four times the number of unique peptide identifications than the pseudo-metagenomes. This suggests that the matched-metagenomes had more matched sequences to the spectra that contained a best fit to only one protein sequence from a single organism than any other database. When comparing the two matched-metagenomes, those without the additional contigs that did not have an assembled genome were found to identify more unique peptides.

Furthermore, as evident in Figure 10, the margin of error shown demonstrates that mouse variability among replicates is smaller than the variability observed between the database used. Database choice was found to matter more than the sampling method for replicates. This also shows that the reproducibility among using different databases for this study is weaker than reproducibility among sampling different mice within the same treatment group. A similar observation that was found in the previous study conducted by Blakeley-Ruiz et. al. in 2022 with these datasets was that the fiber-rich mice had a higher protein richness when compared to the high-fat-fed mice ([ISME Communications](#)). Figure 10 confirmed these same findings that a high-fat diet may reduce the diversity found within the gut microbiome.

Overall, it was clear that the matched metagenomes obtained significantly more quality identifications than all other database builds. This demonstrates that a high-depth

metagenome can outperform a reference database that contains over three times the amount of clustered sequences. It shows that only the most relevant sequences will obtain quality protein identifications and that increasing the size of the database does not always lead to more identifications. Although the catalogs did contain sequencing information for over three hundred mice gut microbiota, they still underperformed in terms of the number of quality identifications in comparison to the high depth assembled metagenomes. Blakeley-Ruiz and colleagues in 2022 assembled 460 new mouse fecal bacterial genomes, of which only seven percent were represented in RefSeq and GenBank ([ISME Communications](#)). The gene products from these novel species were included in the matched-metagenomes in this study but were not present in the catalogs, which may have led to them achieving a higher accuracy in measurement. The additional contigs without an assembled genome in the matched-metagenomes appeared to reduce the number of unique peptides obtained.

Filtering only by protein inference method is not the only metric that should be utilized for reporting the number of accurate identifications. Another way to gauge the accuracy in identification is to measure the completeness of matches obtained for each protein identification, which involves reporting the percent of each protein sequence that was matched by the Sequest HT algorithm from the Proteome Discoverer search. Matches to protein sequences in the spectra may only be partially complete which can affect the confidence in the identifications. These sequence coverage values of the identified proteins from each database were plotted in scatterplots using ggplot2 shown in Figure 11 on the next page. They were graphed with their corresponding amino acid chain length to show that shorter proteins were not biasing the corresponding sequence coverage observed.

Referring to the red line in Figure 11, the mean, no difference between average protein sequence coverage was observed across databases. The average protein sequence coverage obtained across all proteins was under ten percent for every database. It was also found that more proteins were identified which were above sixty percent sequence coverage (horizontal dotted green line) for the matched-metagenomes and catalogs than the pseudo-metagenomes. Given that each organism in an environment has its own unique genome, it is not surprising that the

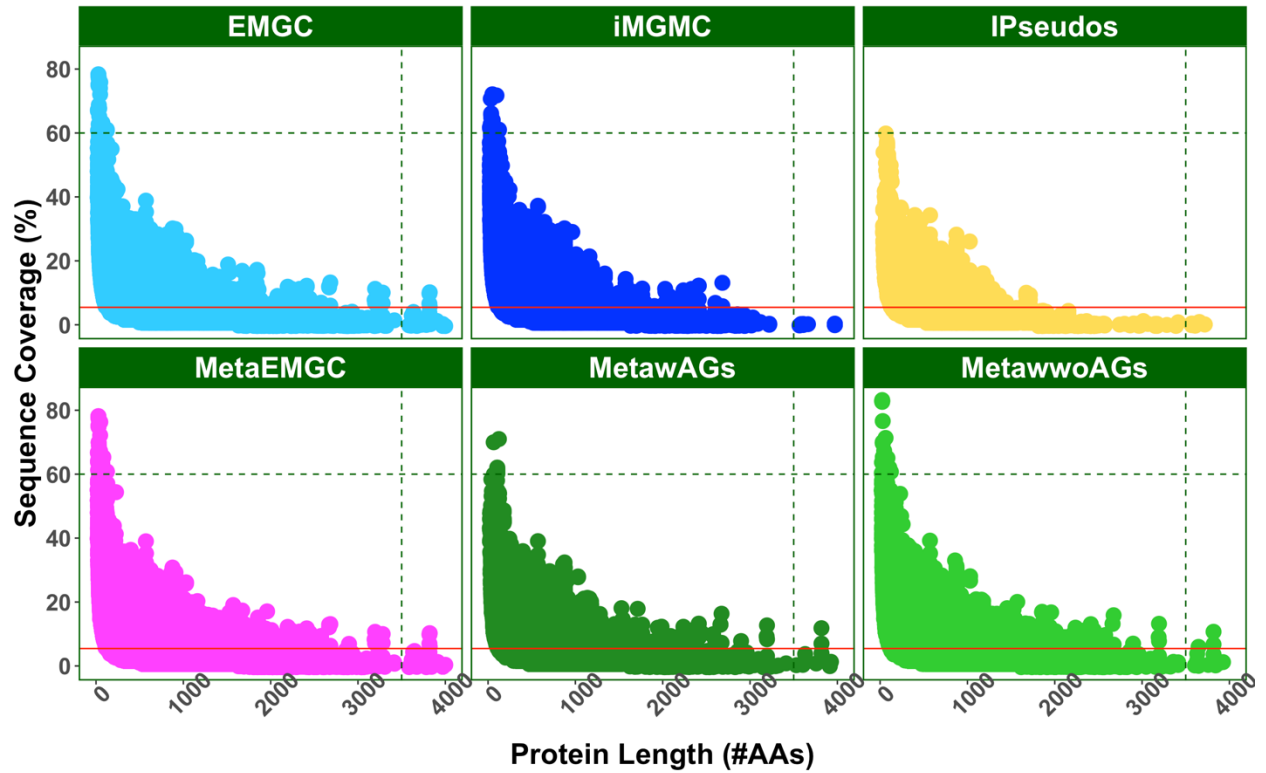


Figure 11 Protein sequence coverage (y axis) and protein length (x axis) obtained across databases for all measured identifications. The scatterplot displays the average in red, and the green dashed lines represent areas of outliers that differed across databases. Plot was generated utilizing ggplot2. Abbreviations: IU= Intersection UniProt Pseudo-metagenome, IR=Intersection RefSeq Pseudo-metagenome, IPseudos=Combined UniProt and RefSeq Intersection Pseudo-metagenomes, UR=Union RefSeq Pseudo-metagenome, MetawAGs=Matched-metagenome with assembled genomes, MetawwoAGs= Matched-metagenome with and without assembled genomes, MetaEMGC=MetawwoAGs+EMGC.

matched-metagenomes obtained a higher sequence coverage than other databases since these metagenomes originated directly from the samples tested in this study.

Additionally, there were longer protein sequences above 3500 amino acids (vertical dotted green line) identified in the matched-metagenome and catalogs than the pseudo-metagenomes. Metagenomes tend to have shorter sequences due to the difficulty encountered during assembly (Ding, 2021). However, for this study, it was found that the length of the matched-metagenome sequences exceeded that of other databases. Long read sequencing performed at 250 million base pairs per mice fecal sample lead to quality translated protein annotations that were well-suited for a metaproteome database.

After assessing the quality of the total identifications across databases, determining the level of accuracy across databases needed to be evaluated. Accuracy in metaproteomics is defined as identifying all the correct proteins and peptides that are in a sample. There are two metrics that can be used to determine if these databases are accurate; false negatives and false positives. False negatives are defined as non-identified proteins, when those proteins were present in the sample. False positives are defined as identifying proteins that were incorrect and not in the sample. A method to determine the number of false negatives has been developed to study synthetic communities of known microorganisms, but none have been developed to assess unknown communities (Starr, 2018).

Studies in proteomics from single organisms have utilized a technique to filter out false positives in samples, which can be applied to metaproteomics studies. This technique involves utilizing what is known as a decoy database where the protein database provided has its sequences reversed (Vaudel, 2012). These reversed sequences can then be utilized to calculate the fitness of a spectral match to the reverse sequences and compared to the fitness of a spectrum matching to the correct, forward sequences to calculate FDR (Vaudel, 2012). This technique has become a standard in metaproteomics and was applied in this study to only identify proteins and peptides with a FDR set to 0.01. A problem with only utilizing the reverse sequences is that it does not account for the sequence redundancy in the database. Efforts to determine this false positive rate in metaproteomics have involved curating what is known as an entrapment database

that contains proteins that should not be found in a sample to gauge how well the decoy database controls the FDR (Vaudel, 2012).

To measure the amount of false positives in this study a nonsensical database, referred to as entrapment, was added to every database in this study. Supplementary Figure 3 shows that this entrapment database obtained matches to the sample's spectra when it was searched by itself showing that sequence redundancy with the entrapment existed also with the true microbial members found in the samples. Entrapment databases are generally used to calculate the FDR. However, this entrapment database could not be utilized for this function when redundancy was observed with the databases and identifications to the sample's spectra were obtained (Vaudel, 2022). Additionally, constructing a database that was larger in size than some of the databases constructed in this study without redundancy would have been nearly impossible to achieve. It is also not known if an entrapment database is the best way to assign the FDR in metaproteomics if those sequences cannot accurately model false positives for spectra (Vaudel, 2012). It has been suggested that entrapment database hits follow closely with the number of reverse decoy database identifications (Vaudel, 2012). However, this has yet to be validated. Instead, here the number of nonsensical identifications obtained from each database are reported not only to show that the false positive identifications can be obtained, but whether erroneous sequences incorporated into a database can have a bias towards identification.

To check the level of entrapment identifications, a bash script was utilized to count the number of entrapment proteins identified by organism, and these results were reported in Table 1 on the next page. From Table 1, it was observed that the entrapment identifications made up less than three percent of all the identifications for each database. Of those three percent, the polar bear accounted for the majority of all the entrapment identifications. Previous groups have suggested that utilizing a eukaryote in an entrapment database is recommended for searching microbial samples because they contain very little sequence redundancy with microbial species (Vaudel, 2012). However, in this study, it was found that the polar bear sequences within the

Table 1 Total entrapment protein identifications across databases at FDR of 0.01 broken down by organism. Databases are listed from left to right in increasing size with MetawAGs being the smallest and MetaEMGC being the largest. These identifications represent the raw proteins that were output from PD prior to filtering. Abbreviations: IU= Intersection UniProt Pseudo-metagenome, IR=Intersection RefSeq Pseudo-metagenome, IPseudos=Combined UniProt and RefSeq Intersection Pseudo-metagenomes, UR=Union RefSeq Pseudo-metagenome, MetawAGs=Matched-metagenome with assembled genomes, MetawwoAGs= Matched-metagenome with and without assembled genomes, MetaEMGC=MetawwoAGs+EMGC.

Type of Protein Identification	MetawAGs	iMGMC	MetawwoAGs	IPseudos	EMGC	MetaEMGC
<i>Mus musculus</i>	856	824	846	714	760	811
<i>Aquifex aeolicus</i>	1	2	1	2	2	1
<i>Pseudothermotoga lettingae</i>	2	1	0	2	1	1
<i>Pyrodictium occultum</i>	1	1	1	1	1	1
<i>Thermotoga maritima</i>	1	0	0	1	0	1
<i>Hyperthermus butylicus</i>	0	0	0	0	0	0
<i>Ursus maritimus</i>	345	309	339	240	300	305
Total Entrapment	350	313	341	246	304	309
Total Contaminants	20	18	20	17	18	18
Total without Contaminants or Entrapment	13370	22015	16325	18037	27106	28963
Total without Mouse, Contaminants, or Entrapment	12514	21191	15479	17323	26346	28152
Total Before Filtering	13740	22346	16686	18300	27428	29290

entrapment database had a great deal of sequence redundancy with the host (mouse), which may have predisposed its identification in this study. A BLASTP search of the polar bear against the mouse proteome was performed to uncover the amount of protein sequence redundancy across the mouse and polar bear proteomes. The results from this search are plotted in Supplementary Figure 4 in the Appendix. It was found that ninety-six percent of the entire polar bear proteome had at least fifty percent sequence similarity across all its proteins. It was also observed that thirty-nine percent of all the proteins found in the polar bear proteome were at least ninety percent similar in sequence. This meant that the probability of obtaining a match to the polar bear proteome was quite high. This explains why the entrapment database may have obtained several hundred protein identifications. Clustering at ninety-eight percent may not have been sufficient to account for redundancy at lower similarity thresholds between the entrapment and these databases leading to a high number of polar bear protein identifications. It should be noted that the polar bear was not removed from this database for the same reason the microbial proteins with redundancy were not removed from the entrapment. The goal of the study was to model real situations in metaproteomics where organisms placed in reference databases might be incorrect and lead to false identifications due to varying levels of sequence similarity, database size and complexity. The sequence similarity of the entrapment compared to each of the databases without mouse present was searched in BLASTP, but it did not show that sequence redundancy influenced the counts observed in Table 1. A correlation was also not observed between the database size as a driving factor for decreasing the number of entrapment identifications as described by Vaudel and colleagues in 2012. A common approach applied in genomics models the complexity of traits in order to predict the FDR (Brzyski, 2017). It is possible the number of organisms represented in each database may have influenced the polar bear entrapment match rate. The matched-metagenomes had a total of 590 assembled genomes, while the catalogs had greater than 830 assembled genomes (Zhu, 2021), and the intersection pseudo-metagenomes had 1,657 proteomes. The lower complexity of the matched-metagenomes may have increased the probability of matching to polar bear from the entrapment.

Unlike the polar bear entrapment identifications which had high redundancy with the mouse proteome, a different trend was observed for the microbes in this database. The

bacteria and archaea identifications from the entrapment database were all similar identifying one to two proteins from each. If the matched-metagenomes provided thorough coverage of the metaproteome for these samples, the entrapment database would have a smaller probability of being identified due to Sequest HT only selecting the best match from the database at an FDR of 0.01. A BLASTP was performed between the entrapment microbial proteins and each database to test this hypothesis. The mean sequence similarity obtained for IPseudos was 44.7 percent, while the means of the MetawAGs and MetawwoAGs were at 45.47 and 46.61 percent respectively. This suggests that the matched-metagenomes had slightly better coverage of the true proteome from the sample which may have lowered the probability that a microbial entrapment protein was identified from each database.

Little difference was observed between the entrapment identification rates across each of the databases representing only a small percentage of the total. Although the false discovery rate cannot be determined by utilizing an entrapment database with redundant sequences that obtain hits to experimental spectra, the low entrapment identifications obtained across the databases suggest that the FDR may be controlled utilizing a decoy database. However, because this entrapment database represented only three percent or less of the entire clustered sequences in these databases, larger entrapment databases might yield more informative results. Incorporating an algorithm that considers the number of organisms and sequence redundancy and not just the number of proteins clustered in a database might provide better predictions for false discovery rates.

3.3.2 Inferring Function and Taxonomy Across Databases Reveals Significant Losses in Annotation and Drastic Changes in Abundance

Although obtaining the highest number of quality protein and peptide identifications is desirable for reference databases used in metaproteomics, it is probably more important to understand whether the functional information gained from a database is relevant and whether they are identifying the right taxa. Comparing the function and taxonomy across different databases would seem like an easy task, however, databases infer protein and taxonomy in different ways. To make this an unbiased comparison, universal protein identifiers needed to be

assigned to the protein sequences since lineage assignment and protein naming was applied differently across the databases.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is one of the largest online repositories utilized for obtaining and visualizing functional information gained from sequencing proteins across all organisms (Kanehisa, 2016). This repository contains a web-based tool called GhostKOALA which allows protein sequences to be submitted and searched against the entire repository (Kanehisa, 2015). The annotations obtained from this search are referred to as KEGG Orthologs (KOs). Each KO number corresponds to a protein that matched a submitted sequence which may be shared across many species but has a similar function (Kanehisa, 2015). Associating protein sequences by function rather than species can allow for the visualization of proteins involved in metabolic pathways which have been previously characterized and experimentally confirmed to occur in nature.

Protein sequences from this study that were identified from each of the databases (excluding contaminants and entrapment) were submitted through GhostKOALA for assignment. The abundance of all proteins was then summed according to its KO or taxonomic level. Since many of the sequences in this study did not have a species level assignment, comparisons were made between the genus and phylum.

For visualization of these differences, UpSetR and Euclidean distance heatmaps were generated. UpSetR plots are great visualization tools for comparing the presence or absence of features from large datasets. They allow more comparisons than traditional Venn diagrams and provide an easier way to visualize which groups show correlation (Conway, 2017). Since UpSetR plots are binary in nature, a heatmap was also utilized to assess how protein abundance varied across the assigned GhostKOALA function and taxonomy that was obtained and allowed visualization of missing values across databases. Combining these two data visualization methods allowed for cross-validation of the missingness and completeness present across databases. Additionally, although the UpSetR plot could not be used to identify abundance of a protein, as is visible in a Heatmap, the UpSetR plot allowed for the number of unique functional and taxonomic

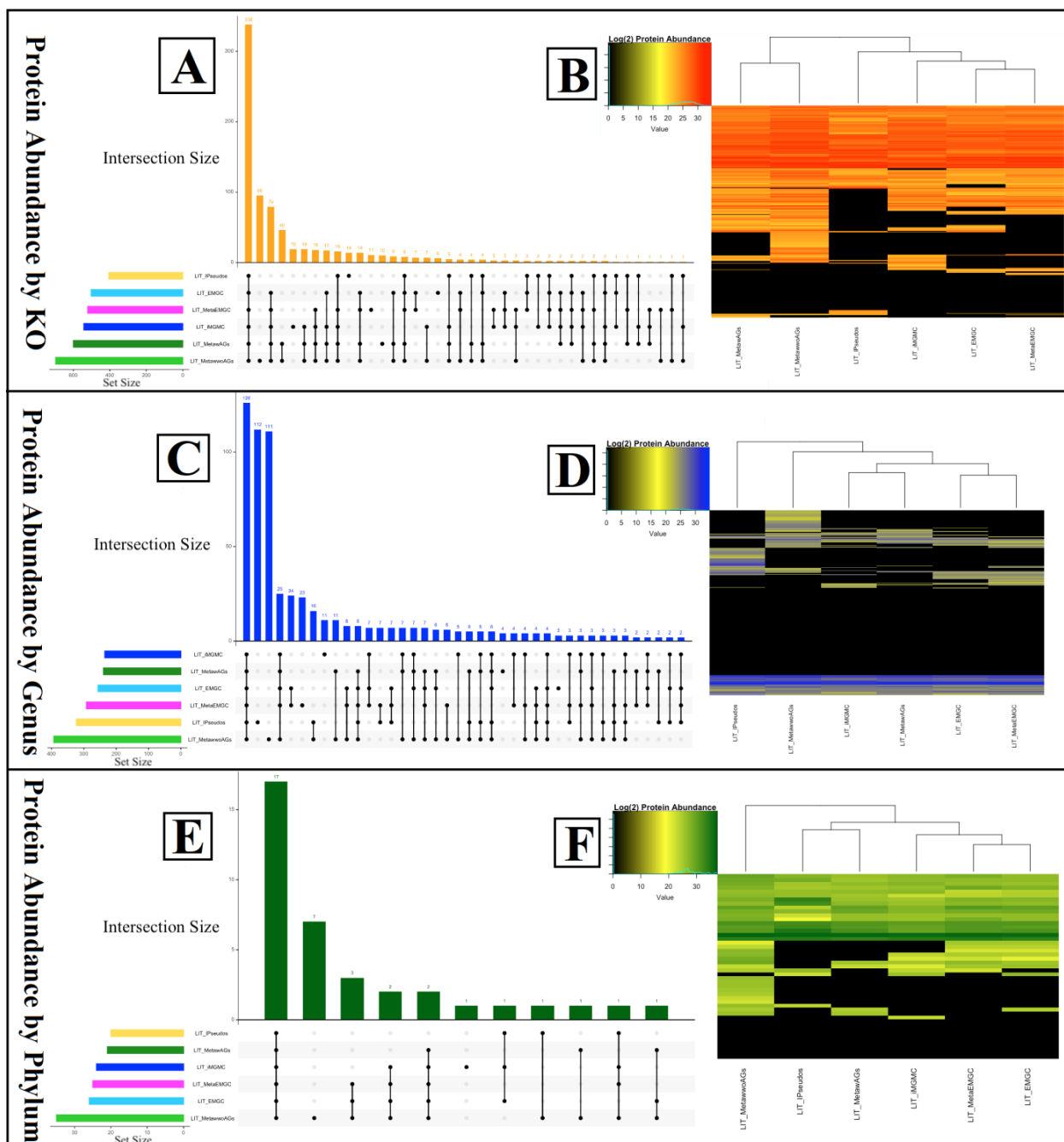
information contributed by each database to be calculated. Since there were several hundred identifications output from GhostKOALA, the heatmap itself had many rows that prevented a viewer from counting this output. Features that the heatmap provided included visualization of the completeness of the protein identifications which were matched to KEGG KO and taxonomy. The colored cells on the heatmaps allowed for visualization of the changes in Log 2 protein abundance per each identification which was not possible with the UpSetR plot itself. Figure 12 on the next page depicts these comparisons for the three categories of GhostKOALA assignment for the fiber-rich/saline (LIT) mice. For simplicity, only the fiber-rich/saline mice are depicted. A similar trend was observed for the high-fat/saline (GHP) treatment group which is reported in Supplementary Figure 5.

From the first KO UpSetR plot in Figure 12 labeled in orange, it was observed the matched-metagenomes contributed the most unique KO functional protein categories. This same observation was confirmed when referring to the heatmap beside this plot. These phenomena were investigated, and it was found that the matched-metagenome with and without assembled genomes contributed 691 KOs while the matched-metagenome without the additional contigs without a genome led to only 600 KO annotations. It has previously been suggested that contigs without an assembled genome might provide additional functional information in a study and should not be removed from the database (Blakeley-Ruiz, Computational and Structural Biotechnology, 2022).

The IMGMC catalog yielded 544 KO annotations, while the EMGC catalog yielded even less at 503. This loss in KO annotation may be due the lack of these sequences residing in the KEGG database as many from the EMGC catalog were obtained through RefSeq. The combination of the EMGC catalog and the matched-metagenome with and without assembled genomes fell in between these catalogs and retrieved 522. The fact that fewer protein identifications were made compared to matched-metagenomes may help to explain this. Lastly, the intersection pseudo-metagenomes from RefSeq and UniProt combined returned only 407 KO terms from the identified protein sequences. Similarly, this phenomenon might be explained by the lower number of identifications obtained. These same trends were observed for the high-fat/saline mice shown in Supplementary Figure 5. All databases from the fiber-rich/saline

Figure 12 Summed GhostKoala KO (A/B), genus (C/D), and phylum (E/F) protein abundance values graphed in UpSetR plot and Euclidean distance heatmaps for high-fiber/saline (LIT) condition. All identified protein sequences across databases were submitted through GhostKOALA to annotate the functional and taxonomic information gained from each database. **A/C/E:** The top bar chart of the UpSetR plot (Intersection Size) displays the total number of proteins belonging to either KO, phylum, or genus which were found in the sets of databases connected by dots and labeled in the plot directly below. The highest numbers from the top bar chart correspond to the sets of databases which identified the most proteins associated with each KO, phylum, or genus from GhostKOALA. The bar chart to the left in each UpSetR plot (Set Size) measures how many unique identifications were contributed by each database. The largest bar on the left shows which database provided the highest amount of functional (KO), and taxonomic (phylum, genus) information. **B/D/F:** Euclidian distance heatmaps depict all the summed protein abundance values for each KO, phylum, and genus identified across databases. Black represents absence of a KO, phylum, or genus assignment. The range in colors depicts how the Log 2 protein abundance changed according to database utilized.

Abbreviations: IU= Intersection UniProt Pseudo-metagenome, IR=Intersection RefSeq Pseudo-metagenome, IPseudos=Combined UniProt and RefSeq Intersection Pseudo-metagenomes, UR=Union RefSeq Pseudo-metagenome, MetawAGs=Matched-metagenome with assembled genomes, MetawwoAGs= Matched-metagenome with and without assembled genomes, MetaEMGC=MetawwoAGs+EMGC.



treatment group shared KO annotations such as K00197, K10907, and K07192 involved in carbon metabolism, amino acid biosynthesis, and insulin signaling. KOs unique to the pseudo-metagenomes included K01358, K01226, K03366 which was a protease, starch and sucrose metabolism enzyme, and butanoate metabolism enzyme. The catalogs contributed unique KOs as well such as K07739, K09800, and K00012 involved in tRNA biosynthesis, membrane transport, and glucuronate biosynthesis. The matched-metagenome contained unique KOs such as K02073, K03217, and K15870 involved in D-methionine transport, quorum sensing, and bile acid synthesis. This suggests that central metabolism seemed to be the core functional information identified using any of these databases but that other less-essential pathways were missing identifications when a different database was utilized.

The functional information gained from each database suggests that overall, the matched-metagenome with and without assembled genomes led to the highest number of KO annotations from the KEGG database. This means that not only were more unique proteins and peptides identified utilizing the matched-metagenome databases, but more functional information was gained that could lead to deeper understanding of the biochemical processes occurring in these organisms. From the heatmap, it was also observed that a great deal of functional information was lost when different databases were utilized. Additionally, abundance is not conserved across databases suggesting that phenotypic information gained from a study is largely database dependent (Figure 12).

Next, in order to assess the taxonomic information gained from each database, the identified taxonomy from these databases needed to be visualized at different levels. Since species-level identifications drastically differed across databases, the genus and phylum level assignments from KEGG were chosen for comparison. From the UpSetR plot generated in Figure 12 in blue corresponding to the genus annotation of the fiber-rich (LIT) mice, it was found that the metagenome had more genus level assignments in KEGG than the pseudo-metagenomes. The matched-metagenome achieved 394 genera annotations in KEGG, while the pseudo-metagenomes only obtained 323. The matched-metagenomes identified unique genera such as *Alicyclobacillus*, *Labilibaculum*, *Desulfurococcus*, and *Thermoactinomyces*. The catalogs also contributed unique genera

such as *Viridibacillus*, *Carnobacterium*, *Planomicrobium*, *Saccharibacillus*, and *Rhodospirillum*. Many common mice gut colonizers such as *Clostridium*, *Eubacterium*, *Candidatus*, *Brevefilum*, *Vibrio*, *Caloramator*, and *Butyricimonas* genera were identified in all the databases.

Interestingly, the complexity and size of the database did not seem to correlate with the annotation obtained. For the high-fat/saline mice, the only difference observed was that the matched-metagenome with assembled genomes had less genus than iMGMC catalog and yielded the least number of annotations. Although the pseudo-metagenomes and catalogs may have had more organisms and genus represented in the database, they retrieved a lower number of identifications to unique genus. Additionally, looking at the heatmap generated, it is important to notice that less than fifty percent of the identifications obtained genus level assignments.

A similar trend for the phylum was observed. For the phylum level assignments in green from the fiber-rich (LIT) mice in Figure 12, it was found that the matched-metagenomes had a more diverse group of phyla, and that abundance varied according to the database selected. The number of phyla annotated for the matched-metagenome with and without assembled genomes was 36 while the pseudo-metagenomes obtained only 21. Some of the most abundant phyla across all databases were Firmicutes and Bacteroidetes with Chlamydiae and Thermodesulfobacteria found only in the matched-metagenomes. A bacterium only found in the catalogs was Atribacterota. These results show that many of the identifications made were not present in KEGG and may require a different annotation tool.

When the phylum and genus richness was compared between the original database annotations, it was found that the pseudo-metagenomes had complete annotation, while the catalogs had less, and the matched-metagenomes obtained very little annotation. Although the pseudo-metagenomes all had taxonomy annotated, GhostKOALA did not retrieve that information. These results show that multiple taxonomic annotation tools may be warranted in mouse gut microbiome studies since KEGG did not contain many of the proteomes annotated in RefSeq. Some studies have suggested using DIAMOND, MEGAN, and UniPept for taxonomic annotation in metaproteomic studies which might improve this coverage (Bağcı, 2021; Mesuere, 2015; Tanca, 2013).

3.4 Conclusions from Study and Closing Remarks

In summary, the results presented herein show that matched-metagenomes with assembled genomes obtained significantly more master proteins, proteins with at least one unique or two unique peptides, and unique peptides than catalogs and reference databases. The total identifications for the matched metagenome with assembled genomes led to more quality matches during the searching process. However, when the functional information was assessed, the matched-metagenomes with the additional unassembled contigs was shown to have the most KO sequences annotated in KEGG. These results suggest that keeping the unassembled contigs from metagenomic assembly may be warranted since more functional information may be gained from this database type.

The entrapment database identifications that were obtained provided evidence that false positive identifications were occurring at a low rate suggesting that the decoy database may perform well for controlling non-sensical protein identifications. If the entrapment is to be used as a method to measure the level of false positives, this suggests that the accuracy in measurement is comparable across databases builds. Larger entrapment databases equivalent to target databases may need to be created to address how the number of organisms clustered and the sequence redundancy might be affecting the false discovery rate in metaproteomics.

Through observing the KO, genus, and phylum assigned to identified protein sequences annotated through GhostKOALA, it was found that the matched-metagenomes contributed the most functional as well as taxonomic information. Additionally, the function as well as taxonomic abundance varied greatly across databases providing further evidence that database curation directly affects the level of information gained from metaproteomics studies. Searching identified protein sequences in KEGG provided evidence that a great deal of mice gut microbiome species lack annotation in this database.

Previous groups have suggested that the multi-step-search strategy might also increase the number of quality peptide and protein identifications (Jagtap, 2013).

Blakeley-Ruiz and colleagues in 2022 constructed a reduced database after searching the matched-metagenome and removed proteins that did not obtain an identification through the Crux/Tide program with Sequest ([ISME Communications](#)). This multi-step-search database was searched against the sample's spectra with the added entrapment database, and it was found to identify more master proteins, proteins with at least one or two unique peptides, and unique peptides for both treatment groups (Supplementary Figure 5). However, the entrapment identification rate was found to be higher for the multi-step-search than that of the matched-metagenomes (Supplementary Table 1). Vaudel and colleagues in 2012 found that multi-step-search strategies drastically inflate the FDR. Future work should be done to determine exactly how much the FDR changes and what effect it has on the annotation of function and taxonomy when this technique is utilized in metaproteomics.

In conclusion, although it is tempting to combine available reference databases in the immediate environment and search the largest space possible for mice gut microbiomes, this study found that utilizing translated, sequenced, and assembled metagenomes with a deep measurement exceeded that of larger reference databases and provided more quality annotations with a similar level of false discovery. The low sequence coverage observed in this experiment should further stress the importance of integrating omics since the FDR in metaproteomics is still largely unknown. Because phenotypic information can drastically differ according to the database utilized, metagenomes should be used to validate the proteins that are identified in metaproteomics for mice microbiomes. As sequencing technology continues to advance, deeper metagenome measurements can be taken to construct databases in metaproteomics that are sample-specific leading to exceptional protein and peptide mapping for diagnosing disease. The 2020, iMGMC, and 2021, EMGC, mice gut microbiome catalogs that were translated with prodigal will be made available to the public to assist in future mice gut studies.

In the future, more RAM for the largest pseudo-metagenomes would need to be obtained to search them in the cloud using the same technique as described for the other databases in this study. Due to insufficient RAM, the large databases containing the union pseudo-metagenome from UniProt were split and searched in PD. It was later recognized that this incorporated not only a duplication bias but resulted in improper protein inference that was not valid for a

comparison with the databases that were searched only a single time in PD and thus were not used in the analyses.

Metaproteomics is a relatively young field that has been evolving every year to help uncover the microscopic life that inhabits this world and to characterize their behaviors as they relate to the health of the host and the microbial community. One of the most critical elements which determines the success of these studies is to curate databases that can give the scientific community the tools to obtain complete coverage of the proteins expressed in these environments with a high level of accuracy (Starr, 2018). Previous groups have compared the use of metagenomes without assembled genomes at a depth of 18 million base pairs for human fecal metaproteome identification and found that combining reference databases increased this coverage (Xaio, 2018; Tanca, 2016; Blakeley-Ruiz, Computational and Structural Biotechnology, 2022). The accuracy in that measurement for human gut microbiome studies, however, was not addressed (Blakeley-Ruiz, Computational and Structural Biotechnology, 2022). These previous findings which suggested the matched-metagenomes underperformed may have been influenced by lack of assembly and poor metagenome sequencing (Blakeley-Ruiz, Computational and Structural Biotechnology, 2022). The study discussed herein found that metagenomes can be a powerfully accurate tool to utilize for characterizing microbiomes if and only if the measurements are deep enough and the genomes are assembled. Metagenome sequencing has now become affordable ranging from \$100 to \$2,000 per sample with a labor cost ranging from a month to a year (Blakeley-Ruiz, Computational and Structural Biotechnology, 2022). These metagenome sequencing resources are no longer inaccessible and should be applied to study the microbiome in metaproteomics.

REFERENCES

- Alharthi, A., Alhazmi, S., Alburae, N., & Bahieldin, A. (2022). The Human Gut Microbiome as a Potential Factor in Autism Spectrum Disorder. *International journal of molecular sciences*, 23(3), 1363. <https://doi.org/10.3390/ijms23031363>
- Allaband, C., McDonald, D., Vázquez-Baeza, Y., Minich, J. J., Tripathi, A., Brenner, D. A., Loomba, R., Smarr, L., Sandborn, W. J., Schnabl, B., Dorrestein, P., Zarrinpar, A., & Knight, R. (2019). Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*, 17(2), 218–230. <https://doi.org/10.1016/j.cgh.2018.09.017>
- Allard, G. (2018). GuyAllard/markov_clustering. Github. Retrieved from https://github.com/GuyAllard/markov_clustering
- Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology*, 37(4), 420–423. <https://doi.org/10.1038/s41587-019-0036-z>
- Anaconda Software Distribution. (2020). Anaconda Documentation. Anaconda Inc. Retrieved from <https://docs.anaconda.com/>
- Aoun, A., Darwish, F., & Hamod, N. (2020). The Influence of the Gut Microbiome on Obesity in Adults and the Role of Probiotics, Prebiotics, and Synbiotics for Weight Loss. *Preventive nutrition and food science*, 25(2), 113–123. <https://doi.org/10.3746/pnf.2020.25.2.113>
- Bağcı, C., Patz, S., & Huson, D. H. (2021). DIAMOND+MEGAN: Fast and Easy Taxonomic and Functional Analysis of Short and Long Microbiome Sequences. *Current protocols*, 1(3), e59. <https://doi.org/10.1002/cpz1.59>
- Bibiloni R. (2012). Rodent models to study the relationships between mammals and their bacterial inhabitants. *Gut microbes*, 3(6), 536–543. <https://doi.org/10.4161/gmic.21905>
- Blakeley-Ruiz JA, McClintock CS, Shrestha HK, Poudel S, Yang ZK, Giannone R, Choo JJ, Podar M, Baghdoyan HA, Lydic R, Hettich RL. (2022). Morphine and High-Fat Diet

- Differentially Alter the Gut Microbiome's Composition and Metabolic Function in Lean and Obese Mice. ISME Communications. Manuscript in press for publication.
- Blakeley-Ruiz, J. A., & Kleiner, M. (2022). Considerations for constructing a protein sequence database for metaproteomics. *Computational and structural biotechnology journal*, 20, 937–952. <https://doi.org/10.1016/j.csbj.2022.01.018>
- Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., & Sabatti, C. (2017). Controlling the Rate of GWAS False Discoveries. *Genetics*, 205(1), 61–75. <https://doi.org/10.1534/genetics.116.193987>
- Burki F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harbor perspectives in biology*, 6(5), a016147. <https://doi.org/10.1101/cshperspect.a016147>
- Childs, D., Hindle, B., & Warren P. (2021). APS 240: Data Analysis and Statistics with R. Github. Retrieved from <https://dzchilds.github.io/stats-for-bio/choosing-models-and-tests.html>
- Church, D. D., Hirsch, K. R., Park, S., Kim, I. Y., Gwin, J. A., Pasiakos, S. M., Wolfe, R. R., & Ferrando, A. A. (2020). Essential Amino Acids and Protein Synthesis: Insights into Maximizing the Muscle and Whole-Body Response to Feeding. *Nutrients*, 12(12), 3717. <https://doi.org/10.3390/nu12123717>
- Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics (Oxford, England)*, 33(18), 2938–2940. <https://doi.org/10.1093/bioinformatics/btx364>
- Deng, Z., & Delwart, E. (2021). ContigExtender: a new approach to improving de novo sequence assembly for viral metagenomics data. *BMC bioinformatics*, 22(1), 119. <https://doi.org/10.1186/s12859-021-04038-2>
- Dill-McFarland, K. A., Tang, Z. Z., Kemis, J. H., Kerby, R. L., Chen, G., Palloni, A., Sorenson, T., Rey, F. E., & Herd, P. (2019). Close social relationships correlate with human gut microbiota composition. *Scientific reports*, 9(1), 703. <https://doi.org/10.1038/s41598-018-37298-9>

- Ejigu, G. F., & Jung, J. (2020). Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. *Biology*, 9(9), 295.
<https://doi.org/10.3390/biology9090295>
- Ezzamouri, B., Shoaie, S., & Ledesma-Amaro, R. (2021). Synergies of Systems Biology and Synthetic Biology in Human Microbiome Studies. *Frontiers in microbiology*, 12, 681982.
<https://doi.org/10.3389/fmicb.2021.681982>
- Feng, X. D., Li, L. W., Zhang, J. H., Zhu, Y. P., Chang, C., Shu, K. X., & Ma, J. (2017). Using the entrapment sequence method as a standard to evaluate key steps of proteomics data analysis process. *BMC genomics*, 18(Suppl 2), 143. <https://doi.org/10.1186/s12864-017-3491-2>
- Fox, G. E., Magrum, L. J., Balch, W. E., Wolfe, R. S., & Woese, C. R. (1977). Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences of the United States of America*, 74(10), 4537–4541.
<https://doi.org/10.1073/pnas.74.10.4537>
- Franklin, S., & Vondriska, T. M. (2011). Genomes, proteomes, and the central dogma. *Circulation. Cardiovascular genetics*, 4(5), 576.
<https://doi.org/10.1161/CIRCGENETICS.110.957795>
- Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(23), 3150–3152.
<https://doi.org/10.1093/bioinformatics/bts565>
- Giai Gianetto, Q., Couté, Y., Bruley, C., & Burger, T. (2016). Uses and misuses of the fudge factor in quantitative discovery proteomics. *Proteomics*, 16(14), 1955–1960.
<https://doi.org/10.1002/pmic.201600132>
- Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature medicine*, 24(4), 392–400.
<https://doi.org/10.1038/nm.4517>
- Simón, Pellicer, A., & Reijo Pera, R. (2013). *Stem Cells in Reproductive Medicine : Basic Science and Therapeutic Potential*. (3rd ed.). Cambridge University Press.

- Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research*, 36(10), 3420–3435. <https://doi.org/10.1093/nar/gkn176>
- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1–8. <https://doi.org/10.1016/j.ygeno.2015.11.003>
- Hediyeh-zadeh, S., Webb, A., Davis, M. (2021). MSImpute: Imputation of label-free mass spectrometry peptides by low-rank approximation. *BioRxiv*. doi: <https://doi.org/10.1101/2020.08.12.248963/>
- Holmes B. (2022). NCBI Datasets is an experimental resource for finding and building datasets. Available online at: <https://github.com/ncbi/datasets>
- Huang, H., McGarvey, P. B., Suzek, B. E., Mazumder, R., Zhang, J., Chen, Y., & Wu, C. H. (2011). A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics (Oxford, England)*, 27(8), 1190–1191. <https://doi.org/10.1093/bioinformatics/btr101>
- Hugenholtz, F., & de Vos, W. M. (2018). Mouse models for human intestinal microbiota research: a critical evaluation. *Cellular and molecular life sciences : CMLS*, 75(1), 149–160. <https://doi.org/10.1007/s00018-017-2693-8>
- 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>
- Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11, 119. <https://doi.org/10.1186/1471-2105-11-119>
- Integrative HMP (iHMP) Research Network Consortium (2014). The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe*, 16(3), 276–289. <https://doi.org/10.1016/j.chom.2014.08.014>
- Issa Isaac, N., Philippe, D., Nicholas, A., Raoult, D., & Eric, C. (2019). Metaproteomics of the human gut microbiota: Challenges and contributions to other OMICS. *Clinical mass*

- spectrometry (Del Mar, Calif.), 14 Pt A, 18–30.
<https://doi.org/10.1016/j.clinms.2019.06.001>
- Jagtap, P., Goslinga, J., Kooren, J. A., McGowan, T., Wroblewski, M. S., Seymour, S. L., & Griffin, T. J. (2013). A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*, 13(8), 1352–1357. <https://doi.org/10.1002/pmic.201200352>
- Jékely G. (2007). Origin of phagotrophic eukaryotes as social cheaters in microbial biofilms. *Biology direct*, 2, 3. <https://doi.org/10.1186/1745-6150-2-3>
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1), D457–D462. <https://doi.org/10.1093/nar/gkv1070>
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1), D353–D361. <https://doi.org/10.1093/nar/gkw1092>
- Kawashima, Y., Nagai, H., Konno, R., Ishikawa, M., Nakajima, D., Sato, H., Nakamura, R., Furuyashiki, T., & Ohara, O. (2022). Single-Shot 10K Proteome Approach: Over 10,000 Protein Identifications by Data-Independent Acquisition-Based Single-Shot Proteomics with Ion Mobility Spectrometry. *Journal of proteome research*, 21(6), 1418–1427. <https://doi.org/10.1021/acs.jproteome.2c00023>
- Kleiner, M., Wentrup, C., Lott, C., Teeling, H., Wetzel, S., Young, J., Chang, Y. J., Shah, M., VerBerkmoes, N. C., Zarzycki, J., Fuchs, G., Markert, S., Hempel, K., Voigt, B., Becher, D., Liebeke, M., Lalk, M., Albrecht, D., Hecker, M., Schweder, T., ... Dubilier, N. (2012). Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19), E1173–E1182. <https://doi.org/10.1073/pnas.1121198109>
- Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 2015(11), 951–969. <https://doi.org/10.1101/pdb.top084970>

- Lesker, T. R., Durairaj, A. C., Gálvez, E., Lagkouvardos, I., Baines, J. F., Clavel, T., Sczyrba, A., McHardy, A. C., & Strowig, T. (2020). An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome. *Cell reports*, 30(9), 2909–2922.e6. <https://doi.org/10.1016/j.celrep.2020.02.036>
- Li, W., Jaroszewski, L., & Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics (Oxford, England)*, 17(3), 282–283. <https://doi.org/10.1093/bioinformatics/17.3.282>
- Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic acids research*, 42(Database issue), D490–D495. <https://doi.org/10.1093/nar/gkt1178>
- MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., & MacCoss, M. J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics (Oxford, England)*, 26(7), 966–968. <https://doi.org/10.1093/bioinformatics/btq054>
- Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., Jarrin, C., Chardon, P., Marteau, P., Roca, J., & Dore, J. (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, 55(2), 205–211. <https://doi.org/10.1136/gut.2005.073817>
- Mesuere, B., Debyser, G., Aerts, M., Devreese, B., Vandamme, P., & Dawyndt, P. (2015). The Unipept metaproteomics analysis pipeline. *Proteomics*, 15(8), 1437–1442. <https://doi.org/10.1002/pmic.201400361>
- Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids research*, 47(D1), D419–D426. <https://doi.org/10.1093/nar/gky1038>
- Milani, C., Duranti, S., Bottacini, F., Casey, E., Turrone, F., Mahony, J., Belzer, C., Delgado Palacio, S., Arbolea Montes, S., Mancabelli, L., Lugli, G. A., Rodriguez, J. M., Bode, L., de Vos, W., Gueimonde, M., Margolles, A., van Sinderen, D., & Ventura, M. (2017). The First Microbial Colonizers of the Human Gut: Composition, Activities, and Health

- Implications of the Infant Gut Microbiota. *Microbiology and molecular biology reviews* : MMBR, 81(4), e00036-17. <https://doi.org/10.1128/MMBR.00036-17>
- Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z., & Dominguez-Bello, M. G. (2015). The infant microbiome development: mom matters. *Trends in molecular medicine*, 21(2), 109–117. <https://doi.org/10.1016/j.molmed.2014.12.002>
- Nath, S., Handsley-Davis, M., Weyrich, L. S., & Jamieson, L. M. (2021). Diversity and bias in oral microbiome research: A commentary. *EClinicalMedicine*, 36, 100923. <https://doi.org/10.1016/j.eclinm.2021.100923>
- Navarro, & Vázquez, J. (2009). A Refined Method To Calculate False Discovery Rates for Peptide Identification Using Decoy Databases. *Journal of Proteome Research*, 8(4), 1792–1796. <https://doi.org/10.1021/pr800362h>
- Nightingale, A., Antunes, R., Alpi, E., Bursteinas, B., Gonzales, L., Liu, W., Luo, J., Qi, G., Turner, E., & Martin, M. (2017). The Proteins API: accessing key integrated protein and genome information. *Nucleic acids research*, 45(W1), W539–W544. <https://doi.org/10.1093/nar/gkx237>
- NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., Baker, C. C., Di Francesco, V., Howcroft, T. K., Karp, R. W., Lunsford, R. D., Wellington, C. R., Belachew, T., Wright, M., Giblin, C., ... Guyer, M. (2009). The NIH Human Microbiome Project. *Genome research*, 19(12), 2317–2323. <https://doi.org/10.1101/gr.096651.109>
- Orsburn B. C. (2021). Proteome Discoverer-A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes*, 9(1), 15. <https://doi.org/10.3390/proteomes9010015>
- Ostrouchov, K. (2022). *Bioinformatic_Tutorials*. GitHub. Available from: https://github.com/kostrouc/Bioinformatics_Tutorials/
- Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825
- Pereira-Marques, J., Hout, A., Ferreira, R. M., Weber, M., Pinto-Ribeiro, I., van Doorn, L. J., Knetsch, C. W., & Figueiredo, C. (2019). Impact of Host DNA and Sequencing Depth on

- the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in microbiology*, 10, 1277. <https://doi.org/10.3389/fmicb.2019.01277>
- Polpitiya, A. D., Qian, W. J., Jaitly, N., Petyuk, V. A., Adkins, J. N., Camp, D. G., 2nd, Anderson, G. A., & Smith, R. D. (2008). DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics (Oxford, England)*, 24(13), 1556–1558. <https://doi.org/10.1093/bioinformatics/btn217>
- Ragan, M. A., Bernard, G., & Chan, C. X. (2014). Molecular phylogenetics before sequences: oligonucleotide catalogs as k-mer spectra. *RNA biology*, 11(3), 176–185. <https://doi.org/10.4161/rna.27505>
- Ram, R. J., Verberkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., 2nd, Shah, M., Hettich, R. L., & Banfield, J. F. (2005). Community proteomics of a natural microbial biofilm. *Science (New York, N.Y.)*, 308(5730), 1915–1920.
- 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>
- Rodriguez-R, L. M., & Konstantinidis, K. T. (2014). Estimating coverage in metagenomic data sets and why it matters. *The ISME journal*, 8(11), 2349–2351. <https://doi.org/10.1038/ismej.2014.76>
- Ruiz, L., Delgado, S., Ruas-Madiedo, P., Sánchez, B., & Margolles, A. (2017). Bifidobacteria and Their Molecular Communication with the Immune System. *Frontiers in microbiology*, 8, 2345. <https://doi.org/10.3389/fmicb.2017.02345>
- 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>
- Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z.,

- Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., Skripchenko, Y., ... Sherry, S. T. (2021). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 49(D1), D10–D17.
<https://doi.org/10.1093/nar/gkaa892>
- Starr, A. E., Deeke, S. A., Li, L., Zhang, X., Daoud, R., Ryan, J., Ning, Z., Cheng, K., Nguyen, L., Abou-Samra, E., Lavallée-Adam, M., & Figeys, D. (2018). Proteomic and Metaproteomic Approaches to Understand Host-Microbe Interactions. *Analytical chemistry*, 90(1), 86–109. <https://doi.org/10.1021/acs.analchem.7b04340>
- Stasiak A. (2003). The first lady of DNA. *EMBO Reports*, 4(1), 14.
<https://doi.org/10.1038/sj.embor.embor723>
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, 28(1), 112–118.
<https://doi.org/10.1093/bioinformatics/btr597>
- Sullivan, B. (2019). Pleased to meet me: How genes, germs, and the environment make us who we are. *National Geographic*.
- Tanca, A., Palomba, A., Deligios, M., Cubeddu, T., Fraumene, C., Biosà, G., Pagnozzi, D., Addis, M. F., & Uzzau, S. (2013). Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PloS one*, 8(12), e82981. <https://doi.org/10.1371/journal.pone.0082981>
- Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., Muth, T., Rapp, E., Martens, L., Addis, M. F., & Uzzau, S. (2016). The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome*, 4(1), 51. <https://doi.org/10.1186/s40168-016-0196-8>
- Tang, Q., Jin, G., Wang, G., Liu, T., Liu, X., Wang, B., & Cao, H. (2020). Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices. *Frontiers in cellular and infection microbiology*, 10, 151. <https://doi.org/10.3389/fcimb.2020.00151>
- The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic acids research*, 47(D1), D330–D338. <https://doi.org/10.1093/nar/gky1055>

- Timmins-Schiffman, E., May, D. H., Mikan, M., Riffle, M., Frazar, C., Harvey, H. R., Noble, W. S., & Nunn, B. L. (2017). Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *The ISME journal*, 11(2), 309–314. <https://doi.org/10.1038/ismej.2016.132>
- Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. (2013). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?. *Briefings in bioinformatics*, 14(3), 315–326. <https://doi.org/10.1093/bib/bbs034>
- Trovato, M., Funck, D., Forlani, G., Okumoto, S., & Amir, R. (2021). Editorial: Amino Acids in Plants: Regulation and Functions in Development and Stress Defense. *Frontiers in plant science*, 12, 772810. <https://doi.org/10.3389/fpls.2021.772810>
- 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>
- UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic acids research*, 49(D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- Van Den Bossche, T., Kunath, B. J., Schallert, K., Schäpe, S. S., Abraham, P. E., Armengaud, J., Arntzen, M. Ø., 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., Jensen, M., ... Muth, T. (2021). Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nature communications*, 12(1), 7305. <https://doi.org/10.1038/s41467-021-27542-8>
- 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>
- Vaudel, M., Burkhardt, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A., Martens, L., & Barsnes, H. (2015). PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature biotechnology*, 33(1), 22–24. <https://doi.org/10.1038/nbt.3109>

- Wang, G., Dassanayake, M. (2020). GOMCL: a toolkit to cluster, evaluate, and extract non-redundant associations of Gene Ontology-based functions. *BMC Bioinformatics*. 21, 139. <https://doi.org/10.1186/s12859-020-3447-4/>
- Weinhold B. (2006). Epigenetics: the science of change. *Environmental health perspectives*, 114(3), A160–A167. <https://doi.org/10.1289/ehp.114-a160>
- Wilmes, P., Andersson, A. F., Lefsrud, M. G., Wexler, M., Shah, M., Zhang, B., Hettich, R. L., Bond, P. L., VerBerkmoes, N. C., & Banfield, J. F. (2008). Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *The ISME journal*, 2(8), 853–864. <https://doi.org/10.1038/ismej.2008.38>
- 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>
- Wilmes, P., Heintz-Buschart, A., & Bond, P. L. (2015). A decade of metaproteomics: where we stand and what the future holds. *Proteomics*, 15(20), 3409–3417. <https://doi.org/10.1002/pmic.201500183>
- Woo, J., Williams, S. M., Markillie, L. M., Feng, S., Tsai, C. F., Aguilera-Vazquez, V., Sontag, R. L., Moore, R. J., Hu, D., Mehta, H. S., Cantlon-Bruce, J., Liu, T., Adkins, J. N., Smith, R. D., Clair, G. C., Pasa-Tolic, L., & Zhu, Y. (2021). High-throughput and high-efficiency sample preparation for single-cell proteomics using a nested nanowell chip. *Nature communications*, 12(1), 6246. <https://doi.org/10.1038/s41467-021-26514-2>
- Woo, P. C., Tsoi, H. W., Leung, K. W., Lum, P. N., Leung, A. S., Ma, C. H., Kam, K. M., & Yuen, K. Y. (2000). Identification of *Mycobacterium neoaurum* isolated from a neutropenic patient with catheter-related bacteremia by 16S rRNA sequencing. *Journal of clinical microbiology*, 38(9), 3515–3517. <https://doi.org/10.1128/JCM.38.9.3515-3517.2000>
- Xiao, J., Tanca, A., Jia, B., Yang, R., Wang, B., Zhang, Y., & Li, J. (2018). Metagenomic Taxonomy-Guided Database-Searching Strategy for Improving Metaproteomic Analysis.

- Journal of proteome research, 17(4), 1596–1605.
<https://doi.org/10.1021/acs.jproteome.7b00894>
- Xiong, W., Abraham, P. E., Li, Z., Pan, C., & Hettich, R. L. (2015). Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics*, 15(20), 3424–3438. <https://doi.org/10.1002/pmic.201400571>
- Yoon, H. S., Ju, J. H., Lee, J. E., Park, H. J., Lee, J. M., Shin, H. K., Holzapfel, W., Park, K. Y., & Do, M. S. (2013). The probiotic *Lactobacillus rhamnosus* BFE5264 and *Lactobacillus plantarum* NR74 promote cholesterol efflux and suppress inflammation in THP-1 cells. *Journal of the science of food and agriculture*, 93(4), 781–787.
<https://doi.org/10.1002/jsfa.5797>
- Zaslavsky, L., Ciufo, S., Fedorov, B., & Tatusova, T. (2016). Clustering analysis of proteins from microbial genomes at multiple levels of resolution. *BMC bioinformatics*, 17 Suppl 8(Suppl 8), 276. <https://doi.org/10.1186/s12859-016-1112-8>
- Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., & Ma, B. (2012). PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics : MCP*, 11(4), M111.010587. <https://doi.org/10.1074/mcp.M111.010587>
- Zhang Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11), 218. <https://doi.org/10.21037/atm.2016.03.37>
- Zhao, Y., & Lin, Y. H. (2010). Whole-cell protein identification using the concept of unique peptides. *Genomics, proteomics & bioinformatics*, 8(1), 33–41.
[https://doi.org/10.1016/S1672-0229\(10\)60004-6](https://doi.org/10.1016/S1672-0229(10)60004-6)
- Zhu, J., Ren, H., Zhong, H., Li, X., Zou, Y., Han, M., Li, M., Madsen, L., Kristiansen, K., & Xiao, L. (2021). An Expanded Gene Catalog of Mouse Gut Metagenomes. *mSphere*, 6(1), e01119-20. <https://doi.org/10.1128/mSphere.01119-20>
- Zommiti, M., Feuilleley, M., & Connil, N. (2020). Update of Probiotics in Human World: A Nonstop Source of Benefactions till the End of Time. *Microorganisms*, 8(12), 1907.
<https://doi.org/10.3390/microorganisms8121907>

Zuo, F., Gu, Q., Li, S., Wei, H., & Peng, J. (2019). Effects of Different Methionine Sources on Methionine Metabolism in the IPEC-J2 Cells. *BioMed research international*, 2019, 5464906. <https://doi.org/10.1155/2019/5464906>

APPENDIX

```
1 FILE="otus.txt" #Create txt file with OTUs and assign variable name
2 while read line; do #Starts a loop to read file of OTUs
3   FNAM="${line/ /_}" #Renames file with _ for a space
4   datasets download genome taxon "$line" --exclude-gff3 --exclude-rna
5   --exclude-seq --exclude-genomic-cds --reference --dehydrated
6   --filename ${FNAM}.zip #Download RefSeq dehydrated package
7   unzip ${FNAM}.zip -d ${FNAM} #Unzipping package
8   dataformat excel genome
9   --inputfile ${FNAM}/ncbi_dataset/data/assembly_data_report.jsonl
10  --fields organism-name,tax-id,assminfo-refseq-assm-accession
11  -o ./lists/${FNAM}.xlsx #Recording org name, taxid, and accession obtained
12  done <${FILE}
13  #Check if these are the files to obtain prior to downloading metadata
14  #Revise list if needed
15  while read line; do
16    REHY="${line/ /_}" #Adding the _ where the space is to find the directory
17    datasets rehydrate --directory ${REHY}/ #Rehydrate directory to obtain metadata
18  done <${FILE}
19  #Creating proteome folder names matching OTU names
20  for f in ./ncbi_dataset/data/*; do
21    out1=$( echo $f | cut -f 2 -d '/')
22    mkdir ./proteins/${basename $out1}
23  done
24  #Moving RefSeq proteomes to new folders
25  for b in ./ncbi_dataset/data/*/protein.faa; do
26    out2=$( echo $b | cut -f 2,5 -d '/')
27    mv $b proteins/${out2}.faa
28  done
```

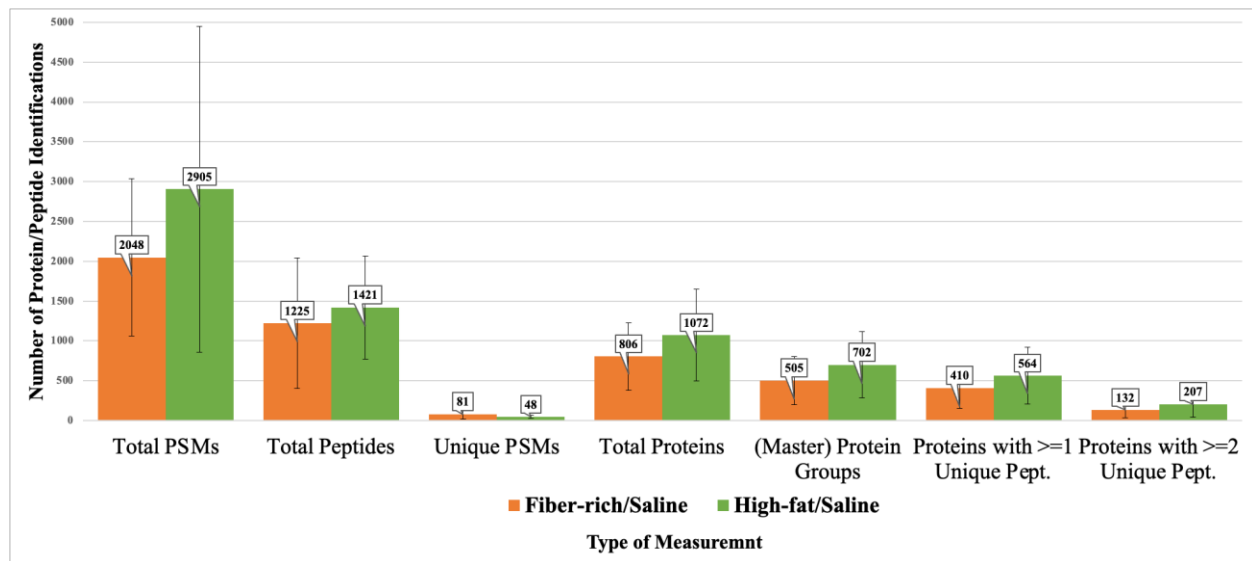
Supplementary Figure 1 Obtaining thousands of RefSeq proteomes using a list of taxonomy or common names in bash. This script was modified from the open-source NCBI Datasets package (Sayers, 2021) and a GitHub tutorial (Ostrouchov, 2022) with an MIT License to obtain RefSeq proteomes. It can be downloaded with conda (Anaconda Software Distribution, 2020) using “conda install -c conda-forge ncbi-datasets-cli”. Command-line help can be accessed with “datasets --h”.

```

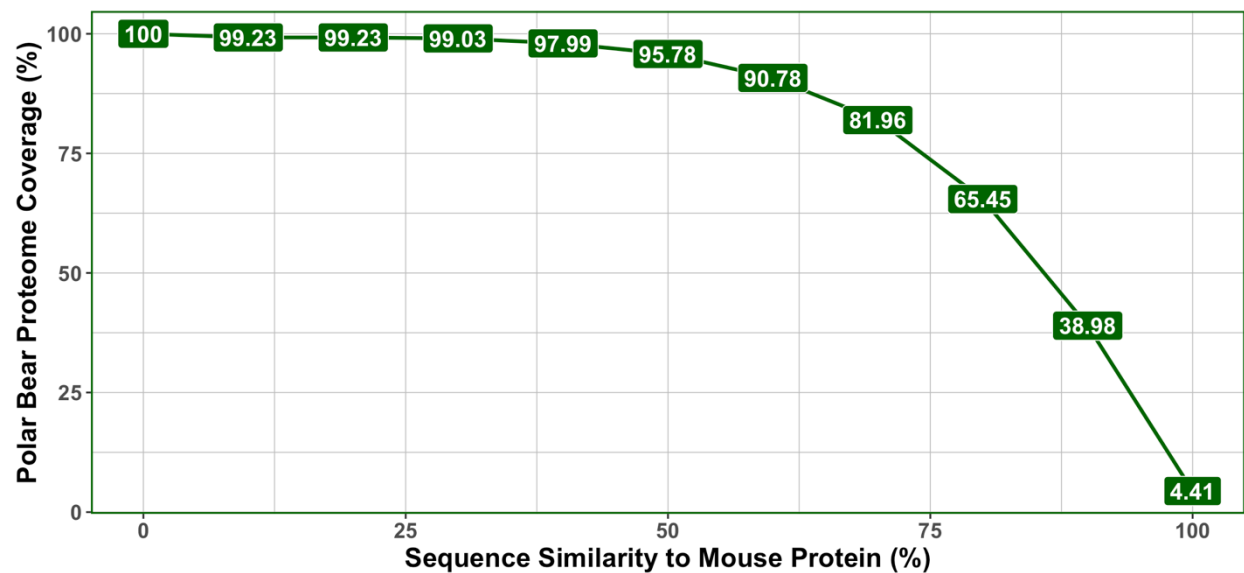
1 #Save Perl script below as apidownload.pl
2 use strict;
3 use warnings;
4 use LWP::UserAgent;
5 use HTTP::Date;
6 my $top_node = $ARGV[0];
7 my $agent = LWP::UserAgent->new;
8 my $query_list = "https://www.uniprot.org/proteomes/?query=reference:yes+taxonomy:$top_node&format=list";
9 my $response_list = $agent->get($query_list);
10 die 'Failed, got ' . $response_list->status_line .
11 ' for ' . $response_list->request->uri . "\n"
12 unless $response_list->is_success;
13 for my $proteome (split(/\n/, $response_list->content)) {
14     my $file = $proteome . '.fasta.gz';
15     my $query_proteome = "https://www.uniprot.org/uniprot/?query=proteome:$proteome&format=fasta&compress=yes";
16     my $response_proteome = $agent->mirror($query_proteome, $file);
17     if ($response_proteome->is_success) {
18         my $results = $response_proteome->header('X-Total-Results');
19         my $release = $response_proteome->header('X-UniProt-Release');
20         my $date = sprintf("%4d-%02d-%02d", HTTP::Date::parse_date($response_proteome->header('Last-Modified')));
21         print "File $file: downloaded $results entries of UniProt release $release ($date)\n";
22     } elsif ($response_proteome->code == HTTP::Status::RC_NOT_MODIFIED) {
23         print "File $file: up-to-date\n";
24     } else {die 'Failed, got ' . $response_proteome->status_line .
25         ' for ' . $response_proteome->request->uri . "\n";}}
26 #In bash terminal run the following script below to obtain list of UniProt proteomes
27 FILE=OTUs.txt
28 while read line; do
29     FNAM="${line/ /_}"
30     mkdir -p $FNAM
31     cd $FNAM
32     perl ./apidownload.pl $line
33     cd ..
34 done <$FILE

```

Supplementary Figure 2 Obtaining thousands of UniProt proteomes using taxonomy or common name. This Perl and bash script was modified from UniProt (Nightingale, 2017) and from a GitHub tutorial (Ostrouchov, 2022). It is freely available under the Creative Commons Attribution 4.0 International Public License and the MIT License respectively.

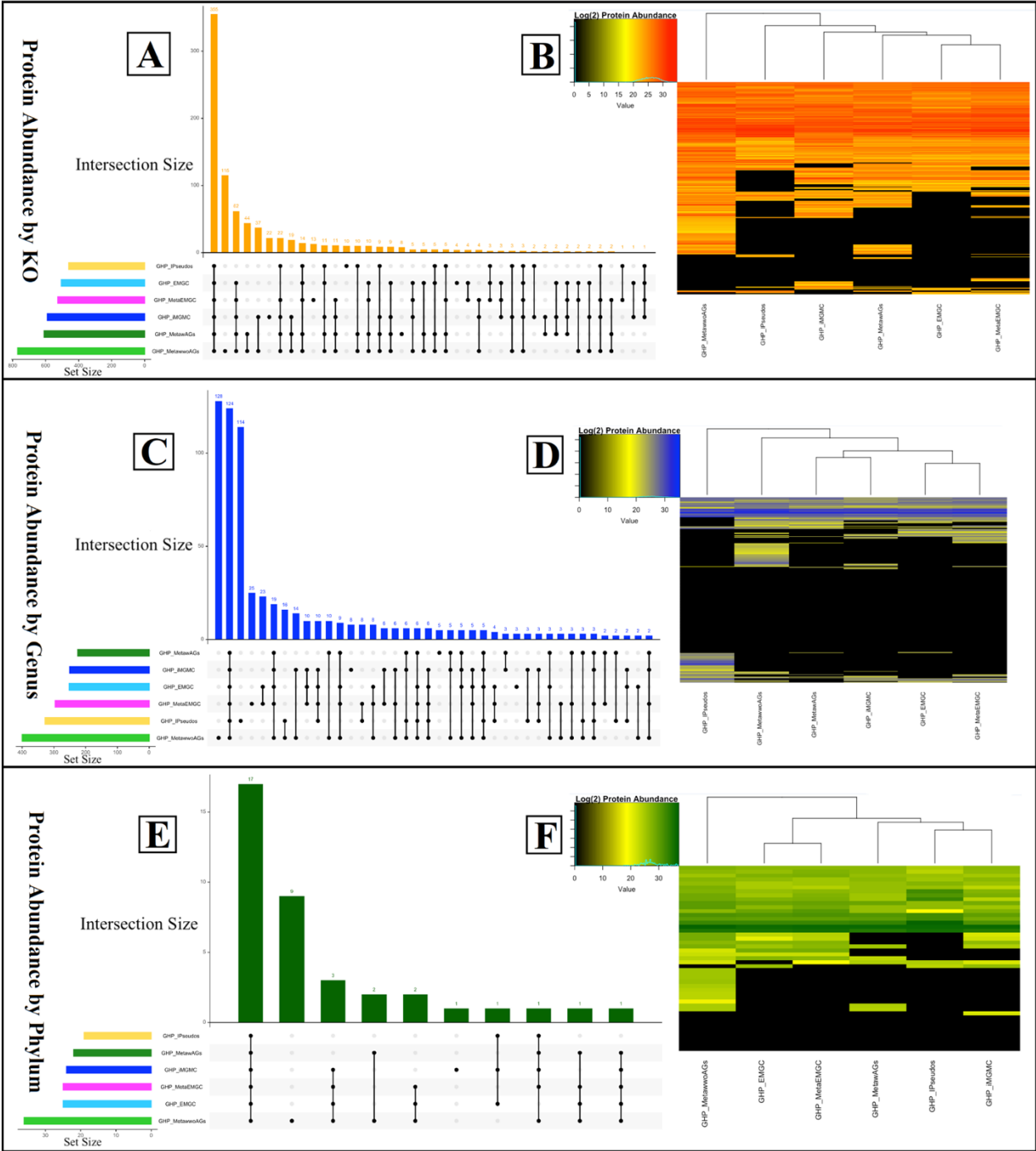


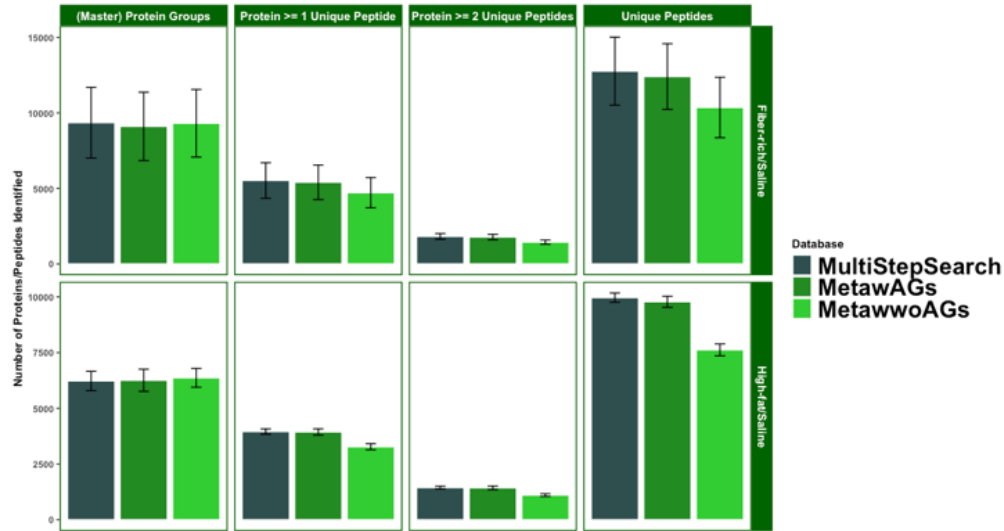
Supplementary Figure 3 Entrapment proteins/peptides identified for fiber-rich/saline (orange) & high-fat/saline (green) mice at FDR 0.01. Contamination was removed from these calculations and error bars represent the standard deviation across the three samples in a condition.



Supplementary Figure 4 BLASTP results comparing microbial and polar bear entrapment protein sequences to all databases. A BLASTP search of the polar bear proteome contained in the entrapment database was searched against the mouse proteome. Sequence redundancy within each database was removed at 98% with cd-hit prior to this search. Points highlighted in green represent the amount of the entire polar bear proteome that had proteins with sequences at certain similarity thresholds to mouse. For example, thirty-nine percent of the proteins found within the polar bear proteome were ninety percent like all the mouse protein sequences.

Supplementary Figure 5 Summed GhostKoala KO (A/B), genus (C/D), and phylum (E/F) protein abundance values graphed in UpSetR plot and Euclidean distance heatmaps for high-fat/saline (GHP) condition. All identified protein sequences across databases were submitted through GhostKOALA to annotate the functional and taxonomic information gained from each database. LIT corresponds to the high-fiber/saline mice, and GHP corresponds to the high-fat/saline mice treatment groups. **A/C/E:** The top bar chart of the UpSetR plot (Intersection Size) is displaying the total number of proteins belonging to either KO, phylum, or genus which were found in the sets of databases connected by dots and labeled in the plot directly below. The highest numbers from the top bar chart correspond to the sets of databases which identified the most proteins associated with each KO, phylum, or genus from GhostKOALA. The bar chart to the left in each UpSetR plot (Set Size) measures how many unique identifications were contributed by each database. The largest bar on the left shows which database provided the highest amount of functional (KO), and taxonomic (phylum, genus) information. **B/D/F:** Euclidian distance heatmaps depict all the summed protein abundance values for each KO, phylum, and genus identified across databases. Black represents absence of a KO, phylum, or genus assignment. The range in colors depicts how the Log 2 protein abundance changed according to database utilized.





Supplemental Figure 6 Total protein and peptide identifications obtained for the multi-step-search versus the matched-metagenomes. Entrapment and contaminants were filtered out utilizing grep on the command-line, and a Python script was run to obtain the master proteins, proteins with unique peptides, and unique peptides.

Supplementary Table 1 Total entrapment protein identifications obtained for multi-step-search versus the matched-metagenomes. Grep was utilized on the command-line to obtain these counts directly from the output PD protein tables.

Type of Protein Identification	MultiStepSearch	MetawAGs	MetawwoAGs
<i>Mus musculus</i>	917	856	846
<i>Aquifex aeolicus</i>	1	1	1
<i>Pseudothermotoga lettingae</i>	1	2	0
<i>Pyrodictium occultum</i>	1	1	1
<i>Thermotoga maritima</i>	1	1	0
<i>Hyperthermus butylicus</i>	0	0	0
<i>Ursus maritimus</i>	371	345	339
Total Entrapment	375	350	341
Total Contaminants	20	20	20
Total without Contaminants or Entrapment	13393	13370	16325
Total without Mouse, Contaminants, and Entrapment	12476	12514	15479
Total Before Filtering	13788	13740	16686

VITA

Katherine E. Ostrouchov was born on April 7th, 1993, in Knoxville, Tennessee, to Dr. George Ostrouchov and Susan Blackford. She completed her undergraduate degree in biochemistry, cellular and molecular biology at the University of Tennessee, Knoxville. While pursuing her undergraduate degree, Katherine served six years in the Tennessee Army National Guard as a military policewoman and technical engineer and achieved the rank of sergeant. Upon completion of her Bachelor's, she pursued an internship at the Oak Ridge National Lab which led to her acceptance into the Genome Science and Technology graduate program. She has accepted a job offer to work as an LC-MS Scientist II starting in September of 2022 and expects to receive her Master of Science at the end of Fall 2022.