

**QUANTITATIVE AND FUNCTIONAL ANALYSIS
PIPELINE FOR LABEL-FREE METAPROTEOMICS
DATA AND ITS APPLICATIONS**

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Lang Ho Lee
August 2015

Copyright © 2015 by Lang Ho Lee
All rights reserved.

ACKNOWLEDGEMENTS

First, I'd like to gratefully and sincerely thank Dr. Nathan VerBerkmoes for guiding me through the completion of my dissertation and subsequent Ph.D. He gave me a chance to work at Oak Ridge National Laboratory as well as Berg Phama Inc. as a research assistant. Also, he never stopped supporting me while he had no easy matters. I believe that he will successfully continue his researches at his new position.

Secondly, I'm very grateful to the remaining members of my dissertation committee, Dr. Arnold Saxton, Dr. Tim Sparer, Dr. Chongle Pan and Dr. Tamah Fridman. Their academic support and input are greatly appreciated. Especially, I appreciate to Dr. Arnold Saxton who has provided me another home as well as academic advises. I'd also like to express my gratitude to Dr. Steven Wilhelm and Dr. Bhanu Rekepalli for their support and help.

Thirdly, I'd like to thank my parents Jae Duk Lee and Ok Hee Kim and parents in law Young Gak Shin and Jung Bae Kwak as well as my relatives. Their prayers and supports have helped me a lot. I couldn't finish Ph.D. without their support and encouragement.

Finally, and most importantly, my upmost gratitude goes to my family, Bo Hee and Yoel. They followed after me and moved to where they never heard before. Their encouragement, support and patience were undeniably precious source of power to keep on studying. Obviously, without these two, it would not

have been possible to complete my study and to build up all things together.
Above all, I am greatly thankful to God who guided, inspired and encouraged me
to go this way.

ABSTRACT

Since the large-scale metaproteome was first reported in 2005, metaproteomics has advanced at a tremendous rate both in its quantitative and qualitative metrics. Furthermore metaproteomics is now being applied as a general tool in microbial ecology in a large variety of environmental studies. Though metaproteomics is becoming a useful and even a standard tool for the microbial ecologist, standardized bioinformatics pipelines are not readily available. Therefore, we developed quantitative and functional analysis pipeline for metaproteomics (QFAM) to help analyze large and complicated metaproteomics data in a robust and timely fashion with outputs designed to be simple and clearly understood by the microbial ecologist.

QFAM starts by running peptide-spectrum searches against resultant MS/MS datasets with mixed metagenome/appropriate protein FASTA database. Its primary search algorithm is MyriMatch/IDPicker. MyriMatch/IDPicker uses multi-CPU's effectively, has an accurate scoring-system, correctly use the high MS accuracy data, and finally has a robust method for protein determination. These are required features for metaproteomics requiring large protein database and complicated peptide-structure.

QFAM has quantitative (QAM) and functional (FAM) analysis to provide dependable protein signatures and confident information for understanding the characteristics of the metaproteome. QAM employs a 'selfea' R package, which

provides probability models as well as Cohen's effect sizes. Our benchmark data test and Monte Carlo simulation results show that selfea can reduce false positives efficiently while losing few true positives; one of the key goals of proteomics and/or metaproteomics experiments.

FAM has two modules: BioSystems and COG analysis. The BioSystems module is most appropriate for well-annotated model organisms, such as humans, whereas the COG module is useful for less-annotated microorganisms and metagenome sequences. Both modules provide an enrichment test using Fisher's exact-test and a significance test using selfea. With two statistics, FAM generates differentially enriched functional terms that are insightful for discerning biological information held behind the metaproteome data.

Two application studies in chapter 4 and 5 show how QFAM can be employed for metaproteomics data analysis. QFAM is distinguished from other proteomics pipelines by multiprocessing as well as quantitative and functional analysis.

TABLE OF CONTENTS

CHAPTER I INTRODUCTION	1
Metaproteomics	1
Metaproteomics in proteomics	1
Shotgun metaproteomics	4
LC-MS/MS analysis	15
Bioinformatics for metaproteomics	28
Peptide-spectrum search and identification	28
Post-process of PSM search results	41
Quantitative and functional analysis pipeline for metaproteomics	51
CHAPTER II QUANTITATIVE AND FUNCTIONAL ANALYSIS PIPELINE FOR METAPROTEOMICS DATA	56
Quantitative and functional analysis for metaproteomics (QFAM)	56
Objectives of QFAM	56
General information of QFAM	56
Module for LC-MS/MS data analysis	59
Automated module for DB search and protein assembly	59
Making contrast file, comparing MS runs, and normalization	62
Module for quantitative and functional analysis	65
Quantitative analysis module (QAM)	65
Functional analysis module (FAM)	69
BioSystems analysis of FAM	71
COG analysis of FAM	76
Other modules of QFAM	83
Conclusion	89
CHAPTER III Selfea: R package for reliable feature selection	93
Introduction	93
Methods	96

Label free LC-MS/MS Proteomics Data	96
Calculation of Cohen's effect sizes and statistics	96
Benchmark data test and fecal proteome data test	98
Monte Carlo simulation	99
User guide for selfea package	100
Algorithm	100
Benchmark test.....	104
Tests using Gregori data.....	104
Tests using fecal proteome data.....	104
Monte Carlo simulation	110
Monte Carlo simulation	112
Conclusion.....	113
CHAPTER IV APPLICATION CASE 1 – HUMAN FECAL PROTEOME STUDY	114
Introduction.....	114
Methods.....	116
Sample preparation and LC-MS/MS experiment	116
PSM search and protein identification	116
Statistical analyses	119
Results.....	121
Data generation from 11 fecal and 5 PBMC samples.....	121
Comparison of 29 fecal and 5 PBMC MS runs	122
Human proteins in 11 fecal proteomes	128
Quasi-Poisson model construction using QFAM	137
Functional analysis using QFAM	140
Conclusion.....	150
CHAPTER V APPLICATION CASE 2 – HUMAN DIET PROTEOME STUDY..	153
Introduction.....	153
Methods.....	154
Generation of Metaproteome data	154
Sample preparation and LC-MS/MS experiment	156
Quantitation and normalization of metaproteome data	157

Quantitative analysis of metaproteome data.....	158
Functional analysis of metaproteome data	159
Identification of translated-metagenome sequences	161
Results.....	161
Generation of diet metaproteome data	161
Construction of generalized linear mixed model	165
Functional analysis using FAM	168
Conclusion	181
CHAPTER VI Conclusions	184
Overview.....	184
Current status of QFAM and remaining challenges.....	186
LIST OF REFERENCES	191
VITA	204

LIST OF TABLES

Table 1 Overview of metaproteomics studies	5
Table 2 Common proteases use in metaproteomics.....	12
Table 3 Accuracy and quantitative coverage of various labeling methods in proteomics	27
Table 4 MyriMatch search time	60
Table 5 MyriMatch run time on the Newton cluster.....	61
Table 6 Run time of LC-MS/MS data analysis in the pipeline	61
Table 7 An example result of the QAM	68
Table 8 An example of a contingency table for a pathway ID using the enrichment test of FAM.....	75
Table 9 Example result of FAM.....	77
Table 10 Comparison of the number annotated proteins between KEGG and COG.....	79
Table 11 A contingency table for one COG ID from the enrichment test in FAM.	84
Table 12 An example COG analysis in FAM.....	85
Table 13 Six methods used for comparing Cohen's ES and FC.....	105
Table 14 Comparison of MS runs of fecal sample and PBMC.....	122
Table 15 Protein identification frequency in fecal and PBMC proteomes	127
Table 16 Top 25 abundant human proteins in 11 fecal samples	129
Table 17 Top 25 abundant immune response proteins in 11 fecal samples.....	135
Table 18 Significantly varied GO terms in human fecal proteins	141
Table 19 Significant COG terms in microbial fecal proteins.....	144
Table 20 Significant COG functional terms in microbial fecal proteins	147
Table 21 The number of identified proteins in different species and in different diets	163
Table 22 166 significant COG terms in different diets.....	172

LIST OF FIGURES

Figure 1 Applications of proteomics	2
Figure 2 Overview of typical shotgun metaproteomics steps.....	8
Figure 3 <i>in silico</i> digestions of the human HeLa cell proteome with assorted proteases	13
Figure 4 These panels display distribution of P-scores, which is the peptide identification score of MaxQuant.....	16
Figure 5 The effect of increasing mass accuracy for unambiguous identification of compounds	20
Figure 6 Data dependent acquisition (DDA) and data independent acquisition (DIA).....	22
Figure 7 Relative amount of proteins in original sample, that are identified and quantified by metaproteomic analysis.....	24
Figure 8. Common quantitative mass spectrometry workflows.....	26
Figure 9 Example of <i>de novo</i> sequencing from MS/MS spectrum	31
Figure 10 Schematic diagram of the DB search for MS/MS spectrum analysis..	33
Figure 11 DB search of MS-GF+.....	36
Figure 12 Target-decoy DB comparison	39
Figure 13 Overview of bipartite graph model and analysis procedure of protein assembly in IDPicker	44
Figure 14 Comparison of the distributions fitted to spectral counts in label-free metaproteomics data	50
Figure 15 The interface of GalaxyP	52
Figure 16 A schematic diagram of workflow in MetaProteomeAnalyzer	54
Figure 17 A schematic diagram of QFAM	58
Figure 18 Boxplots showing the distribution of Pearson correlation coefficients among the different of replicates during the different stages of the proteomics analysis	64
Figure 19 Protein scatter plot for the quasi-Poisson model and Cohen's <i>f</i> combination.....	70

Figure 20 Overview of existing pathway analysis methods using gene expression data as an example.....	72
Figure 21 A diagram of the biosystems analysis of FAM	74
Figure 22 COG analysis of FAM	82
Figure 23 Examples of bar- and line-graph generated by QFAM	88
Figure 24 Example heat-map drew by QFAM's R library	90
Figure 25 Median power of studies included in neuroscience meta-analyses	94
Figure 26 Testing result with Gregori data and fecal proteome data	102
Figure 27 Protein scatter plots comparing three models and FC	106
Figure 28 Boxplots of 1,000 times Monte Carlo simulation of Gregori data	111
Figure 29 Experiment workflow of fecal metaproteome data generation	117
Figure 30 Hierarchical clustering of 29 MS runs of 11 fecal samples	124
Figure 31 Correlation coefficients between fecal and PBMC MS runs	125
Figure 32 Expression location of top 25 abundant human fecal proteins	130
Figure 33 Protein scatter plot of human fecal proteome comparison.....	139
Figure 34 Line plot showing expression profile of 15 COG functions	148
Figure 35 Experiment design of diet proteome study.....	155
Figure 36 Protein scatter plot of diet proteome comparison	167
Figure 37 Stacked bar graphs of accumulated spectral counts of 20 significant KEGG pathways	169
Figure 38 Heat-map of significant 166 COG terms	179

CHAPTER I INTRODUCTION

Metaproteomics

Metaproteomics in proteomics

The goal of proteomics is to obtain an all-inclusive quantitative collection of proteins in a cell, a multicellular organism, an extracellular environment and/or a community of multiple organisms. This comprehensive analysis is accomplished through protein extraction, identification and comparative analysis via liquid chromatography mass spectrometry (LC-MS), one-dimensional or two-dimensional gel electrophoresis-mass spectrometry (1D-GeLC-MS/2DE-MS), affinity purifications-mass spectrometry (AP-MS) and protein- or antibody-based microarray [1]. The goals of proteomics research are various and extensive, such as protein quantification on a large scale, finding drug targets, exploring clinical protein markers, structural characterization of protein complexes, imaging protein traffic in multicellular system and discovery of significant proteins from disease states such as neurodegenerative [2], coronary artery [3] and Lyme disease [1,4,5] (Fig. 1).

Proteomics has rapidly grown in technology, scope, and application over the last 20 years from the first simple proteomics experiments, which attempted to separate all of the proteins from *E. coli* on a 2d-PAGE system followed by the extremely laborious cutting out of spots one by one.

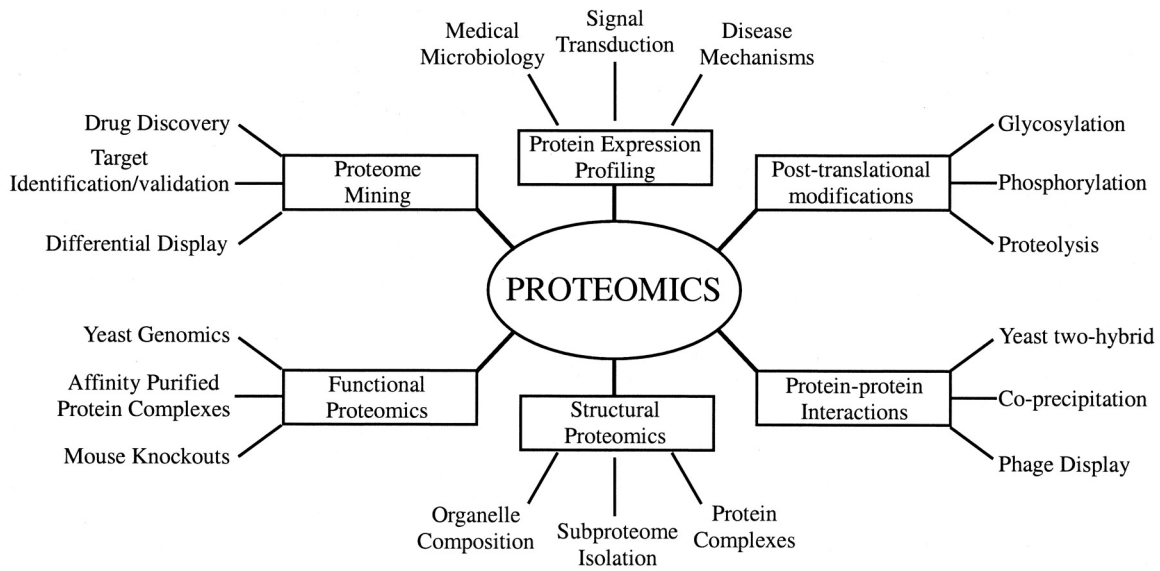


Figure 1 Applications of proteomics. Proteomics is applied to a diverse set of life science fields including basic scientific research, clinical protein marker discovery, and drug discovery. This figure was adapted from a book, 'Molecular Biologist's Guide to Proteomics' [5].

These spots are identified by either MALDI-TOF peptide mass fingerprinting and/or direct infusion nano-spray followed by tandem mass spectrometry (MS/MS) [6,7]. Today in numerous laboratories around the world, proteomics is a massive endeavor with sometimes 10-20 high throughput 2d-LC-MS/MS instruments (i.e., hybrid Velos Orbitraps, QExactives or Fusion MS systems) running 24/7 in “discovery mode” with another 10 or so 1d-LC-MS/MS instruments (i.e., high resolution triple quadrupole instruments) running in very accurate targeted quantitation mode. These labs attempt to use the power of MS-based proteomics to rapidly define protein biomarkers, novel protein therapeutics and many other novel human health applications. Furthermore, high-throughput proteomics laboratories are now expanding into metabolomics and lipidomics, for a holistic “omics” approach to understanding disease, human health, drug mechanism and a host of other biological applications. Clearly it can be stated that mass spectrometry is the fastest growing analytical application in the biological sciences today.

Metaproteomics (also called Community Proteomics, Environmental Proteomics, or Community Proteogenomics) can be defined as “the study of all proteins recovered directly from environmental sources and is an extension of MS-based proteomics [8]. Wilmes primarily used and defined metaproteomics as “the large-scale characterization of the entire protein complement of environmental microbiota at a given point in time” [9,10]. Per Wilmes definition, metaproteomics involves studying the protein complement of microbial communities that consist of multiple species. To date metaproteomics studies

have covered a diverse set of environmental samples including acid mine drainage biofilms [11], wastewaters [12], ocean sampling [13], soil systems [14], permafrost [15] and the human microbiome [16] (Table 1). Since the first large-scale metaproteome data was generated in 2005 [11], metaproteomics has rapidly advanced in terms of scale, depth, speed and quality. For example, Lichtman et al. [17] in 2013 identified 234 proteins human and microbial proteins, but with today's technology and improvements in metaproteomics, now over 56k proteins have been identified in fecal material, as described in our diet proteome study in Chapter 5.

Shotgun metaproteomics

Metaproteomics studies are generally based on bottom up approach, also known as shotgun proteomics [10,18,19]. The bottom-up method aims to identify and quantify peptides that are generally 7-20 amino acids in length, which are obtained from digesting proteins with sequencing grade trypsin although sometimes other proteolysis enzymes are used. The peptides are then injected onto a high performance liquid chromatography (HPLC) coupled with data dependent tandem mass spectrometer (LC-MS/MS) [18-20]. The top-down method directly investigates intact proteins either via LC-MS and/or LC-MS/MS analyses [20-22]. The bottom-up or shotgun proteomics methodologies tend to dominate MS-based proteomics in general and metaproteomics as well.

Table 1 Overview of metaproteomics studies. This table was adapted from VerBerkmoes et al [21].

<i>Microbiome</i>	<i>Number of peptides/ proteins identified</i>	<i>Protein/peptide separation method</i>	<i>MS platform</i>	<i>Peptide identification method</i>
Ocean	184/NA	2D-PAGE, 2D nano-LC	LCQ, MS/MS	Spectral matching, de novo
Acid mine drainage	6,188/2,033 (2p) 6,931/5,090 (1p)	2D nano-LC	LTQ MS/MS	Spectral matching
Lake and soil	NA/513 (1p)	2D nano-LC	Q-ToF, MS/MS	Spectral matching
Estuary	7/3 (2p)	2D-PAGE + LC	Q-ToF, MS/MS	De novo
Ocean	3/1 (2p)	1D-PAGE	MALDI-ToF, MS	Spectral matching
Riftia symbionts	NA/220 (2p)	2D-PAGE, 1D-PAGE + 2D nano-LC	MALDI-ToF MS, Q-ToF MS/MS	Spectral matching
Infant gastrointestinal tract	11/1 (1p)	2D-PAGE	MALDI-ToF MS	De novo
Acid mine drainage	8,137/3,234 (2p)	2D nano-LC	LTQ MS/MS	Spectral matching
Waste water treatment reactor	NA/109 (2p)	2D-PAGE	MALDI-ToF, MS/MS	Spectral matching, de novo
Contaminated soil/groundwater	NA/59(1p)	1D+2D-PAGE + LC	MS/MS	Spectral matching
Sludge	NA/46 (2p)	2D-PAGE	MALDI-ToF MS, Q-ToF MS/MS	Spectral matching
Sludge	4,472/2,378 (2p)	2D nano-LC	LTQ MS/MS, Orbitrap, MS/MS	Spectral matching
Sludge EPS	50/10 (1p)	1D-PAGE + LC	4000Qtrap, MS/MS	Spectral matching
Ocean	6,533/1,042 (1p-2p)	2D nano-LC	LTQ MS/MS	Spectral matching

Table 1 continued.

<i>Microbiome</i>	<i>Number of peptides/ proteins identified</i>	<i>Protein/peptide separation method</i>	<i>MS platform</i>	<i>Peptide identification method</i>
Acid mine drainage	NA/2,752 (2p)	2D nano-LC	Orbitrap, MS/MS	Spectral matching
Gut	NA/2,214 (2p)	2D nano-LC	Orbitrap, MS/MS	Spectral matching

Its advantages include large-scale data acquisition, improved ability to handle high-complexity samples, higher overall sensitivity, higher overall dynamic range, more accurate quantitation and better front-end separation of peptides vs. intact proteins via HPLC [20].

These advantages are all required features to profile highly complicated metaproteomes. Therefore, bottom-up proteomics is currently the most common methodology found in metaproteomics studies [21]. Although bottom-up approach is mainly used in metaproteomics studies, the top-down approach has its unique advantages. Top-down approach provides a holistic view of intact proteins and can assist post-translational modification (PTM) identification. Integration of bottom-up and top-down approaches was successfully applied in studies of microbial isolates *Shewanella oneidensis* and *Rhodopseudomonas palustris* [23,24].

The typical shotgun metaproteomics consists of five crucial steps: sample collection, protein extraction, protein digestion, LC-MS/MS analysis and bioinformatics analysis (Fig. 2). It's essential that each of these steps are optimized, biases minimized and carried out with great care and attention to all details as the entire experimental protocol can fail based on a single mistake in any step in any of the four critical steps.

Sample collection is the first step in metaproteomics studies. Piehowski et al showed that the biggest variation observed in proteomics experiments can be generated from sample collection [25].

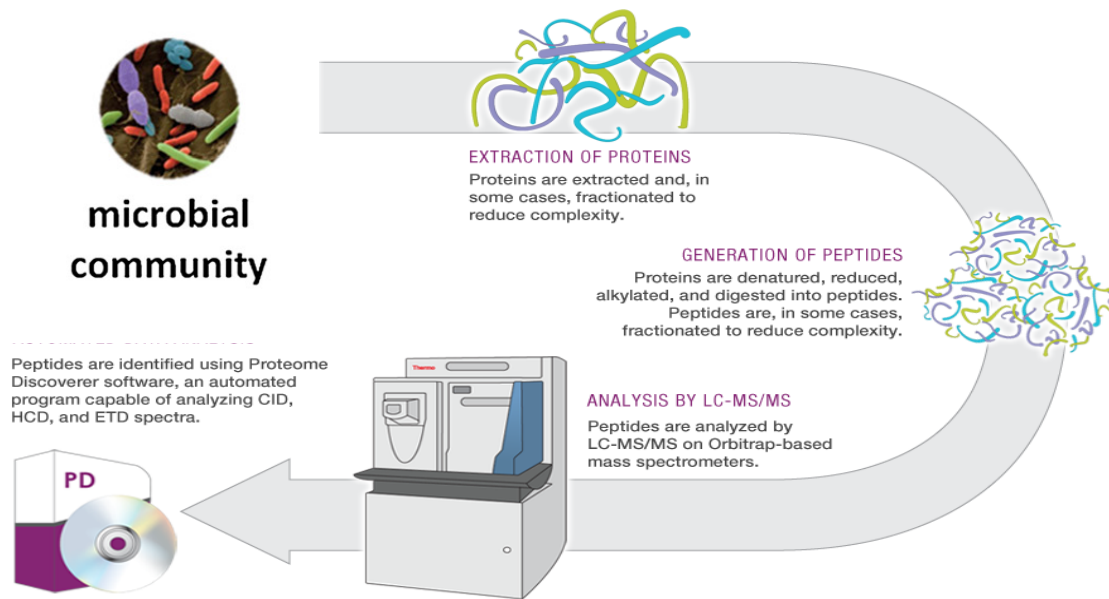


Figure 2 Overview of typical shotgun metaproteomics steps. Microbial ecologist carefully collects a sample containing the desired microbial community with great care to obtain a true *in situ* representation. Proteins are then extracted from the microbial community with an un-biased methodology, with goal of obtaining a true representation of the protein abundances. Then, the collected proteins are proteolytically digested by an enzyme, typically trypsin. The resultant peptides are desalted, concentrated, filtered and analyzed via 2D-LC-MS/MS. Finally, the resultant MS/MS spectra are identified and analyzed by many bioinformatics tools to obtain protein identification and quantification. This illustration was adapted from Tanca et al [10] and <http://www.planetorbitrap.com>.

Also, careful and appropriate sample collection and archiving are important issues in microbial ecology [26,27]. Biomolecules, such as DNA, RNA, and proteins should be protected from chemical and physical shock. Importantly, during the collection of the microbial community, great care should be taken so as to obtain a true *in-situ* snapshot of the proteome as it is in the natural environment; not as it is after extensive sample handling. Therefore, every environmental sample is generally stored in deep freezer at -80°C immediately after collection until protein extraction.

The second step major goal is the un-biased extraction of the proteome followed by complete protein denaturation and di-sulfide bond breakage. This prepares the proteome for enzymatic digestion via trypsin and/or other proteolytic enzymes. There are various methods of protein extraction by combining various detergents (e.g., sodium dodecyl sulfate (SDS), CHAPS or Triton X-100), organic/inorganic compounds (e.g., phenol, NaOH), chaotropic agents (e.g., urea or guanidine hydrochloride), reducing agents (e.g., dithiothreitol (DTT), tributylphosphine) and thermal treatments (e.g., boiling) and/or sonication/bead-beating/French press and etc. Detergents like SDS are used for cell lysis and chaotropic agents can denature proteins. Reducing agents are employed to break disulfide bonds and further denature proteins. Rapid and controlled thermal treatment can inactivate native enzymatic activity and prevent unwanted nonspecific proteome degradation [28]. Because there is no gold standard for metaproteome extraction, it is important to consider the characteristics of the given community and target proteins to maximize extraction yield. Very often

metaproteomics experts will try different methodologies on the environmental sample of interest until one obtains an optimal method. Because environmentally derived samples are often biomass limited due to difficulties and cost of sample acquisition, protein extraction is more challenging than proteomics studies of laboratory-grown microbes [29]. This is even more pronounced in the sub-field of soil metaproteomics, where the presences of humic acids, excess plant material and other abundant small molecules can make an efficient extraction of the microbial metaproteome difficult. Indeed, my PhD advisor, Dr. Nathan VerBerkmoes, has called the complex high organic soil metaproteome the most difficult analytical challenge on the planet to date.

After proteins are extracted from a sample, the next step is protein digestion. Before proteins are digested, chemical compounds, added for protein extraction, should be removed or diluted as not to disturb the following enzymatic digestion. This cleaning step is traditionally fulfilled by protein precipitation using trichloroacetic acid (TCA), followed by excess washing with chilled acetone or ammonium acetate/methanol to the protein extract solution [10]. However, these classical methods may lose certain types of proteins. Another alternative approach recently employed is filter-aided sample preparation (FASP), developed in 2009 by Mann et al [30] and modified by the VerBerkmoes lab for metaproteomics applications. This methodology has been successfully applied in many metaproteome studies [31-34] and it can combine sample cleanup and protein digestion on a centrifuge filter. This FASP methodology showed improve

performance over several competing methods, especially for low abundance proteins [31].

Protein digestion is an important step in shotgun metaproteomics because all the proteins in a bottom up approach are assembled from the digested peptides. Trypsin is widely used for proteolysis in metaproteomics because it is highly specific and generates a positively charged N-terminus and C-terminus, which aids in obtaining high quality MS and MS/MS spectra [35]. Because MS-based metaproteomics are usually run in positive ion mode, the positively charged ends are ideal for obtaining high quality MS and MS/MS spectra. Trypsin specifically targets carboxyl side of arginine and lysine of proteins in a weakly basic environment (Table 2). Nevertheless, trypsin has difficulty digesting tightly folded proteins that may not be denatured by the denaturing reagents and an inadequate distribution of trypsin cleavage sites may generate too many short peptides for proper analysis [36]. Furthermore, membrane proteins tend to be resistant to trypsin due to their limited number of lysine and arginine residues [36]. According to Wu et al, trypsin generates predominantly small peptides shorter than 20 amino acid length (Fig. 3) [37]. The small size of these peptides, which average between 7~20 residues [36], can facilitate MS detection, but cause a number of redundant peptides that are derived from more than two proteins. This is termed the uniqueness problem in metaproteomics and limits the analysis.

Table 2 Common proteases use in metaproteomics. This table is adapted from the website: http://www.proteinsandproteomics.org/content/free/tables_1/table11.pdf.

<i>Protease</i>	<i>Class</i>	<i>Cleavage site</i>
Chymotrypsin	Serin protease	After F, T, or Y
Endoproteinase Arg-C	Serin protease	After R
Endoproteinase Asp-N	Metalloprotease	Before D and cysteic acid
Endoproteinase Glu-C	Serin protease	After E, or after D or E
Endoproteinase Lys-C	Serin protease	After K
Pepsin	Acid protease	Broad specificity
Trypsin	Serin protease	After K or R

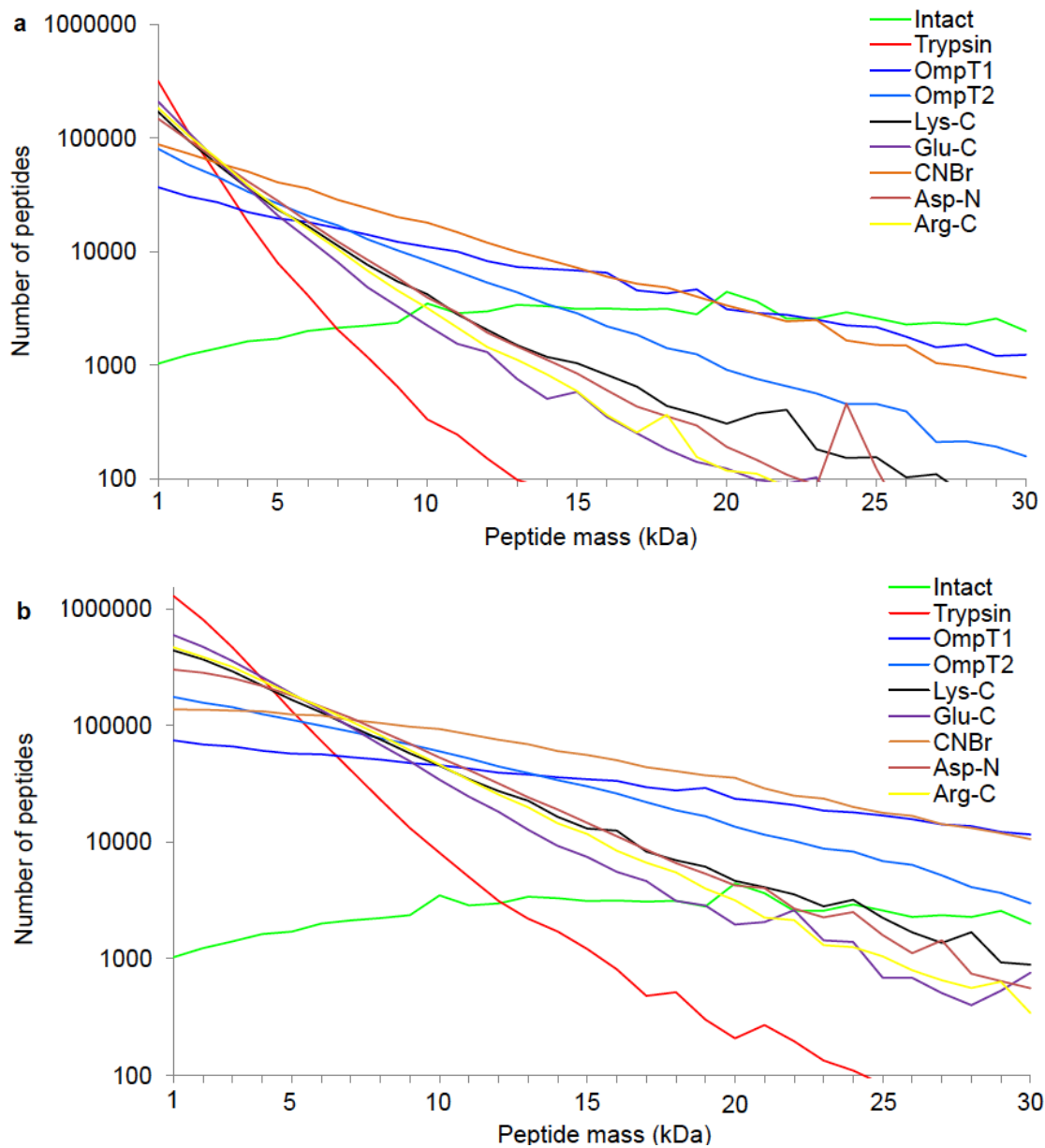


Figure 3 *in silico* digestions of the human HeLa cell proteome with assorted proteases. **A)** The number of peptides predicted without mis-cleavage and **B)** cleavage assuming two mis-cleavages. This figure is adapted from Wu et al [37].

Redundant non-unique peptides, often generated by trypsin digestion, are the biggest hurdle to accurately identifying proteins from the peptides identified by LC-MS/MS. This problem is exaggerated in metaproteomics due to the increase in database size and the fact that very often microbes carry very homologous proteins with very similar sequences. For example EF-TU, ribosomal, GroEL/GroES, other chaperones and glycolysis/TCA/oxidative phosphorylation pathway proteins can be highly conserved across a wide range of microbes in a given environmental sample.

Because redundant peptides can be found in more than two proteins, we cannot be sure where those peptides originate. Hence, these are termed as non-unique. Non-unique peptides limit the confidence of protein identification and quantification. In other words, we cannot know from which proteins these redundant peptides arose from and which protein they should be assigned to. To reduce the number of the problematic peptides, middle down proteomics is new method that attempts to use alternative proteases, which produce longer peptides [37-39]. Wu et al introduced the OmpT enzyme as replacement for trypsin to produce lengthy peptides longer than 3 kDa (Fig. 3) [37]. Peptide redundancy is generally dependent on the average length of produced peptides. While this methodology has not been widely used in metaproteomics, we feel that this is the next step in metaproteomics and would improve protein identification and quantification by limiting the “non-unique” problem, which has plagued metaproteomics studies.

In addition to a decrease in redundant peptides, longer peptides have less probability of false positive identification (Fig. 4) [40]. Figure 4 shows that the P-score Bayesian probability model in MaxQuant [40] changes based on peptide length. P-score distribution of false positive identifications shown in red gets smaller and narrower than true positive distribution shown in blue as the peptide length (L) gets longer. Metaproteomics studies using the middle down approach have not been published, but we believe this could be a potential path forward for solving the “non-unique” peptide problem.

LC-MS/MS analysis

The digested peptides are separated by their chemical and physical characteristics through high-performance liquid chromatography (HPLC), ionized by electrospray ionization and measured on an MS instrument. For most bottom-up proteomics and metaproteomics experiments the mass spectrometer is operated in data-dependent mode. This means that at any given point in time during the LC-MS/MS analysis, 50-100 peptides enter the mass spectrometer. Modern mass spectrometers have very high duty cycle (i.e., the speed going from one scan type to the next). In the data-dependent mode the peptides mass (i.e., m/z) are accurately measured by the mass spectrometer in the full scan mode. For example an Orbitrap mass spectrometer provides mass accuracies at ~1-10 p.p.m or generally the third decimal place of the peptide mass.

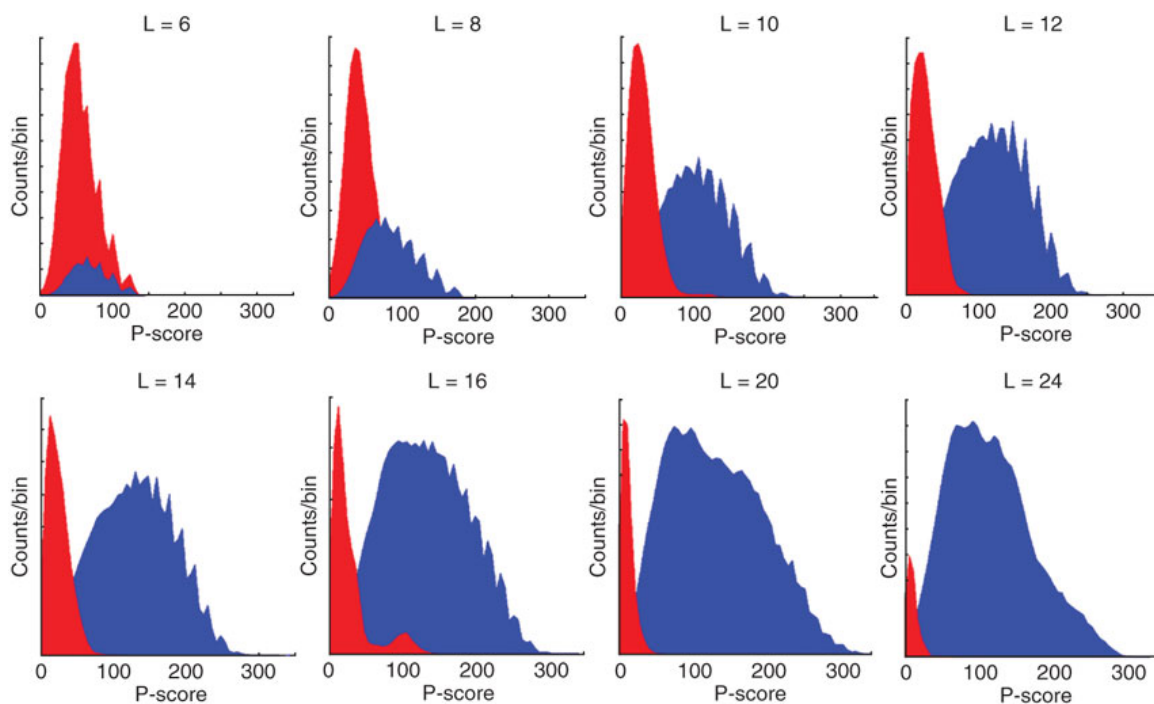


Figure 4 These panels display distribution of P-scores, which is the peptide identification score of MaxQuant. Blue color shows peptide identifications from a forward database search and red colors are from the reverse database, intended to be a false positive peptide database. L is the peptide length. This illustration was adapted from Cox et al [40].

Next the mass spectrometer picks the top 10-20 observed ions from the previous full scan, isolates them one by one and fragments them via energy deposition also termed MS/MS, CID, CAD, or now often called HCD. Peptide sequence information is obtained at this step. The intact peptides are isolated and fragmented down the peptide backbone thus providing sequence information. The accurate mass from the full scan coupled with the sequence information from the MS/MS scan are the two key points of information needed to confidently identify peptides via database searching (described below). The instrument cycles between full scans and data-dependent MS/MS scans during the entire LC-MS/MS analyses.

The key technology for this entire process is the simple but brilliant electrospray ionization (ESI) invented by John Fenn in 1984. This revolutionary technique allowed for the soft ionization of large biomolecules into the gas phase directly from a flowing liquid stream simply by the application of high voltage to a metal needle. The peptides are ionized by Coulomb explosion at Rayleigh limit [41,42]. ESI has revolutionized large molecule measurements using MS instrumentation. ESI does not require any chemical reagents besides water, methanol, and a small amount of organic acid. It is sensitive up to atto-mole levels and can easily ionize proteins up to 100,000 Da and much higher MW can be achieved with special instrumentation. ESI enables the analyses of all major biomolecules (i.e., DNA/RNA, proteins, lipids, sugars, and small metabolites) via mass spectrometry.

Because microbial communities usually have more proteins than single isolated microorganisms, high quality peptide separation is essential to achieve high quality metaproteomic analyses. If a massive number of peptide ions are ionized into the MS instrument at the same time, some of ions may not be appropriately measured or missed entirely due to the limitations of the MS scan speeds and MS dynamic range. Peptide separation should be designed by first considering the ability of MS instrument available to the metaproteomics researcher. Besides the scan speed, good peptides separation can reduce the density of the peptides entering the MS at any given time. In other words, similar m/z peptides can be pulled apart by the LC separation. Close m/z ions within the isolation window, which is generally 1-3 m/z units, are often sequenced together during MS/MS phase of the experiment. This is undesirable to say the least. This co-isolation and resultant MS/MS sequencing makes peptide identification difficult. To obtain better dynamic range, enhanced peptide identification and more confident protein quantification, the level of LC separation should be increased as much as possible. In order to obtain the maximum proteome depth and protein quantitation, the LC-MS/MS was run for 24 hours with 2-dimensions of liquid chromatography separation, strong cation exchanged (SCX) coupled with (reverse-phase) RP separation for the human fecal proteome study (Chapter 4) and influence of diet on the metaproteome (Chapter 5), we have termed this approach “deep and wide metaproteomics”. This means that the goal of this proteome experiment is to measure deep into the proteome of each microbial

species with wide coverage across all the microbial species present in the community.

High resolving power is an important virtue of the MS instrument for unambiguous protein identification in metaproteomics. The higher the mass accuracy, the more exact peptide identification. Low-resolution MS (LRMS) is less expensive and relatively easy to maintain, but cannot provide exact mass measurements. The high-resolution MS (HRMS) can determine several decimal places of mass accuracy. This high precision of HRMS can narrow down candidate peptides to only a few possibilities (Fig. 5). HRMS can give more exact p.p.m. mass accuracy than LRMS. P.P.M is calculated by $((\text{exact mass} - \text{accurate mass}) / \text{exact mass}) \times 10^6$. Lower p.p.m. means small mass error. Smaller p.p.m. (i.e., mass difference between observed and expected mass) can reduce false matches by removing similar weighted false peptide candidates. For metaproteomics experiments high mass accuracy for the full scan MS is absolutely essential. High resolution for the MS/MS spectra provides even better results. Because the potential metaproteome of a complex community is much more complicated than a single species proteome, it requires HRMS to narrow down candidate peptides. Due to the need for high mass accuracy, the Orbitrap MS is widely used in metaproteomics studies because they offer higher resolution/accurate mass analyses in a stable instrument platform.

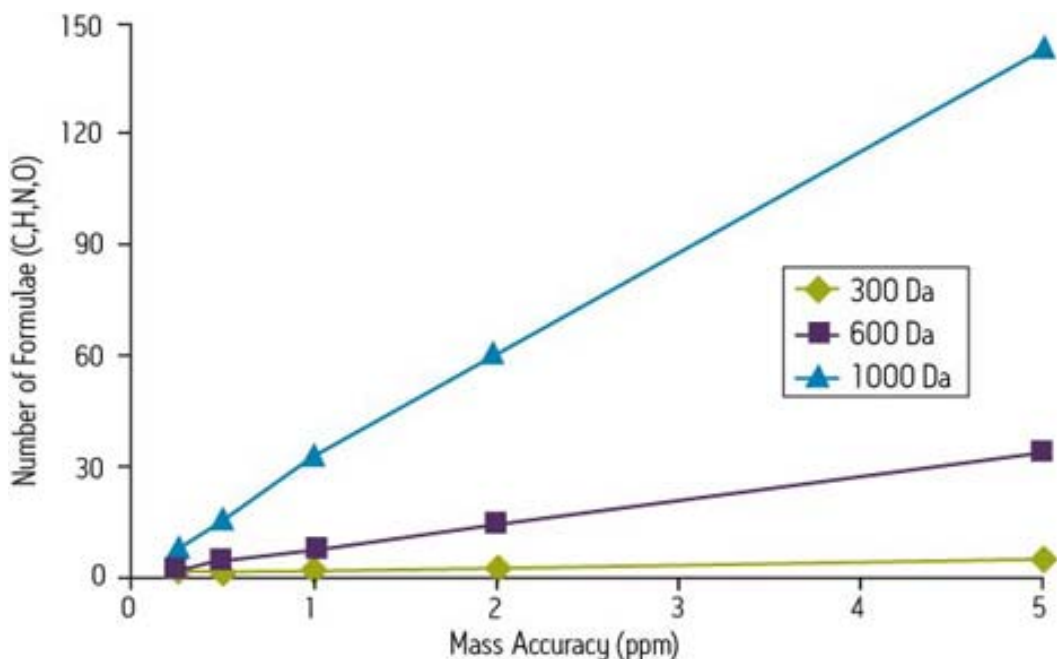


Figure 5 The effect of increasing mass accuracy for unambiguous identification of compounds. P.P.M. is a parts per million ($((\text{exact mass} - \text{accurate mass}) / \text{exact mass}) \times 10^6$). Lower p.p.m. means small mass error. If mass accuracy is low (i.e., p.p.m. is high), the number of candidate compounds increases linearly. As the slope gets steeper, the molecular weight gets larger. This illustration is adapted from Quenzer et al [43].

The Orbitrap MS is easy to use and maintain, produces high quality data, has excellent dynamic range and sensitivity and enables tandem in space and time MS/MS measurements.

Even though high mass accuracy can provide better resolution for peptide identification, it is not enough to identify peptides, especially from highly complex metaproteomics samples. Tandem mass spectrometry or MS/MS is required for exact peptide identification and sequencing. In the MS/MS step, target ions from the survey full-scan mass spectrum (MS1) are selected, isolated, fragmented and the mass analyzed. This MS/MS peak selection can be modified for different acquisition approaches, data dependent acquisition (DDA), and data independent acquisition (DIA) (Fig. 6).

In DDA approach, a fixed number of peaks from the MS1 scan are selected, isolated, and subjected to MS/MS analysis (Fig. 6). DDA is powerful and versatile, but it limits the number of peptides sampled for fragmentation regardless of dynamic range and peak capacity of the MS system [44]. In a complex protein mixture, up to 84% of peptides are not sampled for MS/MS, while at most 30% of peptides are analyzed by MS/MS [45] via the DDA approach. DIA is an alternative approach to DDA. DIA collects holistic MS/MS spectra of all the peaks systematically and independently of the precursor ions (Fig. 6).

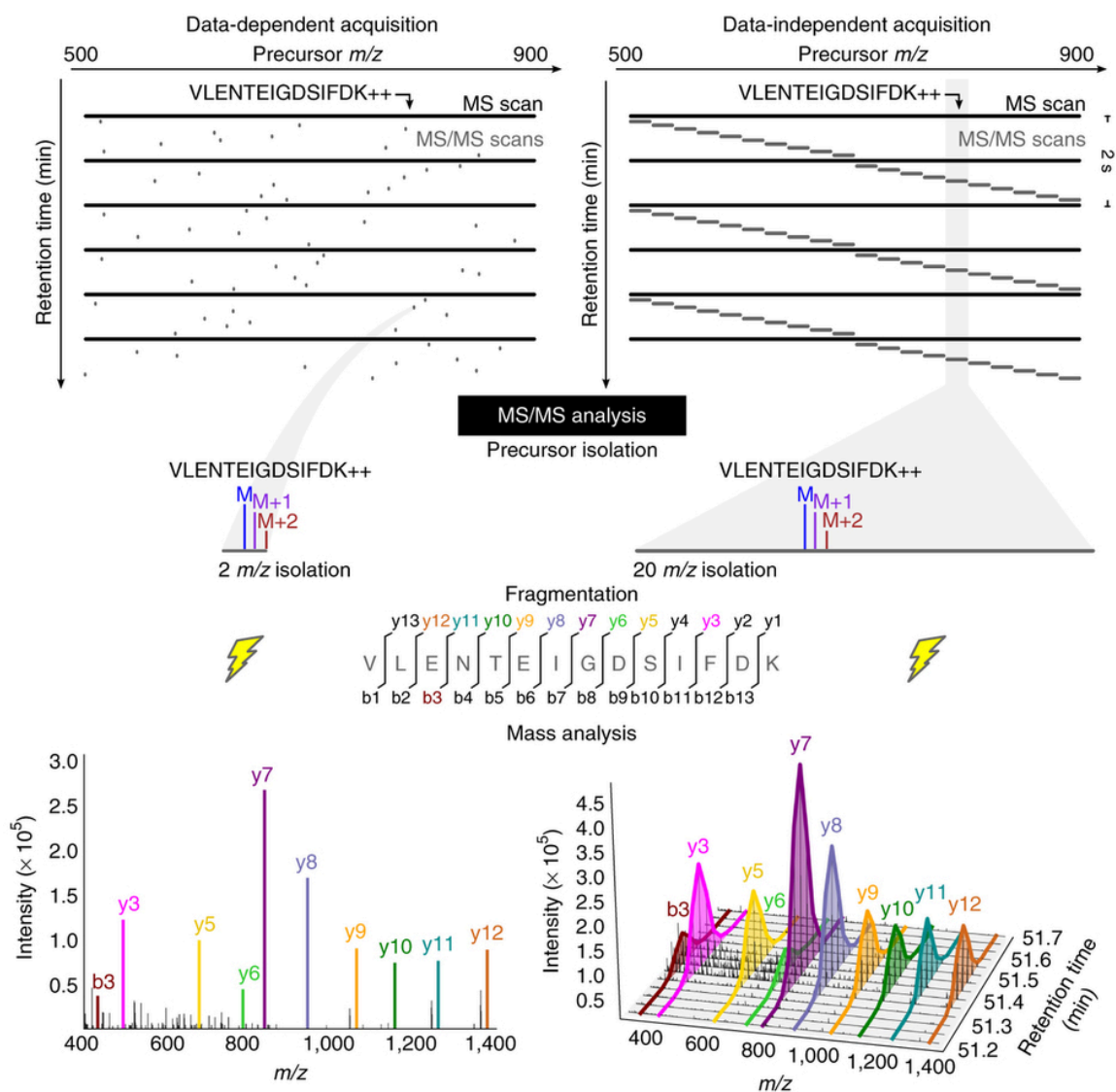


Figure 6 Data dependent acquisition (DDA) and data independent acquisition (DIA). DDA targets MS/MS scans with narrow isolation windows centered on specific precursor ions in a survey MS1 scan. DIA acquired MS/MS scans with wide isolation windows acquiring many peptide precursors. Peptide precursor VLENTEIGDSIFDK++ is present in a single MS/MS spectrum in a DDA analysis, while it can be extracted from all retention times in DIA. This image was adapted from Egertson et al [46].

For example, DIA collects MS/MS data either without precursor ion selection [47], ion mobility–collision-induced dissociation–time-of-flight mass spectrometry [48], wide isolation windows [49], narrow isolation windows combined with many injections [50] or multiplexing strategies [44]. Even though DIA theoretically may acquire more peptide information than DDA, its data complexity is much higher than DDA because the MS/MS spectrum of DIA has fragment peaks arising from a wide number of peptides. Thus data de-convolution and peptide identification for the DIA experiment is difficult to say the least.

Traditional protein reference database search, widely used for DDA data, is not effective for DIA data due to the presence of multi-peptides in the MS/MS data. There are several tools for DIA data analysis, such as Skyline [51], OpenSWATH [52] and UDMS^E [53], but DIA still has a long way to go in spite of its rapid development to be accepted by the proteomics community and employed as routinely as the DDA method [54]. Currently there are no metaproteomics publications using DIA because it makes intricate metaproteomics data even more complicated. However, DIA might enhance the depth of metaproteome experiments if the bioinformatics can de-convolute the highly complex MS/MS data.

Quantification of proteins is a major goal for metaproteomics. The number of proteins identified gets smaller and smaller following each experimental step as proteins are lost during the experimental procedures (Fig. 7).

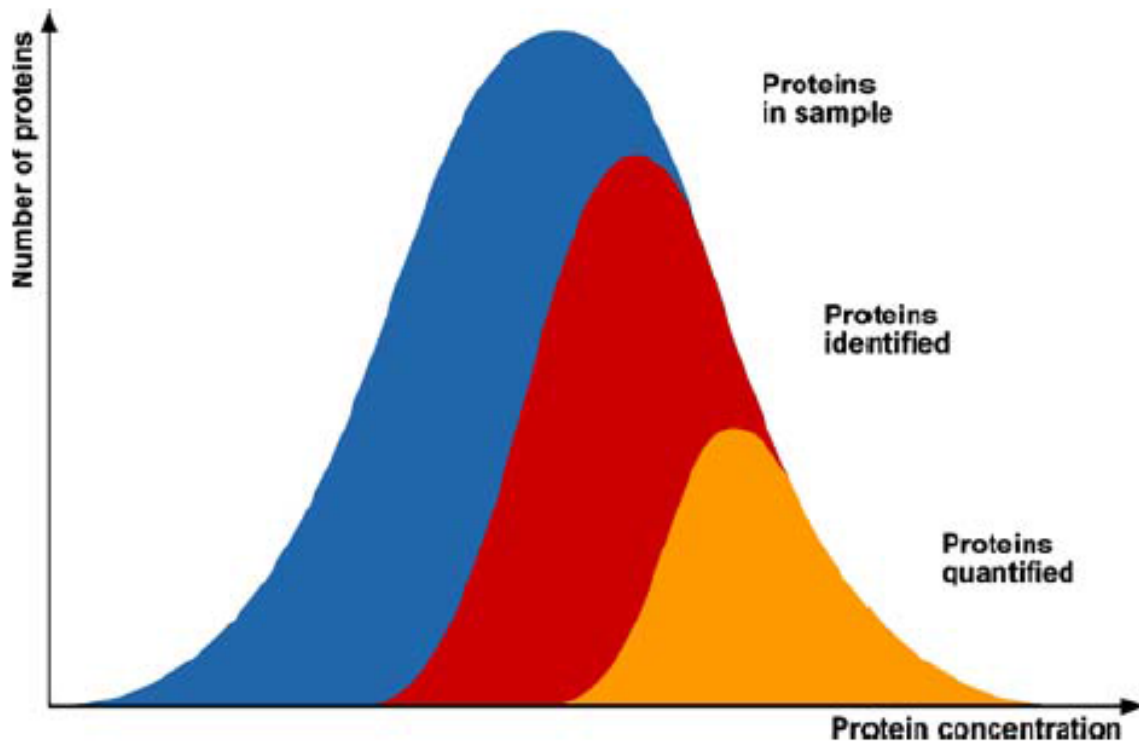


Figure 7 Relative amount of proteins in original sample, that are identified and quantified by metaproteomic analysis. Proteins are lost during protein extraction, digestion, and LC-MS/MS steps. Low-abundant proteins are difficult to identify due to limited dynamic range. Specific types of proteins can be missed by the protein extraction method. Only a fraction of the original proteins is reliably quantified. This illustration was adapted from Bantscheff et al [55].

There are two major methodologies for protein quantification in proteomics. These are termed label-free quantification and labeled quantification. Labeling methods can be categorized into metabolic labeling (SILAC [56] and isotopes), chemical labeling (ICAT [57,58], iTRAQ [58] and TMT[59]) and finally spiked isotopically labeled proteins.

The label free methodology uses intrinsic data obtained during the LC-MS/MS analyses of the sample such as counting spectra identified to each identified peptide. This is termed peptide-spectrum match (PSM) (Fig. 8). The label free methods are the simplest to employ and cheapest but provide limited quantitation accuracy.

Of the labeling methods, metabolic labeling is by far the most accurate (Table 3), but often cannot be employed in metaproteomics experiments due to difficulties in labeling an environmental sample with heavy nitrogen or carbon. Chemical labeling methods target specific residues for ICAT or the N-terminus for iTRAQ and TMT. These methods have medium accuracy and quantitative coverage (Table 3). These approaches can be used in metaproteomics as long as one is sure the environmental sample does not contain contaminants, which could interfere with the labeling process. For example, chemical labeling cannot be used in soil samples due to the presences of humic acids, which will adsorb the entire label. The spiked protein method compares the fixed amount of spiked heavy labeled internal standard proteins with environmental derived proteins. This method has low coverage and medium accuracy, but is rarely used due to cost restraints (Table 3).

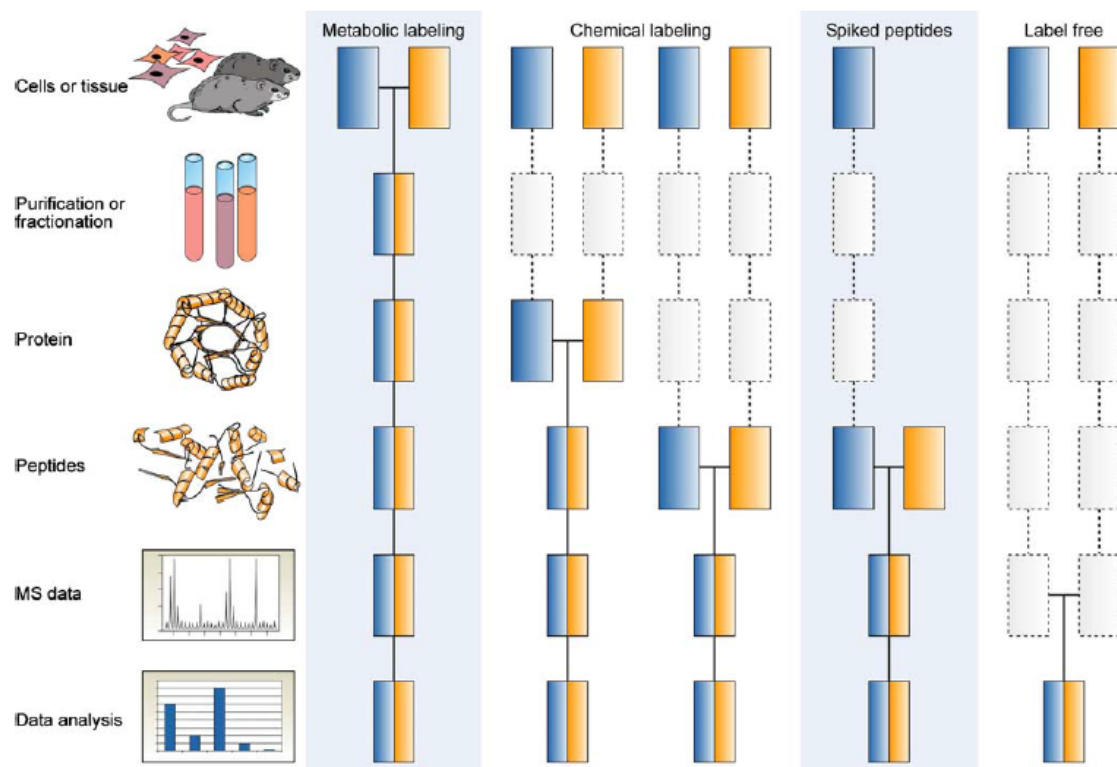


Figure 8. Common quantitative mass spectrometry workflows. Boxes in blue and yellow indicate experiments to be compared. Solid line means that labeled sample is mixed. Dashed line indicates points where experimental variation during quantification errors can occur. This illustration was adapted from Bantscheff et al [55].

Table 3 Accuracy and quantitative coverage of various labeling methods in proteomics. “Accuracy” means how a labeling method can reflect exact difference between experimental groups. “Quantitative proteome coverage” is how many proteins are quantitatively measured and identified by a labeling method. “Linear dynamic range” is the number of digits for measurement of a labeling method. This table was adapted from Bantscheff et al [55].

	<i>Accuracy (process)</i>	<i>Quantitative proteome coverage</i>	<i>Linear dynamic range</i>
Metabolic protein labeling	+++	++	1-2 logs
Chemical protein labeling (MS)	+++	++	1-2 logs
Chemical peptide labeling (MS)	++	++	2 logs
Chemical peptide labeling (MS/MS)	++	++	2 logs
Enzymatic labeling (MS)	++	++	1-2 logs
Spiked peptides	++	+	2 logs
Label free (ion intensity)	+	+++	2-3 logs
Label free (spectrum counting)	+	+++	2-3 logs

Label-free quantification has excellent coverage, but low accuracy. It is easily employed in metaproteomics and the one most widely used (Table 3). Overall, Label-free quantification is the most commonly used in metaproteomics though extensive isotopic labeling of an environmental sample has recently been demonstrated [60-63].

Bioinformatics for metaproteomics

Peptide-spectrum search and identification

Bioinformatics analysis is conducted to identify peptides from MS/MS spectra to reconstruct and quantify proteins. Bioinformatics analysis is the most challenging in all the steps of metaproteomics. The difficulty of computational analysis is due to the intricate data structure of the metaproteome and the lack of prior information about the microbial communities as well as limited or low quality reference metagenome databases.

The species co-habiting a natural community may not be fully known, which means that we cannot construct a protein database collected from reference databases (DB), such as RefSeq [64], UniProt [65] or IMG [66]. Even though metagenome sequencing can replace or enhance an isolated reference DB, the metagenome databases are often not complete, contain unclosed genomes, and many have sequencing errors and/or partial protein sequences.

There are two ways to construct protein sequence DB for metaproteomics. First, the protein DB is constructed by a collection of reference isolated microbial genome DB. Specific microbes can be expected to be in environmental samples

based on previous studies or based on assumptions. For example, we collected gut-isolated microorganisms for the human fecal protein study in Chapter 4 and the diet proteome study in Chapter 5. The NIH funded Human Microbiome Project (HMP) has comprehensive gene and protein databases from microbes isolated from the gastrointestinal tract. We obtained referenced proteins from a subset of isolates from the HMP DB that we used to construct our protein DB.

Next, metagenome sequencing provides deep genome information for the community of interest. Thus, metagenome sequence can complement proteins that were not anticipated to be in the samples. Also, metagenome sequencing can predict potential coding sequences from unknown organism in the environment of interest. However, metagenome sequences are often partial and incomplete. Mixing the reference DB sequences and the metagenome sequences is the best method for construction of a comprehensive potential protein DB for metaproteomics.

Because protein identification of the MS/MS spectrum via the database search method is based on having the correct peptide and thus the correct protein in the composite metagenome/isolate database, deficiencies in this composite database can severely hamper peptide and subsequent protein identification and quantification. If the appropriate reference protein DB is unavailable or has limitations, de novo sequencing could theoretically be an alternative way for identification of the MS/MS spectra. However, de novo sequencing based on MS/MS searches in metaproteomics has been found to be extremely time consuming with limited effectiveness. Cantarel et al [67]

compared the number of peptides identified by DB search using SEQUEST [68], and de novo sequencing using PEAKS [69] and PepNovo [70]. Interestingly, this study showed that de novo sequencing could find almost 9,000 peptides that were not detected by the DB search. At the same time, however, de novo sequencing missed more than 10,000 proteins identified by SEQUEST. Therefore, some PSM search pipeline, such as Proteome DiscovererTM (Thermo Fisher Scientific Inc.), combines DB searches and de novo sequencing to maximize metaproteome coverage.

One of the advantages of de novo sequencing is that it does not require any prior database information. The MS/MS spectrum is a collection of ions derived from the fragmentation of a precursor parent peptide ion but the observable fragmented ions will vary depending on the fragmentation method. Collision induced dissociation (CID) and higher-energy collision dissociation (HCD) produce b and y ions, while electron-transfer dissociation (ETD) generates c and z ions. For example, when the MS/MS peptides are fragmented by CID and HCD, *de novo* peptide sequencing starts from the b and y ions that differ by the molecular mass of the amino acid residues. By comparing ion mass, sequence order is determined (Fig. 9). However, *de novo* sequencing from the MS/MS spectrum is generally not successful for the following reasons: 1. Low ion intensity causes incomplete detection of the sequence ions; 2. Many peptide sequences do not provide the full complement of potential y and b ions upon fragmentation; 3. Unusual amino acid and uncommon covalent modifications can prevent correct amino acid assignment [71].

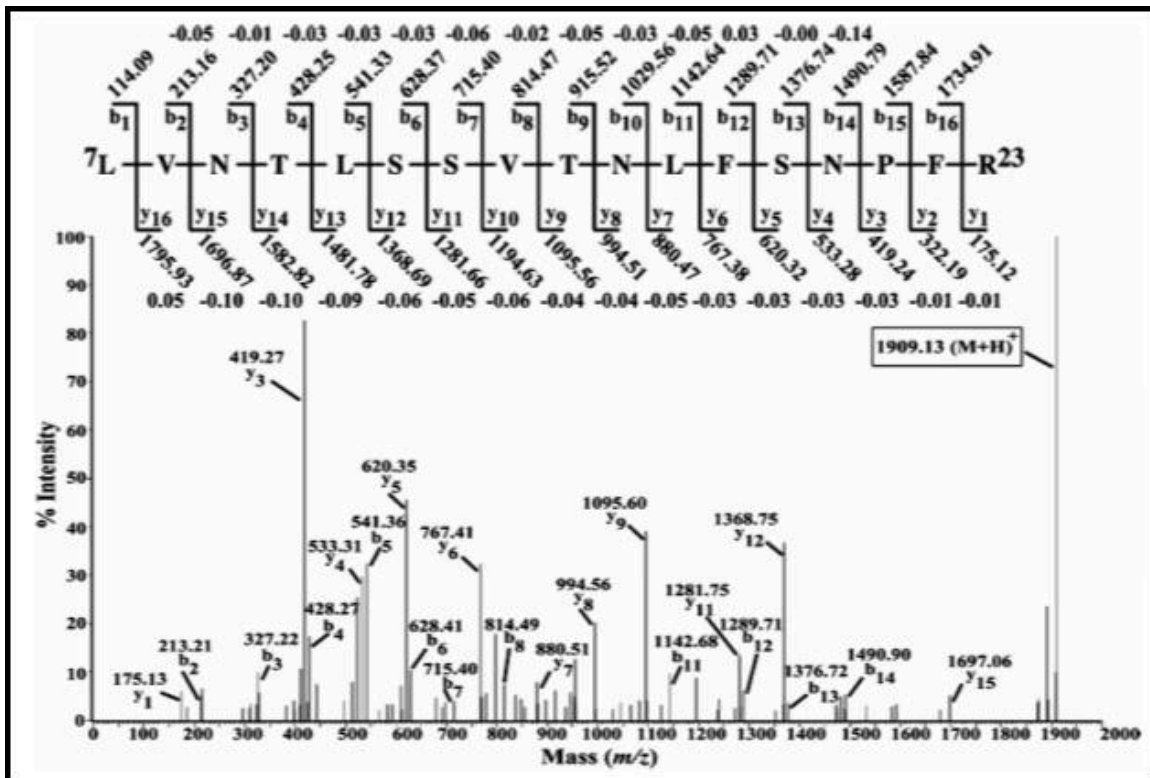


Figure 9 Example of *de novo* sequencing from MS/MS spectrum. This tandem mass spectrum of the $[M+2H]^+$ ion m/z 955.03 in LC/ESI/MS analyses of tryptic peptides from a 70 kDa iPLA2 digest. This illustration was borrowed from <http://msr.dom.wustl.edu/ms-peptide-mapping/>.

De novo sequencing is the ideal way to identify MS/MS, but is dependent on charge state of precursor ions and fragmentation efficiency, which limits its usefulness.

De novo sequencing is often used to support and improve DB-based MS/MS searches (DB search). In contrast to de novo sequencing in metagenomics, peptide de novo sequencing cannot be confirmed by overlapping sequences. Thus building the identified peptides into proteins is a serious bioinformatics challenge. Therefore, de novo sequencing methodologies are currently only used to support DB searches in metaproteomics.

The most widely used method for MS/MS identification in proteomic and metaproteomic studies is the DB search method. The MS/MS search engines, such as SEQUEST[68], MASCOT[72], X!Tandem[73], OMSSA[74], MyriMatch[75] and MS-GF+[76], use an assembled protein DB for peptide identification and subsequently build the peptides into proteins (i.e., assembly). These search methods compare the 100,000s of MS/MS spectra from the LC-MS/MS analyses to the theoretical computational derived peptides MS/MS spectra generated from the protein DB based on the enzymatic digestion method used. The major difference in these algorithms is their use of different scoring systems (Fig. 10). The first step for all search engines generate is to generate all the possible peptides from protein DB based on the experimental proteolytic enzyme used. Then, candidate peptides are selected from the peptide library using the mass of the precursor ion.

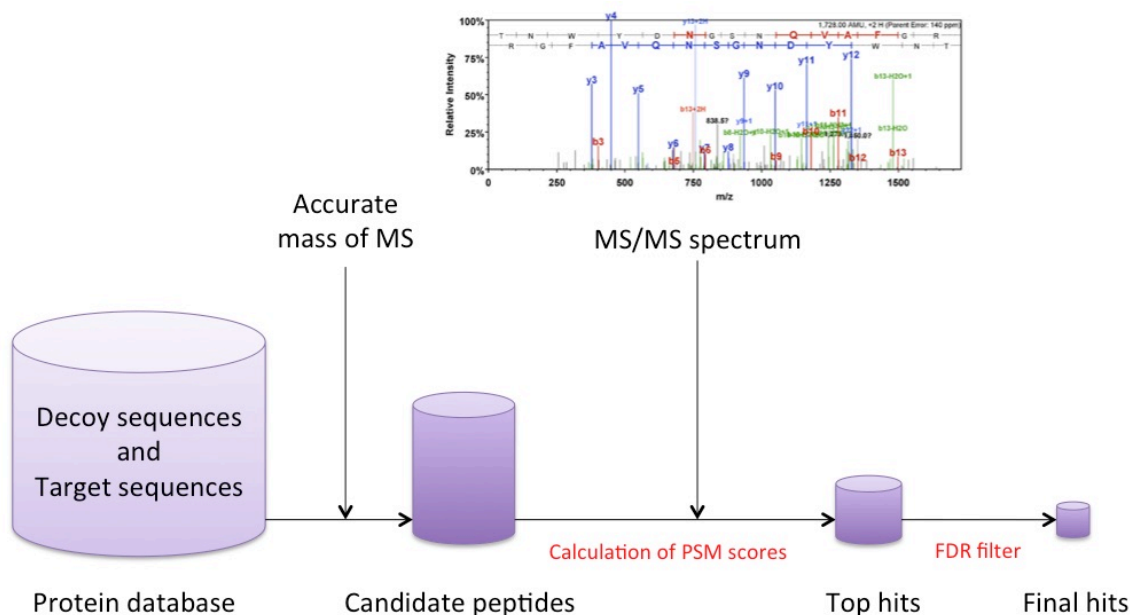


Figure 10 Schematic diagram of the DB search for MS/MS spectrum analysis. First, search engines are used to generate all the possible peptides from the protein DB while figuring in the number of allowed miss-cleavage based on the digestion enzyme. Then, candidate peptides are selected from the peptide library using the mass of the precursor ion. The next step is the calculation of scores, probability, correlation coefficient or similarity, between candidate peptides and MS/MS spectrum. Through calculation of the score matrix top hits are determined. With top hits, FDR is calculated based on decoy identification and final hits are selected.

The next step is to calculate the scores, probability, and correlation coefficient between the candidate peptides and the MS/MS spectrum. It is at this step where the different algorithms show their biggest differences between the different analysis programs.

Since SEQUEST was developed in 1994, many algorithms for MS/MS identification have been developed, such as MASCOT in 1999, X!Tandem in 2004, OMSSA in 2004, MyriMatch in 2007 and MS-GF+ in 2014. SEQUEST is the gold standard DB search engine and was commercialized and is included in the Thermo software package Proteome DiscovererTM 2.0. SEQUEST calculates cross-correlation coefficient (XCorr), similarity score between theoretically generated peaks and query MS/MS spectrum, and DeltaCn, which is the difference of XCorr from the next candidate following an initial screen using the Sp score. MASCOT is the other gold standard search engine and it is based on MOWSE algorithm, is the first probability-based scoring for mass finger printing, and provides a probability of observed matches between the queried MS/MS spectra and the peptide library. Even though MASCOT and SEQUEST were released more than 15 years ago, these are still the most widely used DB search engines in proteomics. However, the basic design of the scoring system in SEQUEST and MASCOT is for traditional CID and low-resolution MS [76].

Because SEQUEST and MASCOT are old and commercialized, there have been numerous attempts to develop more accurate, faster, and free search engines. X!Tandem, OMSSA, MyriMatch and MS-GF+ were developed to fill this need. OMSSA has a probability-based scoring system like MASCOT but

employs a Poisson distribution. OMSSA is faster and much cheaper due to its open source algorithm. X!Tandem is another engine that has a probability-based score (i.e., hyperscore) that is a combination of hypergeometric distribution and dot product. X!Tandem collects common MS/MS peaks between theoretical peaks and actual query peaks which are filtered by their expected intensities, then generates a dot product matrix. This dot product result is modified by multiplying by N factorial for the assigned b and y ions. The factorial modification is determined based on the hypergeometric distribution. X!Tandem calculates an E-value by extrapolating the linear regression line of the log histogram (i.e., the number of results) and measuring the difference between the top hit from the following hit. This E-value plays a similar role to DeltaCn of SEQUEST. MyriMatch uses multivariate hypergeometric distribution (MVH) to calculate probability of matches between query and DB by chance.

MS-GF+ has a different approach from the other DB search engines (Fig. 11). Most DB search engines compare one spectrum to all the candidate peptides filtered from the library according to their preliminary score, such as Sp of SEQUEST and the noise filter of X!Tandem. However, MS-GF+ generates a peptide dictionary of all possible *de novo* sequences from an MS/MS spectrum and searches for their presence in the peptide library [77]. MS-GF+ compares the log likelihood ratio of vectors of peptides and its spectra and calculates E-value to find the best hit. While traditional DB searches compare every spectrum against every peptide in the library, which increases the run time depending on the size of the library,, MS-GF+ execution time is not dependent on DB size.

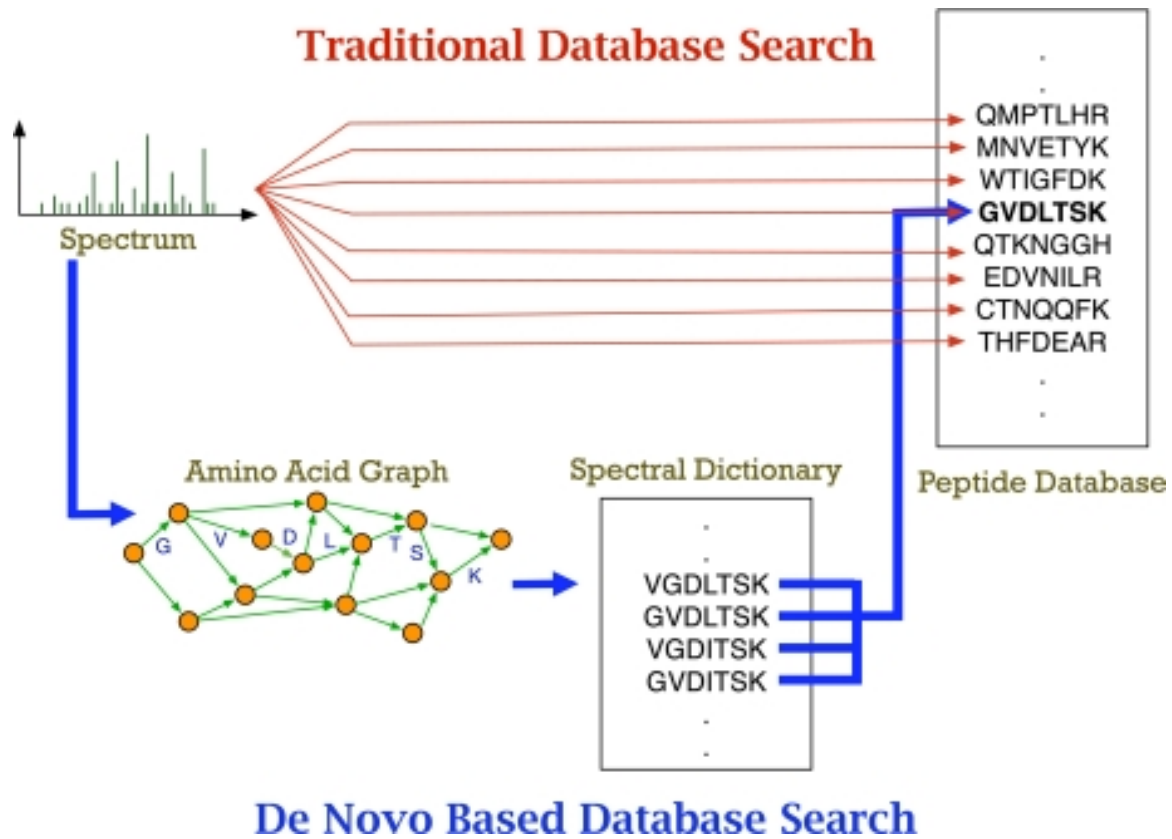


Figure 11 DB search of MS-GF+. MS GF+ is an updated version of the Spectral Dictionary [77]. Red line is a traditional DB search comparing spectra against the peptide library generated from the protein DB. Blue line is the *de novo* based DB search constructing all the possible *de novo* sequences from a MS/MS spectrum. This illustration was adapted from Kim et al [77].

This is a very important advantage for metaproteomics analysis because these experiments generally have a much larger protein DB than single species proteomics experiments. Because MS-GF+ is based on de novo sequencing and is independent of DB size, it can provide more sequences that are not found in the reference DB. This is important in the proteomics analysis because the traditional search algorithms will miss peptides from unknown species in a microbial community because they are not in the predicted protein DB,

In addition to different scoring metrics, every DB search engine has a pre-processing step prior to computing the scoring matrix. This filtering step is to remove noise peaks and to collect valid b and y ions, but can also introduce problems. If the filtering threshold is too low, noisy matches will hinder true matches; if it is too high, true fragment ions may be lost. These different pre-processing methods used by the different DB search engines can result in different peptide identifications.

Besides the accuracy of peptide identification, the speed of the MS/MS search is critical issue because the metaproteome databases tend to be very large. MS-GF+ has an advantage on this issue because its search speed is not slowed by the protein DB size. However, because the MS/MS search is parallel, which means it can be easily divided into small pieces, other MS search engines can make up for its need for searching all of the databases by simultaneously running the search on a cluster server or supercomputer. For example, one LC-MS/MS run consists of many of MS/MS spectra, so if we have 1,000,000 CPUs in a supercomputer, we can search a million MS/MS spectra at the same time.

This means that a million spectra can be processed at the same time it takes for one spectrum search. There are several ways to parallelize the MS/MS peak search engines, such as X!Tandem [78], wrapper for MS [79] and MyriMatch [75]. In contrast to the other programs, MyriMatch does not require additional installed parameters to work effectively in a parallel environment. It can utilize available cores with a simple option when executed. MASCOT and SEQUEST also has a platform for a multi-core environment, but these commercial products are very expensive to put onto large Linux clusters. Due to the very large protein DB size and the large number of MS/MS spectra along with potential PTMs, the analyses of metaproteomics datasets can be near impossible on a single CPU Windows environment. Thus a parallel computing environment is recommended for bioinformatics analysis of metaproteomics.

The target-decoy search strategy has revolutionized peptide and protein identification. Before this methodology was introduced proteomics datasets were plagued with high false positive rates putting the whole field of proteomics in a questionable light. Elias et al. suggested generating decoy databases. These databases have the same characteristics to the protein DB, but contain a competing target DB, and a decoy DB during the MS/MS search [80]. The decoy DB is generally engineered by reversing or random mixing of protein sequences of target DB (i.e., the original protein DB). By reversing sequences, decoy proteins have same characteristics as the target DB in terms of digested peptides length and monoisotopic mass (Fig. 12).

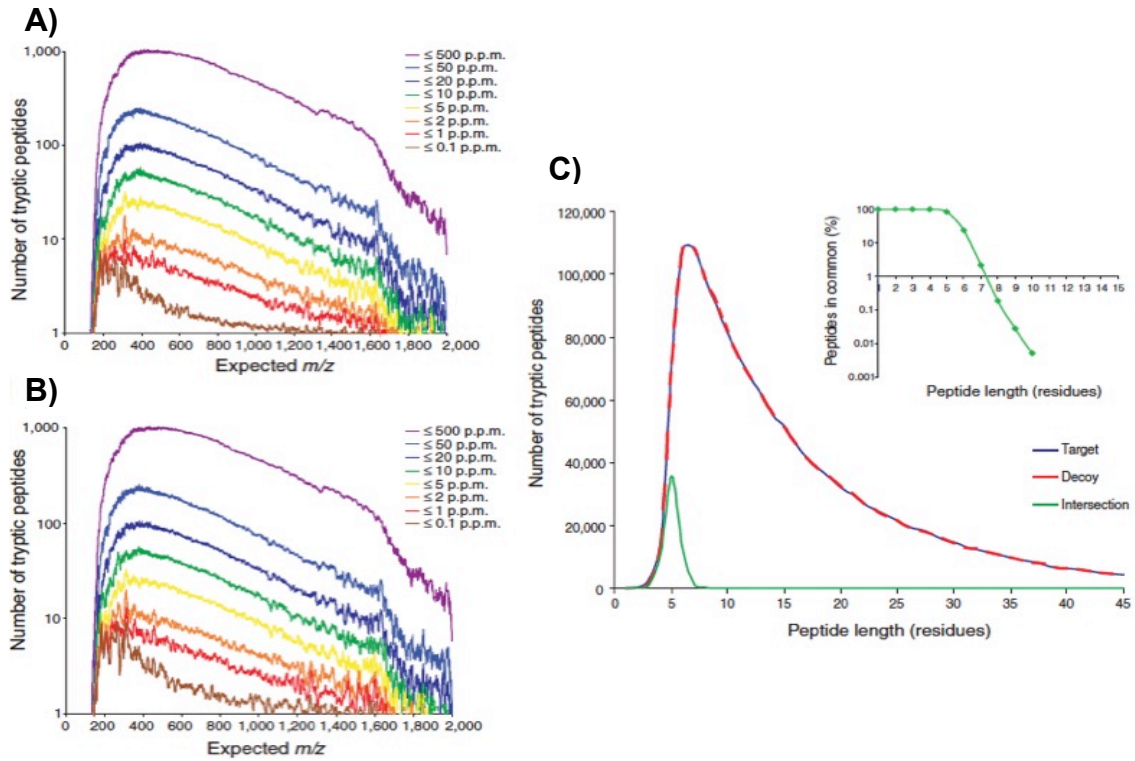


Figure 12 Target-decoy DB comparison. The number of trypsin-derived peptides plotted against the calculated monoisotopic mass of A) the target DB and B) a decoy DB from *in silico* tryptic digestion of IPI human protein DB. C) The number of trypsin-derived peptides using the target (red line) and decoy (green line) DB. The inset is a plot of the percentage of proteins that were identified in both target and decoy DB searches versus peptide length. This illustration was adapted from Elias et al [80].

Because the decoy sequences are not in the sample, if a protein is identified during a search, it is considered a false positive. The number of false positives from decoy hits can then be used to calculate the false discovery rate (FDR) using the following formula:

$$FDR = \frac{2 \times \text{Decoy hits (FP)}}{\text{Decoy hits (FP)} + \text{Target hits (TP)}}$$

Generally, a < 5% FDR threshold is considered a publishable proteomics dataset. The FDR formula is different for different analysis. Software packages such as Peptide Protein Prophet, MASCOT, and Percolator have their own modified FDR formulas. After the decoy search a FDR calculation is necessary. However, the arbitrarily generated decoy can not cover all the possible false positives, so there are concerns about using FDR as the gold standard method for MS/MS search assessment [81].

Applying FDR to metaproteomics is tricky because of the large number of non-unique peptide sequences and the amount of unknown genomic sequence in the samples. With unique peptides, FDR is often overestimated (e.g., 6-8% in Wilkins et al [82]) because of the large reduction in real peptides that are redundantly matched. With all the unique and non-unique peptides the rate is decreased to 0.6-0.9% [82] which is probably an underestimation due to the addition of many redundant peptides. An adjustment of the FDR is required when evaluating the metaproteome considering its complex characteristics.

As more and more peptide-spectrum matches (PSM) are published by proteomics laboratories world wide, a new method for searching MS/MS spectra

is becoming more popular. The Spectral Library Method has a greatly reduced DB search time. This is achieved by not using redundant searches of the same post-translational modifications (PTM) and peptides, especially from abundant common proteins found in all samples. The National Institute for Standards and Technology (NIST) and the Global Proteome Machine (GPM) provides spectral libraries for proteomics studies. X!Hunter [83], Spec2Spec [84] and Pepitome [85] use PSM searches against spectral libraries that do not need to generate theoretical MS/MS spectrum. However, a spectral library search is not an appropriate method for metaproteomics because the spectral library requires samples that have been analyzed by a large number of labs with a high number of replicates such as serum/plasma studies.

Post-process of PSM search results

After a PSM search, post-processing is required to validate the PSMs and assemble the appropriate proteins. Assessing the MS/MS identification is always important. The threshold scoring metrics that gives reliable peptide identification is ambiguous. Since the development of the first MS/MS search engines in 1994, different evaluation methods for peptide identification have been evaluated. DTASelect [86], developed in 2002 and widely used for protein assembly of SEQUEST results, suggests the following guideline for a SEQUEST search: > 1.8 XCorr for +1 charged ion, > 2.5 XCorr for +2 charged ion, > 3.5 XCorr for +3 charged ion and > 0.08 DeltCN for every charged ions. By providing filters, and simple protein assembly, DTASelect is widely used for

assembling SEQUEST search outputs. Protein Prophet [87], also developed in 2002 and incorporated into the trans proteomics pipeline (TPP), calculates probability based on a mixture-model EM method using a Bayesian approach for MS/MS identification. Protein Prophet gives statistical guidelines for assessing the correct and incorrect peptide assignments. It also estimates the likelihood of proteins corresponding to those peptides in a sample. To maximize peptide identification, Scaffold [88] employs various search engines and integrates its results from Protein Prophet. Scaffold obtains MS/MS search results from SEQUEST, MASCOT, OMSSA, X!Tandem and etc. Then, it evaluates peptide identifications of each engine using Protein Prophet. These results are aggregated and re-evaluated by Protein Prophet. By aggregating multiple evaluation methods, Scaffold provides a platform-independent result with maximized peptide and thus protein identifications.

Protein assembly is a challenging step after peptide identification of the MS/MS. Generally, assembly is more difficult than disassembly because some parts dissociate from a protein. This may or may not have arisen from that specific protein. Redundant peptides, found in two or more proteins, makes protein reconstruction and quantification difficult. Metaproteomes generally have more shared peptides due to orthologous and paralogous proteins from multiple species. Orthologous and paralogous proteins have similar protein sequences, so they generate shared or redundant peptides.

Meyer-Arendt et al. demonstrated that the human protein database has 89,486 entries in the international protein index (IPI) DB v3.75. This would have

almost 2 million shared peptides with the 4 million trypsin-derived peptides (length > 7 amino acids allowing up to 2 miss-cleavages) [89]. Moreover, a metaproteome has a number of microbial species, which means there will be a greater chance for overlapping peptides than Meyer-Arendt's human database only example. Therefore, metaproteomics datasets have many shared peptides making ambiguous protein identifications common and problematic.

How ambiguous protein identifications are dealt with has a direct impact on the final protein identification and quantification. If an identified peptide A can be generated from protein 1, 2 and 3 in the protein DB, several interpretations are possible. For example, peptide A may be derived only from 1, from 1 and 3 together or from 1, 2, and 3 at the same time. The source of peptide A is uncertain due to lack of information, but different final protein identifications are possible based on how the ambiguous protein identifications are handled. Using DTASelect, would identify all possible proteins, such as protein 1, 2 and 3 of above example. However, IDPicker [90], developed in 2007 and linked to MyriMatch, reports a minimal protein list. Protein assembly of IDPicker is based on a bipartite graph and greedy algorithm (Fig. 13). First, IDPicker connects peptides and proteins in a bipartite graph (Fig. 13A). Next, indistinguishable peptides and proteins are collapsed (Fig. 13B). The collapsed bipartite graph is deconvoluted into separate protein clusters with connected peptides (Fig. 13C). Proteins that have uniquely matched peptides or supported by more than two different peptide groups are identified and added to the minimal protein list (Fig. 13D). IsoformResolver [89] accesses shared peptides differently.

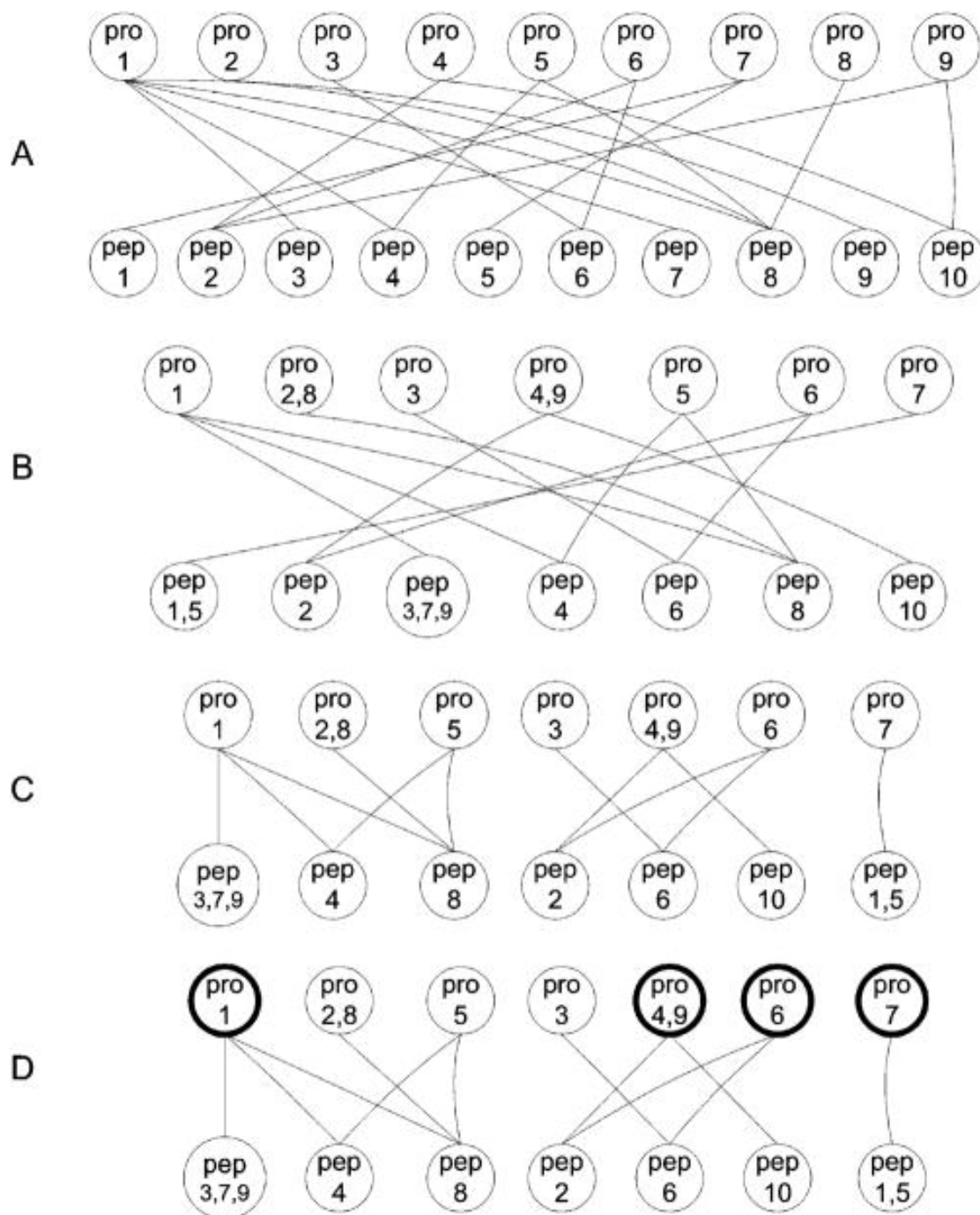


Figure 13 Overview of bipartite graph model and analysis procedure of protein assembly in IDPicker. Pro: protein, pep: peptide. This figure was adapted from Zhang et al [90].

It uses a peptide-centric approach. This approach is similar to the algorithm of MS-GF+ used for generating all possible de novo sequences. IsoformResolver constructs unique peptide libraries from the protein DB and assigns proteins to the corresponding peptides. IsoformResolver keeps shared peptides as unique entries. However, it cannot provide any insight on how to interpret the problematic peptides. Percolator [91] employs semi-supervised machine learning to discriminate correct from incorrect PSMs, and calculates an accurate FDR and posterior error probabilities.

Protein quantification varies based on the handling of ambiguous protein identifications. Since shotgun metaproteomics explores peptides, not proteins, the quantitative measurement of peptides should be transferred to the corresponding proteins. Processing of unique peptides, matched solely to one protein, makes it easy to assign the matched protein. However, assigning shared peptides to candidate proteins is ambiguous due to the uncertainty of its origin. DTASelect redundantly assigns SC of overlapping peptides to every candidate protein. This leads DTASelect to amplify the expression level according to the numbers of ambiguous peptides in the DB search results. IDPicker employs a greedy algorithm following a winner takes all approach. After protein and peptide groups are separated, IDPicker prioritizes proteins by whether a unique peptide is linked or not and the number of SC from linked unique peptides. Greedy algorithm iteratively chooses proteins by order of priority and gives all the shared peptides to the winner. IDPicker generates parsimonious

protein identification and amplifies their expression level by casting all the available votes.

In label-free shotgun proteomics, SC of redundant peptides are sometimes distributed to the corresponding proteins (i.e., spectral balancing) [92,93]. Spectral balancing considers only proteins that have more than one unique peptide (i.e., unique proteins). SC of a redundant peptide is divided into unique proteins by calculating the ratio of their unique peptide SC. Divided SC of a shared peptide for its member unique protein is calculated as described below:

$$Divided\ SC = \frac{uSC}{\sum_{i=1}^n uSC_i} \times sSC$$

Where uSC is SC of a unique peptide, sSC is SC of the shared peptide and n is the number of unique proteins linked to the shared peptide. This spectral balancing can be connected to the normalized spectral abundance factor (NSAF) if desired [92-94].

If we have more than two MS runs, technical replicates and/or biological replicates, as well as different biological or ecological samples, we need to normalize the total number of spectra to quantitatively compare across samples and LC-MS/MS runs. There are systematic errors including instrumental error and human error even though MS runs were carried out by the same experimenter using the same instrument for the same sample. The goal of normalization is to adjust and minimize these errors. Generally, normalization methods are designed for a label-free experiment. NSAF is the most widely used normalization method. It considers spectral balancing as well as protein length

effect on probability of peptide assignment. NSAF assumes that long proteins have a better chance of having identified peptides associated with it than short proteins. Distributed NSAF (dNSAF) for i-th protein is calculated as described below:

$$dNSAF_i = \frac{dSAF_i}{\sum_{i=1}^N dSAF_i} \text{ where, } dSAF_i = \frac{uSC_i + \frac{uSC_i}{\sum_{m=1}^M uSC_m} \times sSC_i}{Length_i}$$

Where, N = the number of total identified proteins, uSC is SC of a unique peptide, sSC is SC of a shared peptide, Length is the protein length and M is the number of unique proteins. The sum of dNSAF within an MS run is 1, so the dNSAF number is very small. This small number of dNSAF is often logged to make it easy to analyze.

In addition to NSAF, there are a few normalization methods including total PSM normalization, exponentially modified protein abundance index (emPAI), and quantile normalization. Normalization by total sum of PSM is also widely used in label free metaproteomics studies [95] and is calculated as described below:

$$Normalized\ SC_i = \frac{\frac{\sum_{j=1}^M \sum_{k=1}^N SC_k}{M}}{\sum_{k=1}^N SC_k} \times SC_i$$

Where, N = the number proteins and M = the number MS runs. The total sum normalization assumes that all the MS runs should have same amount of PSMs theoretically. This makes MS runs have same total PSM. However, the total sum normalization tends to amplify a small sum of PSM too much. A small sum

of PSM runs generally has a small number of proteins, so the total sum normalization method can make false positives by distorting the data. To complement this disadvantage of total sum normalization, average normalization can be used as described below:

$$Normalized\ SC_i = \frac{\frac{\sum_{j=1}^M \frac{\sum_{k=1}^N SC_k}{N}}{M}}{\frac{\sum_{k=1}^N SC_k}{N}} \times SC_i$$

Where, N = the number proteins and M = the number MS runs.

Normalization using emPAI method is used in MASCOT searches [96]. emPAI is also designed for label-free proteomics and considers the number of experimentally observed peptides and observable peptides calculated as described below:

$$emPAI = 10^{\frac{N_{observed}}{N_{observable}} - 1}$$

Where, $N_{observed}$ = the number of experimentally observed peptides and $N_{observable}$ = the number of theoretically observable peptides. $N_{observable}$ is calculated by *in silico* protein digestion considering the digestion enzyme, allowed miss-cleavages, and the mass range of the instrument. Even though emPAI normalize SC based on the potential number of peptide identifications, it does not correct systemic errors between different MS runs.

Quantile normalization is a technique for making two distributions identical in rank statistics. Quantile normalization considers the rank of SC within an MS run and makes the same rank proteins have the same SC across MS runs.

Since quantile normalization makes MS runs and SC have the same rank, it can make the data to lose variation within replicates, which is important to ANOVA.

Statistical analysis in proteomics is not really any different from other large-scale biology data analysis. Student's t test, ANOVA, and quasi-Poisson modeling are used for statistical comparisons between different experimental groups. Student's t test and ANOVA requires a normal distribution assumption, but typical SC distributions from MS runs do not follow a normal distribution (Fig. 14). In Figure 14, the empirical estimation (red line) is not similar to normal-fitted line (blue line). The left half of normal line is missing because negative values are not in SC. Therefore, other distributions besides a normal distribution are required for proteomics data analysis. However, the different distributions are not tested in proteomics data analysis. Recently Li et al [97] tested the quasi likelihood Poisson modeling (quasi-Poisson). They compared the quasi-Poisson model to many methods including Fisher's exact test, Wilcoxon rank test, and Student's t test for processing LC-MS/MS data. They concluded that quasi-Poisson could be applied to proteomics data like other methods.

Because Student's t test is not appropriate for multiple comparisons, ANOVA is the analysis of choice for large biology data sets. Whenever a t test is conducted, there is a chance of Type I error, which is typically 5%. If we a run t test three times, the potential for a Type I error increases up to 15% [98]. If we conducted t test over 1,000 proteins collected from LC-MS/MS runs, the accumulated Type I error approaches 100%, which makes this analysis unreliable because it contains too many false positives.

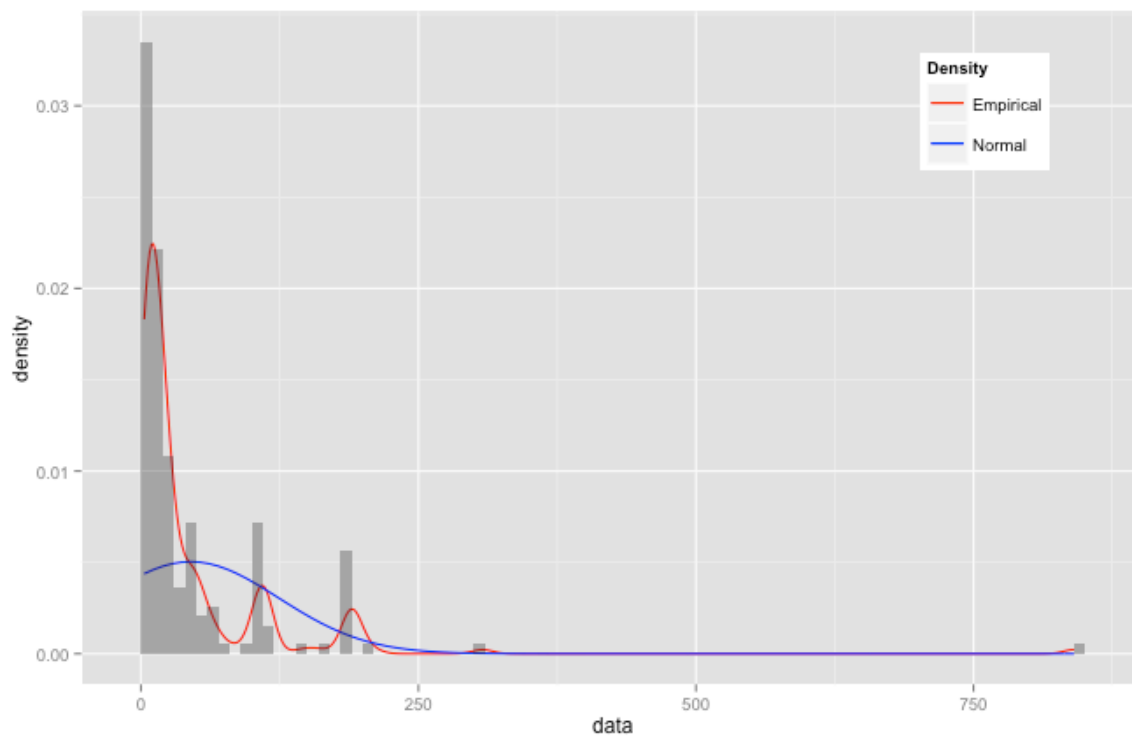


Figure 14 Comparison of the distributions fitted to spectral counts in label-free metaproteomics data. Blue line is a normal distribution fitted to spectral counts and the red line is an empirically fitted line.

However, ANOVA can limit Type I errors even when it is repeatedly applied to a data set. The difference between ANOVA and a t test is based on what these analysis tools test. Student's t test compares the mean of each sample, while ANOVA analyzes the variance within groups and between groups. Therefore, t test is not the appropriate statistical method for proteomics as well as other large biology data analysis.

Quantitative and functional analysis pipeline for metaproteomics

Even though there is a growing sense that metaproteomics studies are important, very few bioinformatic analysis pipelines are publicly available. GalaxyP (<https://usegalaxyp.org/>) is a proteomics application of the Galaxy platform [99]. Galaxy platform is a scientific workflow and data analysis pipeline for computational biology for wet-lab biologists who have limited programming skills. Galaxy is written in Python and JavaScript and provides a simple three-column interface (Fig. 15). Galaxy is easy to install and free so any user can use it. Adding bioinformatics tools to Galaxy is possible. However many Galaxy servers already have many publicly available tools (<https://wiki.galaxyproject.org/PublicGalaxyServers>). Current GalaxyP provides X!Tandem and Protein Prophet as well as data management. However, GalaxyP may not be appropriate for metaproteomics because X!Tandem is not run in a parallel environment and Protein Prophet is not the best tool for handling redundant peptides. Also GalaxyP does not have functional analysis modules, which is limiting.



Figure 15 The interface of GalaxyP. Its three-column interface consists of available tools, main screen for analysis for results and parameter input and status of analysis steps from left to right. This illustration is screen-captured from <http://www.usegalaxyp.org>.

MetaProteomeAnalyzer (MPA) [100] was recently published and is designed for metaproteomics (Fig. 16). While GalaxyP has a web-interface and server-oriented pipeline, MPA is simply a client-server communication system. In MPA the client-side uploads MS/MS experimental data and the server-side deposits and analyzes the uploaded data. MPA focuses on data interpretation rather than parallelized MS/MS searches. MPA employs four DB search engines (X!Tandem, OMSSA, Crux [101] and InsPecT [102]) and aggregates four DB searches to increase PSMs. MPA groups proteins into meta-proteins due to redundant peptides. Based on meta-proteins, MPA provides functional analysis using the Kyoto encyclopedia of genes and genomes (KEGG) pathways [103], enzyme commission (EC) [104] and National Center for Biotechnology Information (NCBI) Taxonomy [105]. MPA generates visualized reports drawing figures and produces analysis reports. MPA is the first analysis pipeline, equipped with a visual interface, but it does not provide multiprocessing and sequence-based functional analysis. MPA's functional analysis is based on the reference DB at the current time. However, metagenome sequences, which are used for metaproteome DB construction, cannot be annotated and many of microbes within the reference DB are not well annotated. Therefore, MPA's functional analysis is very limited in its current form.

An analysis pipeline for metaproteomics should have two important requirements. First, simultaneously running MS/MS search engine is necessary (i.e., multi-processing on multiple computer cores).

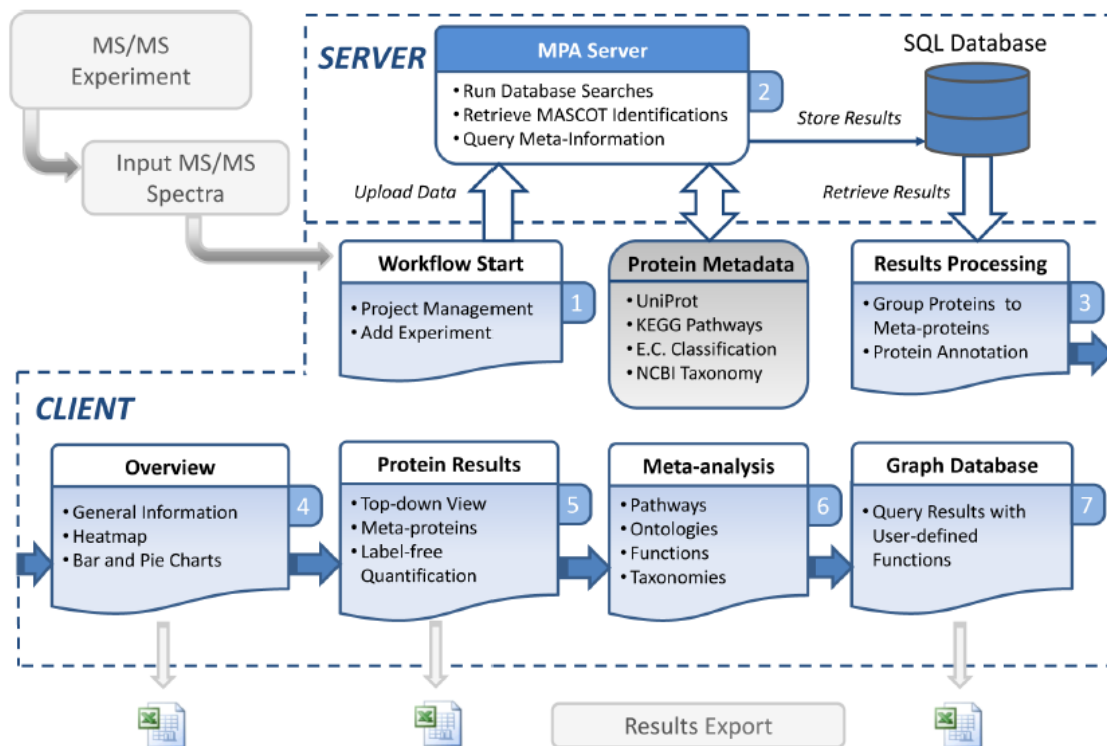


Figure 16 A schematic diagram of workflow in MetaProteomeAnalyzer.

Server side and client side communicate and the uploaded data is deposited in server SQL DB. The illustration is adapted from Muth et al [100].

Since metaproteomics experiments generally have a highly complicated protein sequence DB, DB search take much longer than single species proteomics. Even though the search speed of MyriMatch is also dependent on protein DB size, it is designed for multi-threading on computer clusters. Therefore, it can use multiple CPUs simultaneously for DB searches. Secondly, an analysis pipeline for metaproteomics should provide functional analysis based on the protein sequences, not based on the reference DB only. Human, mouse, *Escherichia coli* and yeast DB are manually curated with high quality annotation. However, most microbes in an environmental sample are unknown and do not have basic annotation. Functional analysis based on the reference DB that is used in MPA is not very valuable in metaproteomics analysis. Therefore, a metaproteomics pipeline should have at least these two required features.

CHAPTER II

QUANTITATIVE AND FUNCTIONAL ANALYSIS PIPELINE FOR METAPROTEOMICS DATA

Quantitative and functional analysis for metaproteomics (QFAM)

Objectives of QFAM

Quantitative and functional analysis pipeline for metaproteomics (QFAM) is an analysis pipeline for the metaproteome, which is highly complicated and large data. The goal of QFAM is to offer fast and reliable functional analysis of metaproteomics data. Since the metaproteome consists of a large proteomics database, QFAM enables multiprocessing where many jobs can be processed at the same time. Also, QFAM provides sequence-based functional analysis because the metaproteome cannot distinguish proteins with similar sequences and less-annotated microbial proteins. Lastly, QFAM produces reliable statistics including a significance test, an enrichment test, and the effect sizes in order to reduce potential false positives in the analysis.

General information of QFAM

QFAM is installed on the Newton cluster at the University of Tennessee-Knoxville. The Newton cluster has 4,200 processor cores (400 nodes). All nodes are connected via 1Gb Ethernet and Infiniband interconnected (up to 40Gb/sec). The operating system is Scientific Linux 6.5, which is Redhat Enterprise compatible, and sun grid engine (SGE), also known as N1 grid

engine, controls batch-queue system. Newton is designed for scientific high performance computing and is freely available to students and faculty of the University of Tennessee. Generally Newton allows only a 24 hour run for single job, but QFAM is allowed to use 48 cores for unlimited time because the PSM search of the metaproteomics analysis sometimes takes more than two weeks in spite of using 48 CPUs, which is the maximum number allowed by the server administrators.

QFAM is designed for label-free shotgun metaproteomics data obtained from microbial communities in environmental samples. The pipeline consists of five modules: PSM search module, contrast file module, normalization module, statistical analysis module, and functional analysis module (Fig. 17). The PSM search module employs MyriMatch and IDPicker for fast and exact peptide and protein identifications. The PSM search module produces tab-delimited files containing protein ID, spectrum counts, and decoy-target information. If running multiple MS runs, spectrum counts of multiple MS runs are summarized into one contrast file that has protein IDs in rows and spectrum counts in columns. This output file is tab-delimited and can be opened in excel. The module that makes the contrast output file also compares each of MS run by calculating the Pearson coefficient correlation (PCC) for technical replicate. If the experimental datasets has many technical replicates and thus a need to pick high quality technical runs, the algorithm preferentially selects MS runs that have high PCC.

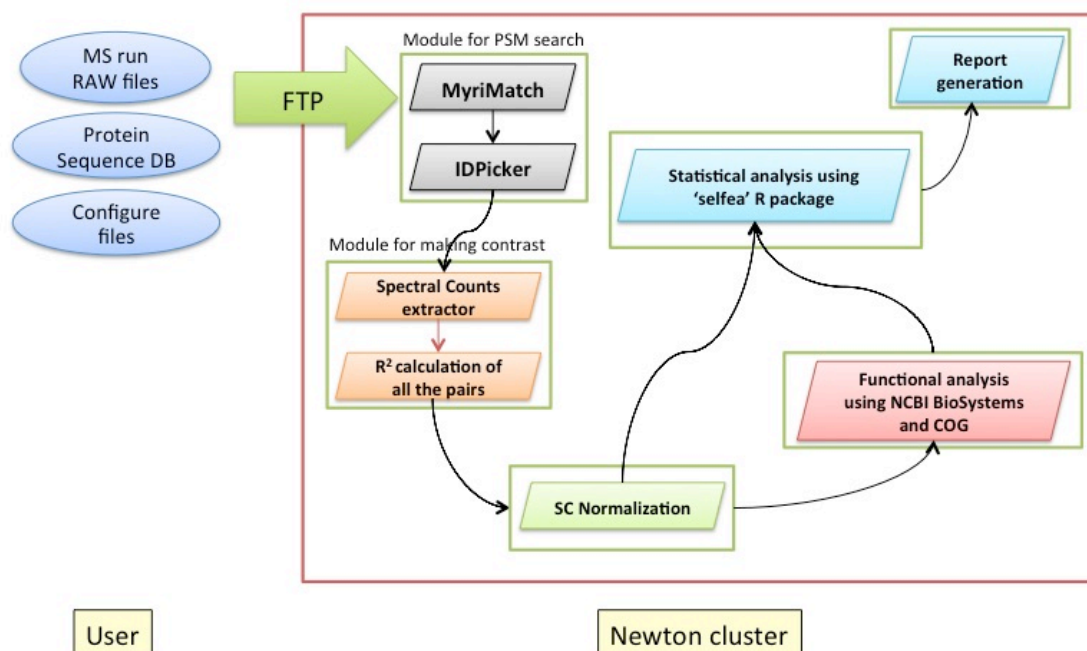


Figure 17 A schematic diagram of QFAM. QFAM consists of seven modules and is designed to process metaproteome data rapidly to give functional reports to users.

This repeated selection process could improve statistical analysis using ANOVA. The next module is for normalization. The module provides three options: total sum normalization, same average normalization, and NSAF. The statistical analysis module is programmed by an R statistics script to apply the three methodologies: quasi-Poisson model with Cohen's f , negative binomial model with Cohen's w , and Normal ANOVA with Cohen's f . The next module applies functional analysis based on the NCBI Biosystems DB [106] and cluster of genes (COG) terms [107]. The last module draws charts/graphs for publication. This module draws bar and line graphs using the 'ggplot2' R package [108].

QFAM is distinguished from other proteomics pipelines, such as MPA and GalaxyP, by its PSM search speed, quantitative analysis module (QAM) and functional analysis module (FAM). QFAM controls the PSM search module using MyriMatch and IDPicker, which are able to do multithreading processing, which is absolutely necessary for large databases common to metaproteomics studies. QAM of QFAM uses the 'selfea' R package for discerning false positives efficiently. Selfea tests the significance of FAM using NCBI BioSystems DB and COG terms. FAM provides a COG enrichment and significance test, which are not incorporated in other pipelines.

Module for LC-MS/MS data analysis

Automated module for DB search and protein assembly

PSM search speed is dependent on protein DB size because each PSM comparison checks the protein DB and this process is repeated for each MS/MS

spectra in a run. Since metaproteomics has large protein DB (> 100 megabytes), the PSM search speed is much slower than single species proteomics studies. The slow search speed increases the total time of the PSM search. This time-consuming process is the first bottleneck of the analysis pipeline for metaproteomics. To improve search speed and the overall process timing, a multiprocessing technique is needed. Multiprocessing executes multiple processes within a single process, which means running multiple jobs simultaneously using multiple CPU resources. Because the PSM search is an independent process, it is simple to parallelize. This means that it is possible to run each PSM searches individually on multiple CPUs. Multiprocessing is applied using MyriMatch for QFAM.

MyriMatch and IDPicker v3.0 are incorporated to QFAM for DB searches, post-processing, and protein assembly. Gross et al [109] showed that using MyriMatch can efficiently reduce run times (Table 4). We also tested MyriMatch in our Newton cluster where QFAM is installed and could confirm that MyriMatch can reduce PSM search times by using more CPUs (Table 5).

Table 4 MyriMatch search time. This table was adapted from Gross et al [109]

<i>Protein Database</i>	<i>1 CPU</i>	<i>2 CPU</i>	<i>20 CPU</i>	<i>40 CPU</i>
SwissProt Human (20,348 proteins)	02:24:54	01:30:23	00:07:28	00:03:57
IPI Human (87,925 proteins)	06:51:21	04:31:19	00:22:15	00:11:06

Table 5 MyriMatch run time on the Newton cluster. We tested MyriMatch using a 500-megabyte protein DB containing 176,620 proteins to search 41,079 peaks. Increasing the number of CPU cores dramatically decreases running time.

	<i>1 CPU</i>	<i>2 CPUs</i>	<i>12 CPUs</i>	<i>48 CPUs</i>
Run time for single RAW file	49:03:45	40:53:07	08:27:34	02:51:06

Since the PSM search process is the major bottleneck of LC-MS/MS data analysis, saving significant time in the search process is particularly important for metaproteomics studies. IDPicker has three sub-modules: IDPQonvert, IDPAssemble and IDPQuery. IDPQonvert creates idpDB from pepXML or mzIdentML files produced by MyriMatch as PSM search results. IDPAssemble analyzes peptide-protein matches by using bipartite graphs and calculates spectrum counts following greedy algorithm. IDQuery is a data-retrieving tool from the idpDB processed by IDPAssemble. Total execution with five RAW data runs took almost ten and half hours when tested on a 500 megabytes protein DB (Table 6).

Table 6 Run time of LC-MS/MS data analysis in the pipeline.

	<i>MyriMatch</i>	<i>IDPQonvert</i>	<i>IDPAssemble</i>	<i>IDPQuery</i>	<i>Total</i>
Run time	10:24:55	0:01:07	0:00:31	0:00:15	10:26:48

DTASelect generally takes more than a week to assemble peptides and calculate spectrum counts, while IDPicker only took two minutes to do the same job (Table 6). This enhanced speed is essential for processing large and complicated metaproteome data. IDPicker's MaxFDR is 0.05 as the default setting. QFAM reports peptides and proteins at a 0.05 FDR level.

This module for MyriMatch and IDPicker is fully automated. Only one job submission to Newton is needed. This job controller is programmed in Python 2.7.6 and requires one configure file named as "BMI_pipeline.cfg". This configure file contains information on the data file path, program (MyriMatch and IDPicker) path, the number of CPUs for pipeline execution, and a job list file location. The PSM search module produces a tab-delimited file that has the protein ID, spectrum counts from unique peptides, the number of unique peptides, DB source (i.e., target or decoy), protein length, percentage of protein sequence coverage, and spectrum counts.

Making contrast file, comparing MS runs, and normalization

Every metaproteomics experiment has more than two MS runs for comparison unless it is a test run. Multiple MS runs are included in the experimental design for comparisons of the treatment and control groups. Multiple MS runs and technical or biological replicates are needed for each experiment. Multiple MS runs are summarized into one data file (i.e., contrast file), usually in a tab- or comma-delimited file, and contain the MS run names in

columns and protein IDs in rows. QFAM provides a module for making a contrast file from multiple PSM search results. Also, QFAM reports the protein FDR, the number of decoy protein identifications divided by the total number of proteins, and removes decoy protein identifications.

In addition to making a contrast file, QFAM calculates the Pearson correlation coefficient (PCC) of every possible pair in one contrast file. By comparing PCC within replicates and between different groups, users can choose the more similar replicates and different MS runs from other experimental groups. This module is necessary for the comparison of technical replicates and biological replicates in Chapter 4. The PCC comparison module is programmed in R statistical programming script. Also the PCC comparison can group MS runs that are very similar ($PCC \geq 0.90$). Generally technical replicates have much better PCC than biological replicates and other experimental groups. Picking MS runs for further analysis is not automated and requires user input.

The next module is normalization followed by contrast file creation and picking MS runs. Normalization is necessary step for label-free proteomics data to standardize different MS runs for comparison. Piehowski et al examined source of errors during proteomics data generation [25]. They considered four variance sources, protein extraction, digestion, instrument stability and instrument variance (Fig. 18). Dissecting different parts from the same sample made different protein extraction replicates. Protein digestion, instrument stability and instrument variance were measured by repeating experiments at different stages.

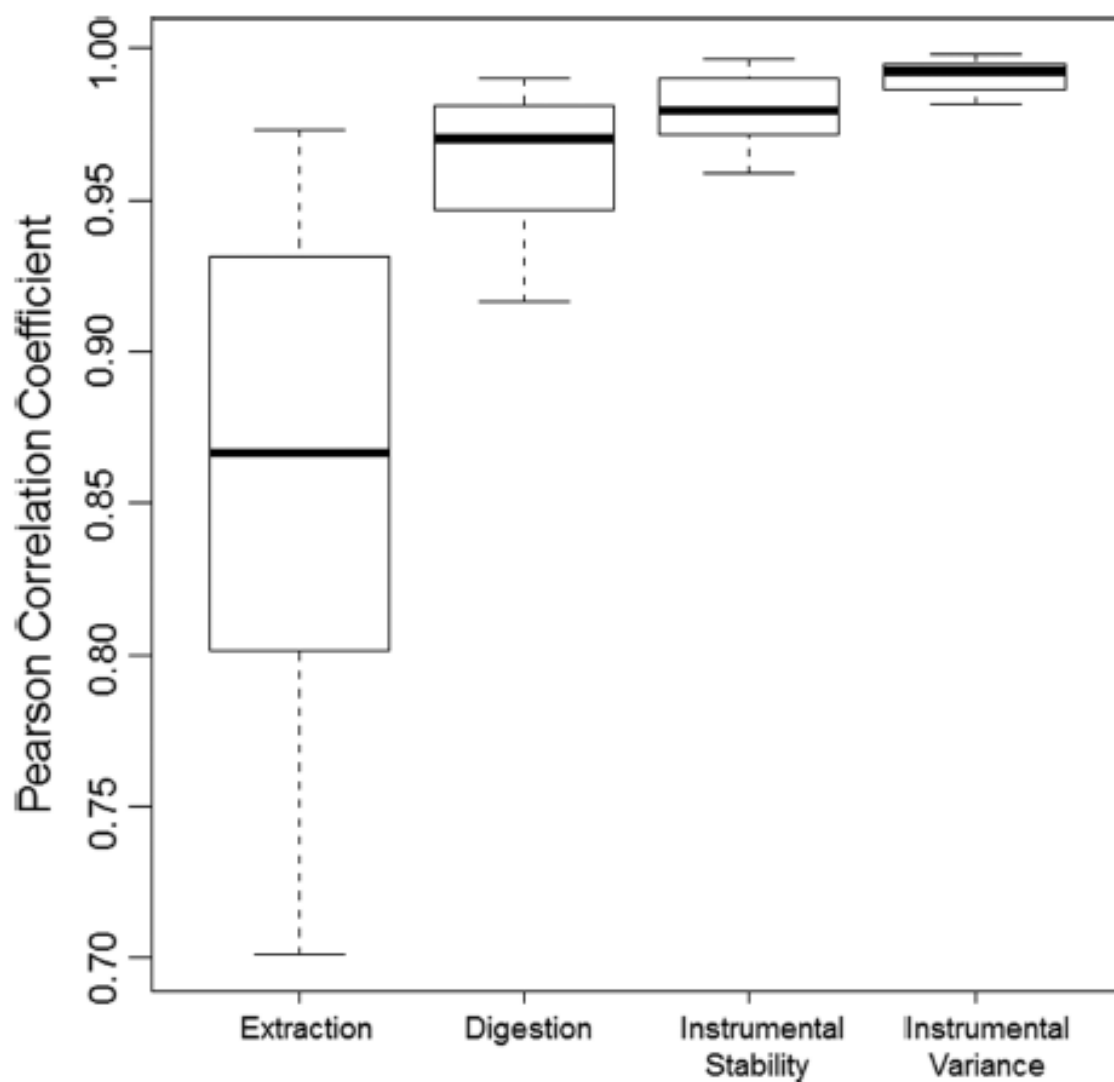


Figure 18 Boxplots showing the distribution of Pearson correlation coefficients among the different of replicates during the different stages of the proteomics analysis. This illustration was adapted from Piehowski et al [25].

Interestingly, MS instrument related variances were negligible when compared to protein extraction and digestion (i.e., average PCC: 0.99 for instrument variance and 0.97 for instrument stability). Despite of using the same patient's sample, protein extraction variability was much higher than others (i.e., average PCC: 0.86). Normalization is used mainly for sample variation rather than instrument variability. Our module for normalization provides total sum normalization and average normalization. Since QFAM employs IDPicker, which uses greedy algorithms for assigning redundant peptide matches, spectral balancing is not included in the pipeline as IDPicker already accomplishes it.

Even though normalization is employed to correct potential systematic errors, sometimes normalization can make serious changes in a dataset. This means that normalization can make unwanted differences between experimental groups. When we standardize our data, we should be aware of this potential data distortion. Interestingly, normalization can have a huge impact on data analysis, but only a few papers discuss data normalization in proteomics. As with any step of computational analyses the user should check the results to be sure that obvious errors have not been introduced into the dataset by an automated bioinformatics process.

Module for quantitative and functional analysis

Quantitative analysis module (QAM)

QFAM's quantitative analysis module (QAM) is a major improvement compared to other proteomics analysis pipelines, in our opinion. QFAM uses

'selfea' R package as the core statistics calculator. Selfea will be explained in Chapter 3 in greater detail. The analysis module has several R and Python scripts to execute 'selfea' simultaneously using multiple cores. As with any analysis on a metaproteome, the running time for the statistics computation takes a relatively long time. It took more than 12 hours to calculate the statistics of 55k proteins with single core. However 1 hour was enough to finish the computation via the parallel-running module.

The QAM has four steps. First, the normalized expression profile is divided into small pieces. Users can decide the size of data pieces. Python script for the fragmentation produces job files for parallel execution as well as reduced data chunks. Job files have information for module loading, change directory and R script execution. R script is prepared for statistical calculation using 'selfea' package. Next, job files are submitted to SGE of the Newton cluster. All the submitted jobs are executed on computing cores that is dedicated only for computation, not for systems administration or control. After all the jobs are completed, a Python script for gathering the results of the executed jobs is carried out. All the separately calculated statistics results are collected and compiled. P-values, calculated from three models (i.e., quasi-Poisson model, negative binomial model and Normal ANOVA model) are adjusted by user-selected methods including Benjamini and Hochberg correction (FDR) [110]. At the same time, insignificant proteins are filtered out and a protein scatter plot drawn.

QFAM provides six statistics for every protein: Cohen's w , Cohen's f , maximum fold change and adjusted p-values from quasi-Poisson model, negative binomial model, and normal ANOVA model (Table 7). Cohen's w and f measure effect sizes (ES) by comparing how many of the detected proteins exist in the different experimental groups. The measurement of the ES is important to maintain the statistical power in the resultant signature. Maximum fold change is the biggest fold change among all the possible pairs in the experimental groups. Selfea provides three probability models as well as various ES calculations. Li et al [97] use the quasi-Poisson model for label-free proteomics data analysis. The negative binomial model is similar to quasi-Poisson and is used for analysis of count-type data, such as spectrum counts of label-free proteomics and the number of reads in RNA sequencing (RNA-Seq) data.

Selfea calculates the ES threshold according to the statistical power where the default is 90% ~ 0.90, but the user can change this. By combining selected methods, such as quasi-Poisson and Cohen's f , significant proteins are selected at the designated statistical power. Statistical power is the probability correctly of rejecting the null hypothesis when it is false. Hence, it is directly related to reliable protein selection. Selfea enhances the statistical power by removing underpowered hypotheses tests using Cohen's ES. The higher the Cohen's ES is, the stronger the statistical power, which generates higher quality data. Making a higher statistical power can reduce the number of false positives in the final output dataset.

Table 7 An example result of the QAM. “Cohen’s w ”: Calculated Cohen’s w for expression profile of a protein, “Cohen’s f ”: Calculated Cohen’s f for expression profile of a protein, “Max_FC”: maximum fold change, “QP_Pval_adjusted”: adjusted p-value from quasi-Poisson model, “NB_Pval_adjusted”: adjusted p-value from negative binomial model and “Normal_Pval_adjusted”: adjusted p-value from Normal ANOVA model.

<i>Protein</i>	<i>Cohen's w</i>	<i>Cohen's f</i>	<i>Max_FC</i>	<i>QP_Pval_adjusted</i>	<i>NB_Pval_adjusted</i>	<i>Normal_Pval_adjusted</i>
SERPINA1	0.4988	1.7646	3.1710	3.9790E-04	4.9878E-05	3.0741E-04
UBC	0.8655	0.7157	7.4570	1.3285E-01	5.2853E-02	4.3099E-01
IGLL5	1.2780	1.0265	96.4830	2.4177E-03	5.1797E-07	7.6510E-02
FCGBP	0.9688	4.9266	266.0565	2.9330E-11	0.0000E+00	1.8945E-10
CELA3A	0.4067	1.5158	6.8717	4.7076E-04	1.7161E-05	1.8453E-03
ACTB	0.6098	1.5480	5.0740	3.2750E-04	2.6138E-06	1.4676E-03
ACTG1	0.6098	1.5480	5.0740	3.2750E-04	2.6138E-06	1.4676E-03
UBB	0.8595	0.7446	7.4849	1.1914E-01	4.8966E-02	3.6476E-01
DMBT1	1.5035	3.7198	296.4068	3.4727E-09	1.9259E-12	1.2735E-08
IGJ	1.0418	1.3838	Inf	4.1234E-05	2.7157E-11	4.9763E-03

However, many biological experiments may have limited statistical power due to limited budgets and time. Selfea can complement low statistical power experiment by increasing ES. For example, significance is usually determined based on p-value that is calculated from a Student's *t*-test. P-value can indicate statistical significance at high statistical power. Therefore, Cohen's ES is required to complement the p-value's weakened selection power. By applying Cohen's *d* statistics into a Student's *t*-test at higher than 90% statistical power in concert with a p-value significance test, there is increased confidence in those values. Selfea provides probability-based thresholds and ES thresholds at the same time for reliable selection.

Finally, the QAM module supplies a protein scatter plot showing how significant proteins were selected from the total proteins (Fig. 19). In this example the red dots are insignificant proteins, while blue dots are significant. The horizontal line shows the p-value threshold, while vertical line indicates the ES threshold.

Functional analysis module (FAM)

Since metaproteome experiments generally have a large number of undistinguishable proteins, functional analysis improve the metaproteome's output rather than focusing on individual proteins. For this QFAM provides the FAM module. This consists of two modules: biosystems and COG analysis. Both modules use two statistical tests: the enrichment test and significance test.

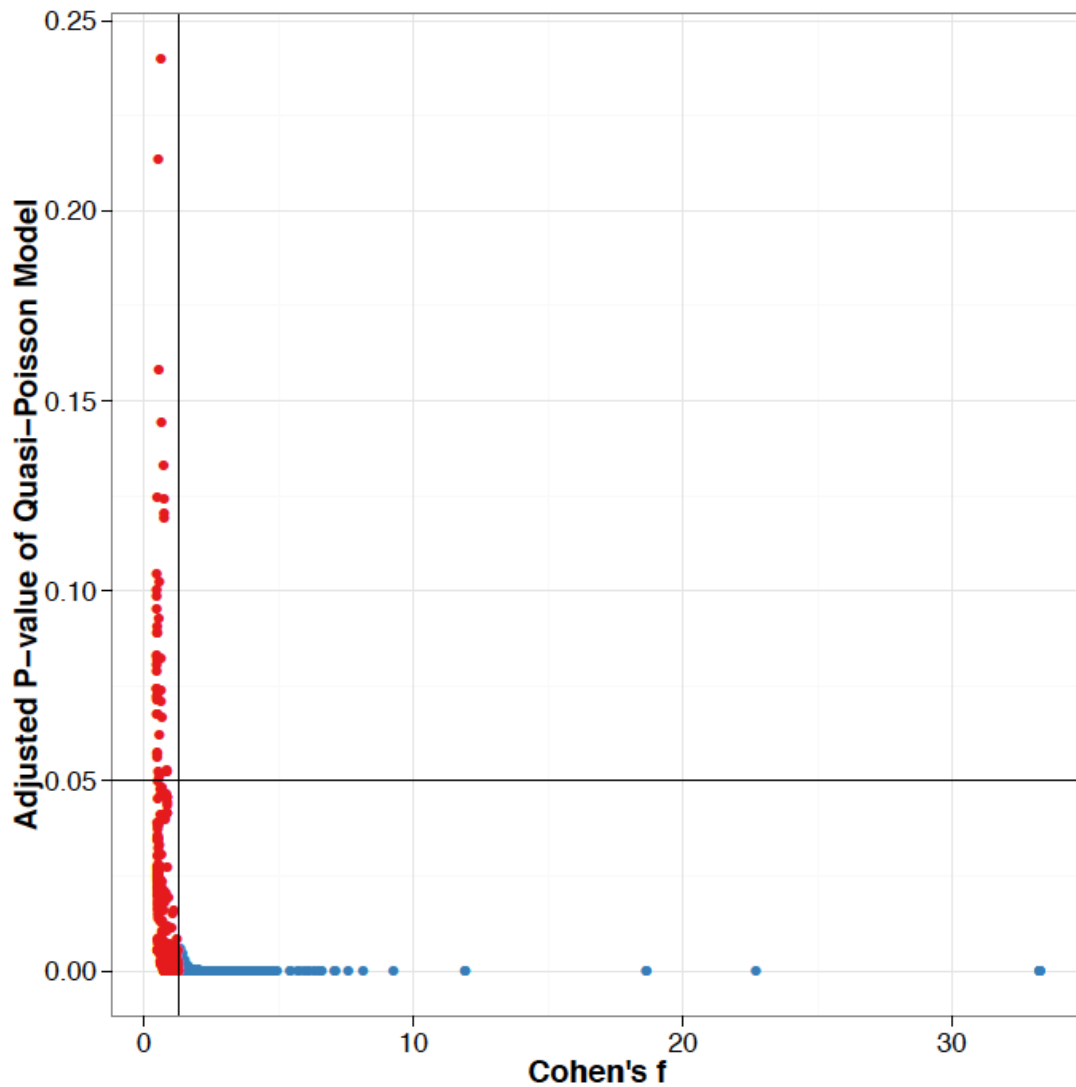


Figure 19 Protein scatter plot for the quasi-Poisson model and Cohen's f combination. Selfea provides a protein scatter plot for explaining how significant proteins were distinguished. Horizontal black line is false discovery rate (FDR) = 0.05, user-defined p-value threshold, and vertical black line is Cohen's f at 80% statistical power, which is typically the recommended power for reliable feature selection. Red dots are insignificant proteins, while blue dots are significant.

The enrichment test measures how significantly specific functional terms are enriched from total proteins. The significance test examines how the different terms are detected between independent experimental groups. These two statistical methods can report differential and abundant functions in these microbial communities.

Of the many functional analysis tools, FAM could be categorized as a second-generation functional class-scoring module (FCS). Khatri et al [111] grouped pathway analysis tools into three categories: over-representation analysis (ORA), FCS, and pathway topology (PT) (Fig. 20). FCS tools utilize expression profiles for finding significant functions, which is similar to FAM's significance test using selfea. The first generation ORA relied on pre-selected features such as differentially expressed genes or proteins. FAM's enrichment test is close to the ORA method. Therefore, FAM functions by using two kinds of analytical methodologies, ORA and FCS. The enrichment and significance test are used to identify important and differential functions. FAM does not consider network topology or different weights that are important issues in TP.

BioSystems analysis of FAM

The application of FAM to biosystems uses KEGG, BioCyc [112], Reactome [113], National Cancer Institute (NCI) pathway interaction database (PID) [114] and GO [115] to analyze the pathways based on the proteomics data.

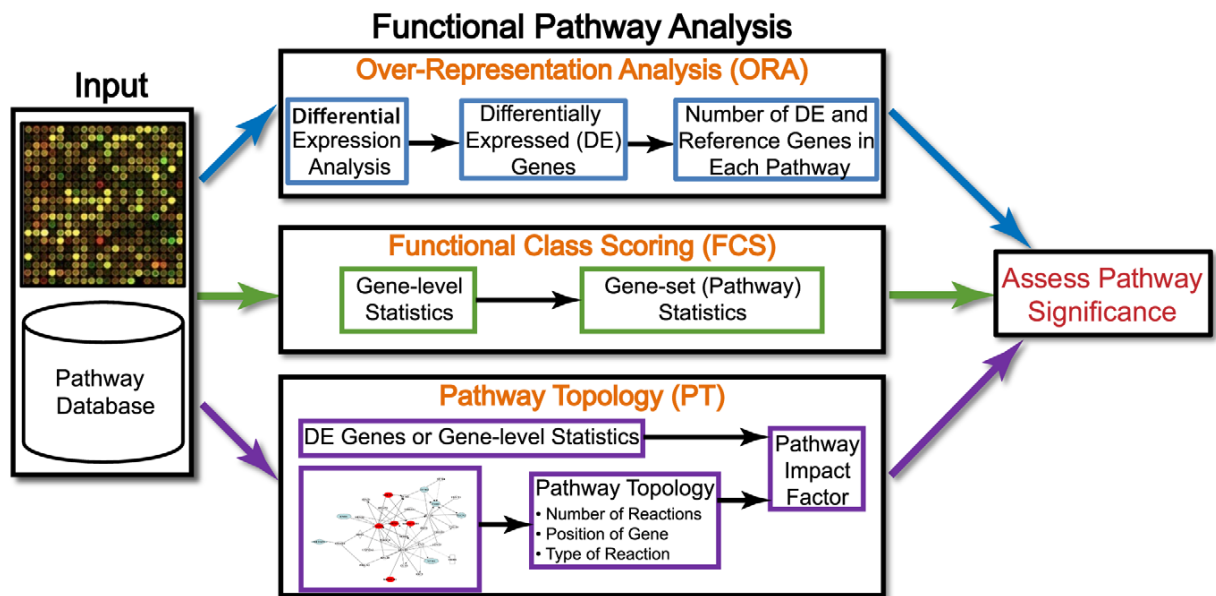


Figure 20 Overview of existing pathway analysis methods using gene expression data as an example.

These are functional annotations of biosystems, such as biological pathways and disease, which can involve any components of the cell (i.e., genes, proteins, metabolites and drugs). Of these programs, KEGG and GO are most widely used. BioCyc, Reactome and NCI PID have their additional advantages, such as pathway integration in BioCyc, reaction-focused responses in Reactome and PID for pathway interactions.

FAM consists of an enrichment test and annotation pipeline, downloads of source files, parsing and updating reference files for FAM analysis. This helps to maintain up-to-date information (Fig. 21). FAM annotation pipeline downloads and parses NCBI Entrez Gene [116], RefSeq [117] and BioSystems through file transfer protocol (FTP). Then, FAM parses the downloaded files and updates flat-files for NCBI BioSystems analysis. This process is automated and can be run regularly to keep up-to-date status.

Next, FAM executes enrichment tests for the different pathways. Proteins in the user's expression profile will have NCBI RefSeq, Entrez Gene or UniProt IDs because FAM's annotated reference files are based on NCBI GenBank and UniProt's SwissProt and TrEBML IDs. These IDs are transformed to an Entrez Gene ID that is directly connected to NCBI BioSystems IDs. After transformation, a Python script collects all the BioSystem terms found in the Gene IDs of the query expression profile. The number of proteins is counted for every BioSystem term.

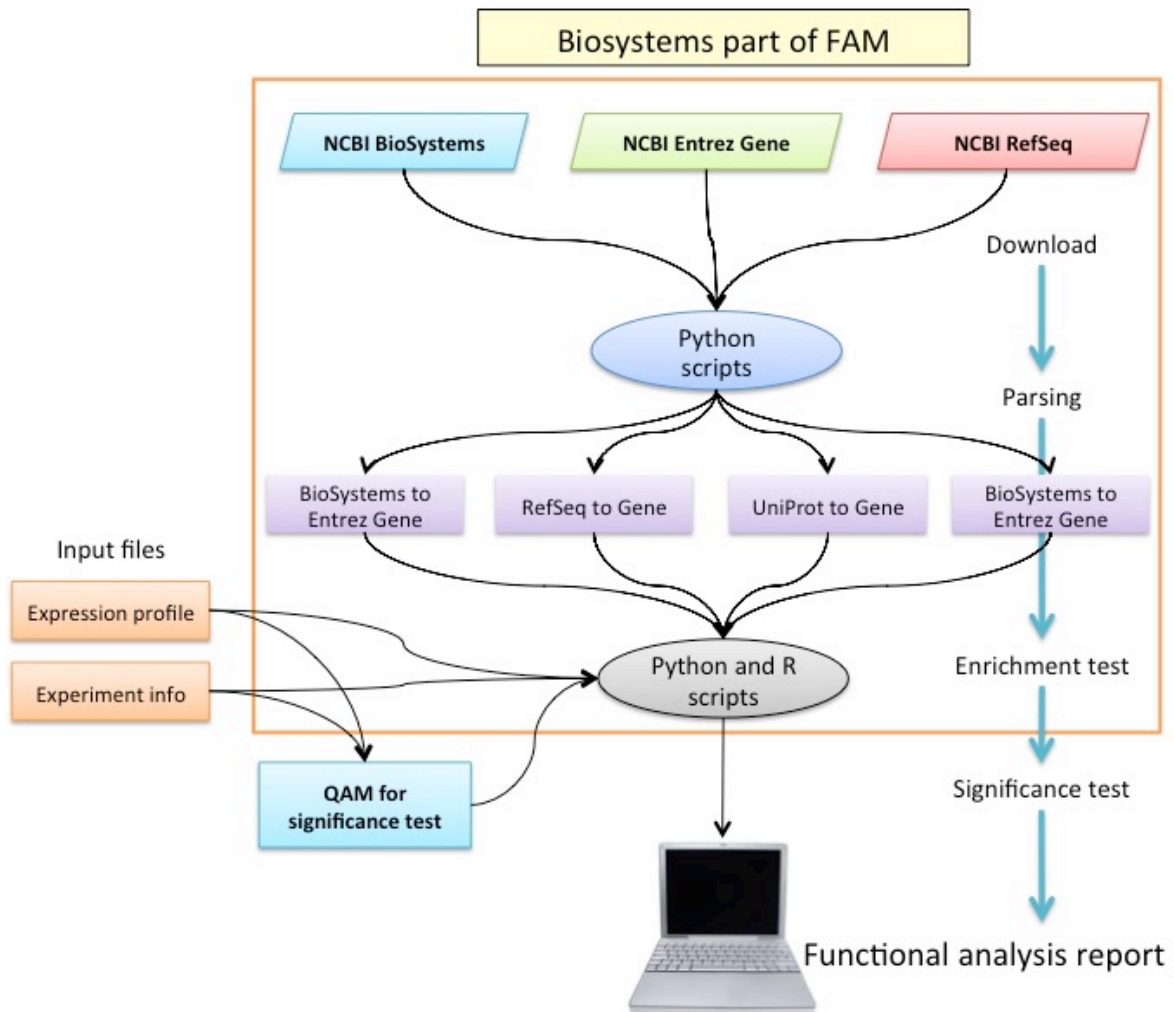


Figure 21 A diagram of the biosystems analysis of FAM. FAM has two parts, a script updating reference file and enrichment test. QAM is incorporated into FAM for significance tests. The user receives results from the enrichment and significance tests.

Table 8 An example of a contingency table for a pathway ID using the enrichment test of FAM. The numbers listed are the numbers of proteins in each category.

	<i>Expression profile data</i>	<i>BioSystems DB</i>
Matched	19	37
Unmatched	52,934	978,925
Total	52,953	978,962

Then, an enrichment test makes contingency tables using the number of found pathway proteins, the total number of proteins in the expression profile, and the total number of proteins in the BioSystems DB (Table. 8). With the constructed contingency table, Fisher's exact test is calculated using NumPy [118] and SciPy [119] packages. The acquired p-values for every spotted pathway ID are adjusted according to the FDR. The FDR was calculated using the Benjamini-Hochberg method for controlling of false positives caused by repeated hypotheses tests. This method is different from the one was used in PSM identification.

The significance test is conducted by calling QAM subsequent to the enrichment test. FAM produces an expression profile for all the detected pathways during the enrichment test. A pathway expression profile is constructed by transferring spectrum counts of identified proteins to the different pathways.

QAM is executed in concert with FAM to compute p-values and Cohen's ES for the pathway expression profile. The significance test identifies the different pathways, while the enrichment test how quantifies how many proteins are discovered. Hence, those two tests are mutually combined to report pathways that should have the most significance in a biological context. Basically, significant pathways are identified that have p-values less than 0.05 for both of tests and a Cohen's ES larger than 90% statistical power (Table 9). There are options for significance test methods, which selfea provides, such as quasi-Poisson model and Cohen's f and negative binomial model and Cohen's w .

COG analysis of FAM

COG analysis is a better approach to interpret the characteristics of the microbial community than pathway analysis. COG is not dependent on database annotation so it has better coverage than pathway annotations. Model organisms, such as human, mouse, yeast and *Escherichia coli*, have very good annotation because of they have been investigated for a very long time. However, since almost all the microbes are not cultivable in laboratories, most microbes encountered in environmental proteomics studies have limited or no annotation available. Hence, there are limited annotations of microorganisms in reference databases, such as NCBI Gene and BioSystems, compared to the model organisms mentioned above. Many annotations of microbial proteins are inferred by sequence alignment, but this has limited usefulness.

Table 9 Example result of FAM. FAM provides three statistics: adjusted p-value of enrichment test, significance test, and effect size. The table represents filtered pathways from a whole FAM analysis result using KEGG pathway p-value adjustment method for FDR, significance test method using quasi-Poisson and Cohen's *f*, enrichment test p-value < 0.05, significance test p-value < 0.05 and 80% statistical power. "Enrich FDR" is the FDR of the Fisher's exact test, "Source ID" uses the original DB ID in Biosystems (KEGG ID in this table), "# of prot" is the number of proteins and "Quasi-Poisson FDR" is the FDR calculated by quasi-Poisson model.

<i>BS ID</i>	<i>Description</i>	<i>Enrich FDR</i>	<i>Source ID</i>	<i># of prot</i>	<i>Cohen's f</i>	<i>Quasi-Poisson FDR</i>
287	Glycolysis / Gluconeogenesis	2.49E-12	ko00010	38	2.5236	4.24E-07
296	Fatty acid degradation	1.65E-06	ko00071	22	2.6618	4.81E-10
817567	Carbon metabolism	8.94E-11	ko01200	44	2.6143	2.41E-07
795174	Biosynthesis of amino acids	3.57E-09	ko01230	34	2.5554	6.84E-07
478	Focal adhesion	3.03E-06	ko04510	52	2.5316	1.81E-07
498	Olfactory transduction	2.39E-06	ko04740	11	1.6099	1.25E-04
169295	Pancreatic secretion	3.01E-10	ko04972	41	4.2238	9.48E-10
171868	Protein digestion and absorption	1.61E-09	ko04974	38	2.9432	2.82E-07
516	Pathogenic Escherichia coli infection	8.45E-08	ko05130	27	1.6479	1.07E-04
167191	Amoebiasis	4.87E-08	ko05146	39	2.2770	1.16E-06
662842	Viral carcinogenesis	6.77E-09	ko05203	59	1.5133	2.16E-04

Table 9 continued.

<i>BS ID</i>	<i>Description</i>	<i>Enrich FDR</i>	<i>Source ID</i>	<i># of prot</i>	<i>Cohen' s f</i>	<i>Quasi- Poisson FDR</i>
468194	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	4.87E-07	M00001	18	2.4862	6.24E-07
468195	Glycolysis, core module involving three-carbon compounds	9.19E-06	M00002	12	2.3486	1.03E-06

Table 10 Comparison of the number annotated proteins between KEGG and COG. The total number of identified microbial proteins is 5,354.

	<i>KEGG Pathway</i>	<i>COG</i>
The number of proteins annotated	2,254 (42.1%)	4,905 (91.6%)

For example, the coverage of KEGG and COG is different for the microbial proteins discussed in application study in Chapter 4 (Table 10).

There are two reasons why COG has better coverage than pathways. First, sequence comparisons are not an efficient way to annotate pathways, while COG was originally based on sequence comparisons. Although there are shared pathways in different species, many microbes have different pathways adapted for survival in diverse environments. Therefore, sequence comparisons for pathway annotation is restricted to universally shared pathways. However, proteins can functions in ways that are not covered by the pathways analysis. This limits the usefulness of using this pathway analysis. The pathways are manually generated by reading manuscripts and organizing the accumulated research findings into the determined pathways. Theoretically, all the proteins should be involved in at least one pathway, but often it takes multiple studies to uncover even single pathways. Unfortunately, most microbes are not well studied at this point. This limits pathway annotation to the many microbial proteins observed in metaproteomics studies.

COG assignment is built on protein sequences that are available in metaproteomics. COG terms are annotated to proteins through comparing query protein sequences with protein sequences in the COG DB. Each COG has protein paralogs from at least three lineages, so species information and sequence similarity are required to decide the COG assignment. COG is widely used to find microbial orthologous proteins, but NCBI stopped updating it after 2005. From 2010 NCBI released an updated COG DB by expanding species coverage and introducing new a COG assignment algorithm (COG software) [120]. Until the new COGnitor, a tool for COG allocation, came out, to find a similar COG ID one would run the data through RPS-BLAST [121] or PSI-BLAST [122] at certain E-value level, such as 0.00000001. COG requires at least three proteins from different lineages and needs to have enough similarity to other members of COGs, not to only one member. However, the simple method using RPS-BLAST and PSI-BLAST were widely accepted because they were easy to use and the original COG assignment algorithm, SymBet, was not. SymBet was complicated and slow to run on a local desktop.

The newly released COGSoftware employs the EdgeSearch algorithm that is faster and less complex than SymBet. This enables the new COGnitor to assign COG IDs to proteins in our dataset and avoids violating COG construction rules. COG analysis of FAM employs COGSoftware for COG allocation. COGSoftware goes through two steps to assign COG to new proteins. Prior to COG assignment, COGSoftware requires PSI-BLAST results to be compared to the query protein sequences run against COG DB. With these multiple sequence

alignments, the COGSoftware finds the best COG terms related to the query proteins.

FAM controls the COGSoftware and distributes the PSI-BLAST jobs to 48 CPUs for increased speed (Fig. 22). Sequence comparisons using PSI-BLAST take most of time of COG analysis. Basically, PSI-BLAST execution time depends on BLAST DB size and query protein size. In the FAM module, the size of the BLAST DB and the query protein FASTA file are large because FAM requires a whole protein for PSM searches of LC-MS/MS data. In metaproteomics, the DB is several hundred megabytes. To run PSI-BLAST simultaneously FAM divides the large protein DB into small chunks and runs PSI-BLAST simultaneously. In spite of parallel operations, if the protein DB is large PSI-BLAST sometimes takes up to several days because of the limited number of available CPUs (e.g., 12 hours for the human fecal proteome study in Chapter 4).

The next step of COG analysis is the enrichment test (Fig. 22). This enrichment test is similar to the BioSystems analysis. However, COG identification of the whole protein DB is limited or not provided by publicly available databases, while BioSystems access this information from NCBI BioSystems. Therefore, FAM identifies all the possible COGs in the whole metaproteomics protein DB despite of long PSI-BLAST running time (Fig. 22). Because the enrichment test measures how significantly proteins in the expression profile are enriched from the total DB, FAM needs COG identification of the whole protein pool.

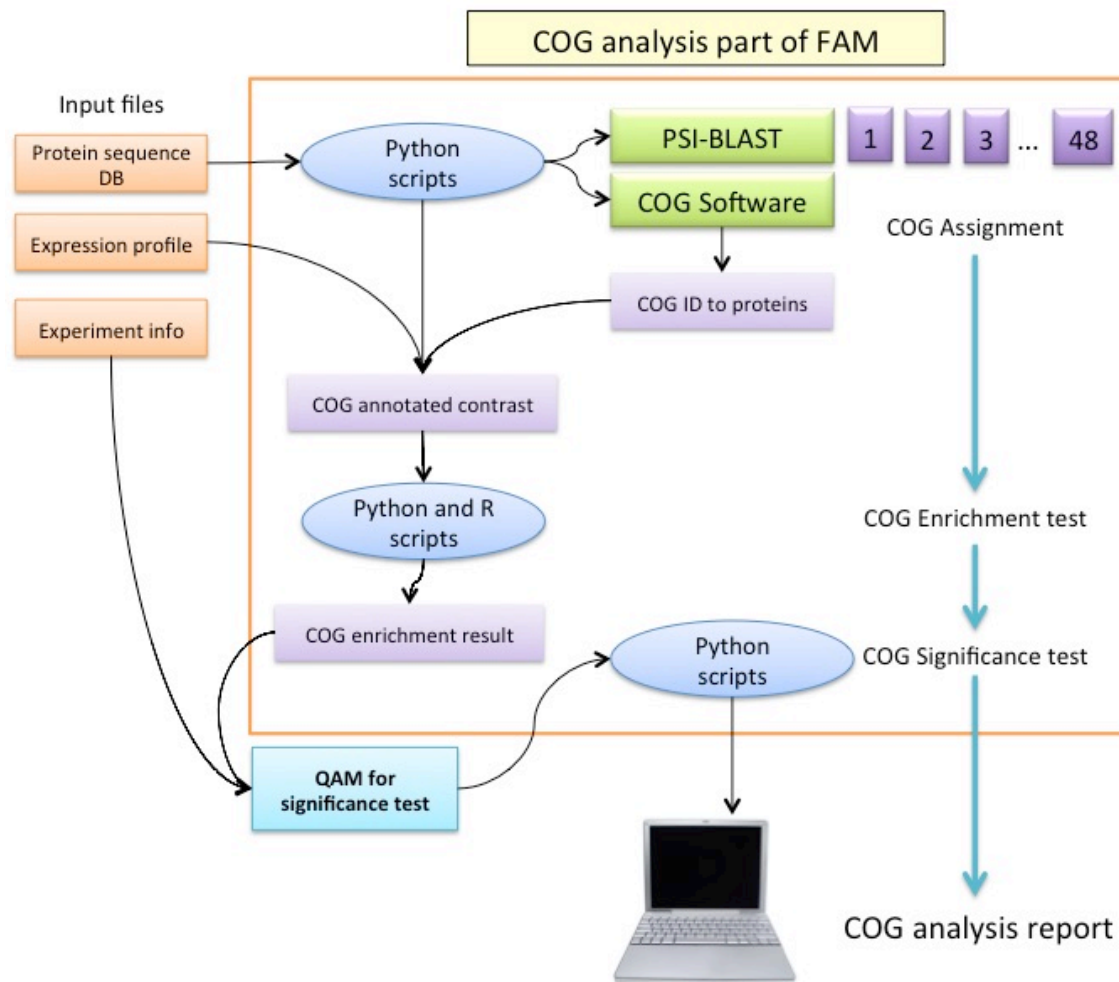


Figure 22 COG analysis of FAM. COG analysis runs PSI-BLAST and the COGSoftware to identify COG IDs of the query proteins. With the identified COG IDs, enrichment and significance tests are conducted by the R and Python scripts of FAM. Orange color box: input files; blue oval shape: python and R scripts; green box: COGSoftware and PSI-BLAST; purple box: distributed jobs; and light purple box: output files.

FAM makes a contingency table for every COG (Table 11) and calculates a p-value using Fisher's exact test.

The COG enrichment and significance tests are combined to provide significant and differentially enriched COG terms. All the p-values are adjusted by user-selected method, such as FDR, following the enrichment test. Three statistics, adjusted p-values from two statistics tests, and Cohen's effect size are integrated and finally significant COGs are reported as described in table 12. COG analysis of FAM enables us to obtain differentially and abundantly expressed proteins. The same analysis used for COG and BioSystems is repeated for assigning the COG functional categories to provide a wider view of the functional features of the microbial community. Because the COG analysis tools do not provide enrichment and significance tests at the same time, FAM is unique tool among functional analysis software.

Other modules of QFAM

QFAM has other modules besides the PSM search, QAM, and FAM. First, it has a module for metagenome proteins, translated proteins from metagenome sequences. Metagenomes have a wide variety of un-annotated sequences. Thus, metagenome sequences are not characterized even though those are differentially detected in metaproteomics studies. The metagenome can be functionally analyzed by COG identification in QFAM. However, annotation is required for BioSystems analysis of FAM. Therefore, QFAM has a module for protein annotation of metagenomes via sequence comparison.

Table 11 A contingency table for one COG ID from the enrichment test in FAM. For every COG term, FAM counts the number of matched and unmatched proteins in the expression profile data and entire metaproteomics DB. With these counts FAM generates contingency table. P-value is calculated using Fisher's exact test on the contingency table.

	<i>Expression profile data</i>	<i>Entire proteins in metaproteomics DB</i>
Matched	34	77
Unmatched	48,934	263,921
Total	48,968	263,998

Table 12 An example COG analysis in FAM. Same thresholds applied in table 9 were used for this table. “Enrichment FDR”: FDR of Fisher’s exact test; “Source ID”: Original DB ID in Biosystems (KEGG ID in this table); “Func cat.”: COG broad functional category; “# of prot”: the number of proteins; and “Quasi-Poisson FDR”: FDR calculated by quasi-Poisson model.

<i>COG ID</i>	<i>Description</i>	<i>Enrichment FDR</i>	<i>Func cat.</i>	<i># of prot</i>	<i>Cohen’s f</i>	<i>Quasi-Poisson FDR</i>
COG00542	ATPases with chaperone activity; ATP-binding subunit	2.33E-13	O	36	3.1885	4.27E-07
COG00050	GTPases - translation elongation factors	5.32E-12	J	26	1.5264	8.32E-05
COG00480	Translation elongation factors (GTPases)	5.32E-12	J	32	2.5506	9.50E-07
COG00055	F0F1-type ATP synthase; beta subunit	1.42E-11	C	24	2.7013	1.14E-06
COG00465	ATP-dependent Zn proteases	8.86E-11	O	27	3.3691	5.58E-08
COG00653	Preprotein translocase subunit SecA (ATPase; RNA helicase)	1.99E-10	U	23	1.4936	2.55E-04
COG00086	DNA-directed RNA polymerase; beta' subunit/160 kD subunit	1.99E-10	K	22	1.5654	1.30E-04
COG00048	Ribosomal protein S12	1.99E-10	J	21	1.6682	1.27E-04
COG00100	Ribosomal protein S11	1.99E-10	J	21	2.0664	4.58E-06
COG00085	DNA-directed RNA polymerase; beta subunit/140 kD subunit	2.75E-10	K	22	1.9143	4.63E-06

Table 12 continued

<i>COG ID</i>	<i>Description</i>	<i>Enrichment FDR</i>	<i>Func cat.</i>	<i># of prot</i>	<i>Cohen's f</i>	<i>Quasi-Poisson FDR</i>
COG00459	Chaperonin GroEL (HSP60 family)	2.75E-10	O	22	2.9585	7.46E-08
COG00052	Ribosomal protein S2	7.43E-10	J	20	1.5196	7.50E-05
COG00185	Ribosomal protein S19	7.43E-10	J	20	1.6309	6.39E-04

This module uses BLAST for sequence comparison with UniProt, which has an extensive protein sequence pool collected over the years. The most similar UniProt ID is assigned to a metagenome protein if their alignment does not have open gaps and shows more than 95% sequence similarity. Sequence similarity is calculated as described below:

$$Similarity = \frac{Aligned\ length}{Protein\ length} \times Percent\ identity$$

Where the aligned length and percent identity of the BLAST result and protein length is the number of amino acid residues. Because generally the metaproteome DB has a lot of metagenome proteins, running of the BLAST assignments is divided into multi-cores and merged after all the jobs are completed. With this annotation, we can apply the reference DB-based analysis to metagenome proteins.

QFAM provides an R library for drawing graphs including bar- and line-graphs and heat-maps using 'ggplot2' package. Because most biologically meaningful images are manually drawn, QFAM does not provide an automated module for making figures. However, if manually curated information is provided, QFAM can produce publication quality figures. Bar- and line-graphs showing expression profiles of significantly differential proteins across different experimental groups can be generated (Fig. 23). Line graphs can show expression profiles of multiple proteins over different samples. Proteins are colored differently to make clear distinctions between identified proteins.

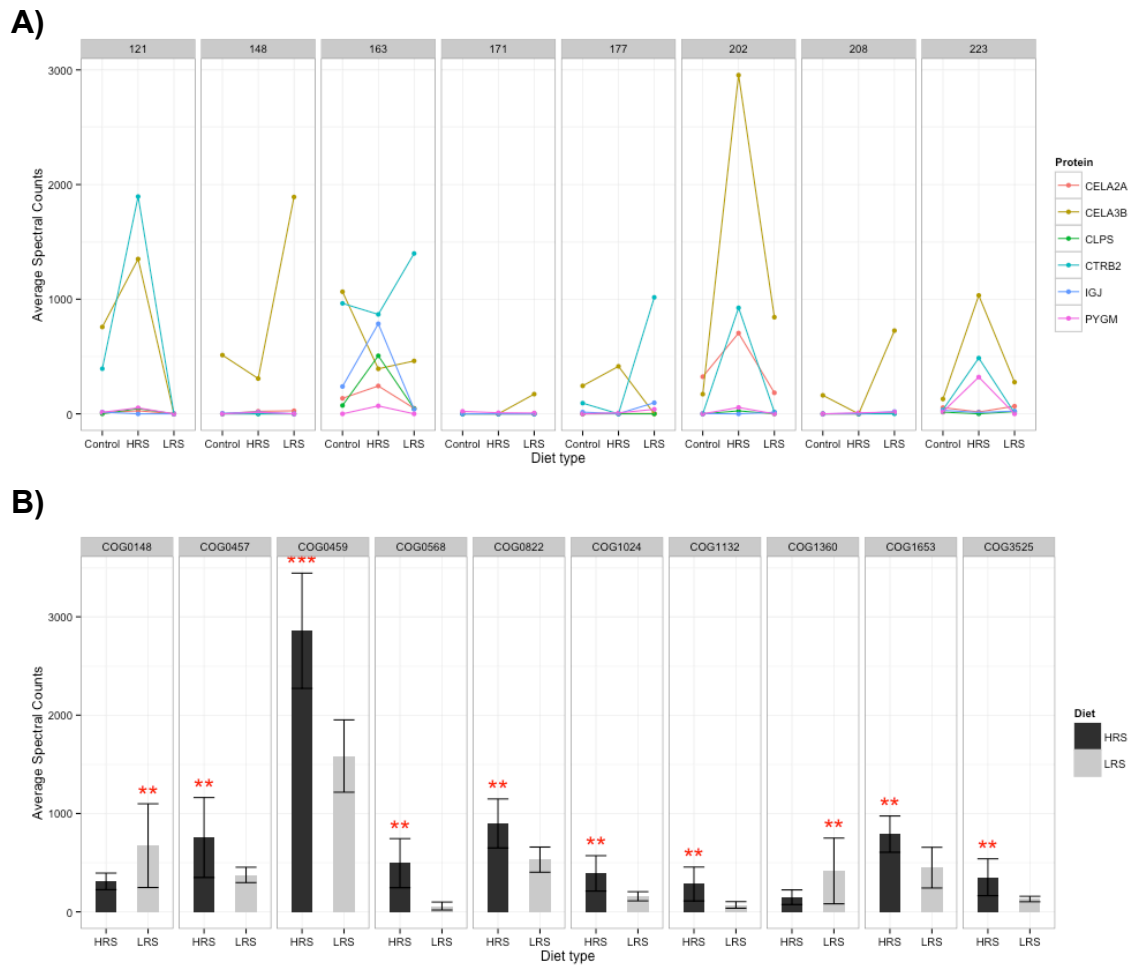


Figure 23 Examples of bar- and line-graph generated by QFAM. QFAM provides R library for drawing A) line graph and B) bar graph from analysis result of QFAM.

QFAM's bar graph for showing multiple proteins, pathways, or COGs in different experimental groups. Significance can be marked with red asterisks and error bars can be added to show data variation.

QFAM's R library Drawing heat maps are provided by (Fig. 24). Heat map figures are widely used in Omics based studies to display expression profile of many proteins or genes. Although heat-map can show global trends in whole dataset, it is not able to exhibit detail information. Heat map graphs are often combined with hierarchical clustering to group similarly expressed proteins. QFAM provides protein and experiment group clustering. Coloring of expression is standardized spectral counts calculated by $(\text{spectral counts} - \text{mean}) / \text{standard deviation}$. Colored bars located at upper side of heat-map can display experimental groups. Color-key is also provided by QFAM.

Conclusion

We developed QFAM for quantitative and functional analysis of metaproteome data. Since metaproteome data is highly complicated and less annotated, analysis pipeline should be able to run its processes in parallel computing environment and functionally interpret metaproteome proteins from limited annotated metagenomes. In terms of these requirements, QFAM is excellent pipeline for metaproteome data analysis. QFAM can run MyriMatch and IDPicker with 48 CPUs and automatically control all the processes of PSM search. MPA is another pipeline for metaproteome, but it doesn't provide multithreading and multiprocessing execution of PSM search steps.

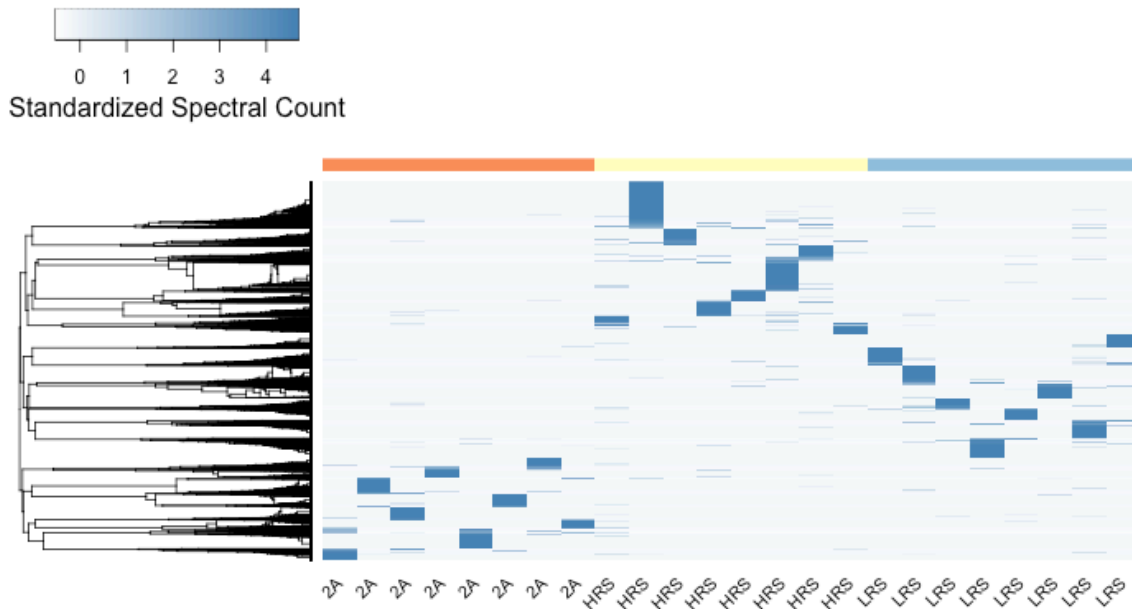


Figure 24 Example heat-map drew by QFAM's R library. Heat-map, colored by steel blue and white, can display global trend of protein expression. Hierarchical clustering for proteins and for experimental groups are available in QFAM's R library. In this example, only proteins are clustered to show how protein expressions are changed in different groups (color bars of upper side of heat-map). Heat-map's cells are standardized spectral counts by mean and variance.

Therefore, QFAM has an important advantage to MPA in terms of PSM search speed. Furthermore, QFAM is simple to install and optimize on any Linux cluster.

In addition to fast search speed, QFAM has QAM module for reliable feature selection. QAM employs selfea R package for finding of differentially expressed proteins. Selfea provides three widely used probability models, quasi-Poisson, negative binomial and normal distribution, with Cohen's effect-sizes, such as Cohen's f and w . Combination of p-values and effect-sizes is able to produce reproducible research findings. Detail explanation about selfea is in chapter 3.

QFAM services functional analysis module, FAM, for metaproteome data interpretation. FAM has two parts, BioSystems and COG analysis. BioSystems is based on reference databases of NCB, while COG is derived from sequence comparisons. BioSystems analysis is more appropriate for well-annotated organisms, such as human, mouse and yeast, while COG is the best analysis approach to metagenome proteins that takes biggest portion of metaproteome data. Through COG analysis and BioSystems analysis, FAM can cover from well-known species to undiscovered species. Since metaproteome has mixed proteome with deeply studied organisms and potential microbes that have never been observed to date, FAM is appropriate analysis pipeline for metaproteomics.

FAM provides two statistical tests, enrichment test and significance test, for every functional term including pathways and COGs. Enrichment test measures how many proteins are enriched from total proteins, while significance

test is for getting differentially detected terms over different experiment groups. Both of two tests can be mutually analyzed to provide differentially and meaningfully enriched characteristics of microbial communities. This FAM's approach is unique in many of published functional analysis tools because FAM integrates enrichment test and significance test together.

CHAPTER III

Selfea: R package for reliable feature selection

All of the data presented below has been adapted from the following submitted journal article:

Lang Ho Lee, Nathan VerBerkmoes and Arnold Saxton “Using Cohen’s Effect Sizes for Reliable Feature Selection of Quantitative MS-based Proteomics Data”. *Nature Methods* (In review).

Introduction

In data analysis from biological studies it’s often a struggle to obtain reproducible confident results. Most scientific discoveries, in the biological sciences, are statistically supported by p-value that is not always reliable evidence unless statistical power is high enough ($\geq 90\%$) [123]. However, many of biology experiments are not conducted with parameters that have enough statistical power due to limited budget and time. Button et al showed that 42 of 49 neurological studies, published in 2011, have less than 90% power that is typically recommended level for reliable hypothesis tests [123,124] (Fig. 25). Since p-value is fickle at insufficient power, it is often accompanied by effect size (ES), such as fold change (FC), in biological studies, to increase result confidence[125,126]. ES can show the magnitude of difference between experiment groups, while p-value assesses if an effect exists[125]. Combination of ES and significance filter can improve reproducibility[126]. FC is, however, not sufficient to supplement p-value.

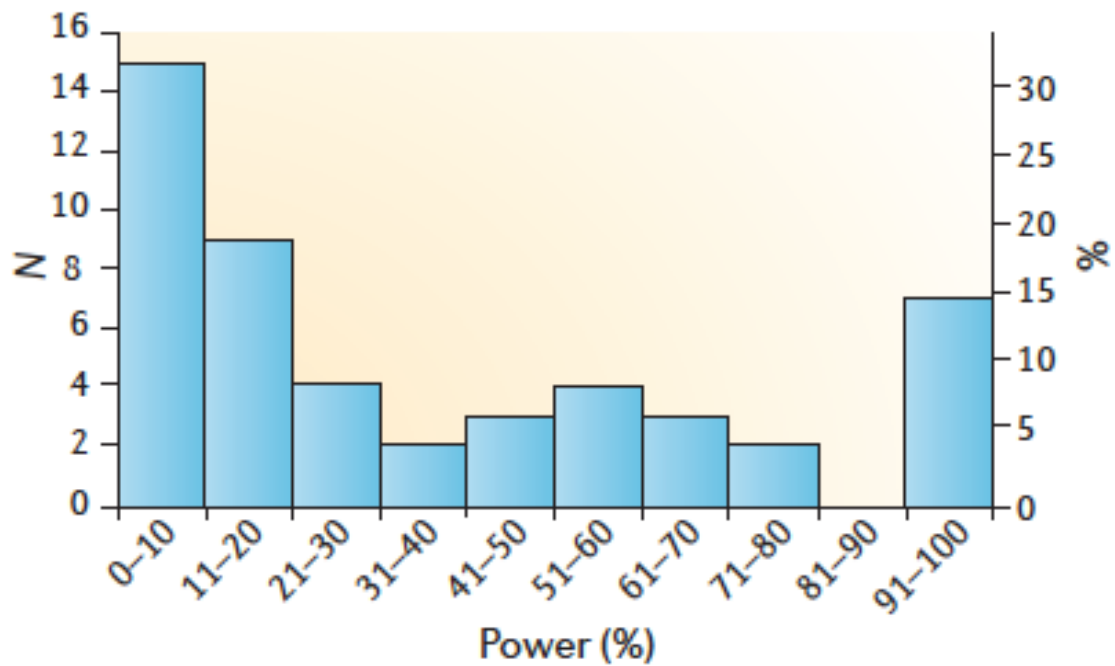


Figure 25 Median power of studies included in neuroscience meta-analyses. This figure shows a histogram of median statistical power of neuroscience papers published in 2011. If this figure was based on minimum power as we used, the power of papers should be far more below than 80%. This illustration was borrowed from Button et al [124]. N: the number of papers, Power(%): median statistical power

FC is easy to calculate and a ratio for relative comparison, familiar to biological research, but it has obvious limits. First, FC threshold is rooted on empirical evidence of biologists. FC is traditionally accepted because it means observable changes that are clear to the naked eye. However, this visible difference doesn't always mean significant effect, considering signal intensity and variation within replicates. The same FC threshold is generally applied to various data regardless of dataset's characteristics, such as variance and sample size. Furthermore, since FC is a relative ratio, it cannot distinguish absolute quantitative difference, which is main purpose of ES[125]. For instance, in label-free proteomics experiment, if expression level of proteinA has an observed 1 spectrum count (SC) in control and 2 SC in test group, FC of proteinA is 2. It is still same 2-FC, however, to another proteinB whose expression level is 1,000 SC in control and 2,000 SC in test group. Difference in the two groups for proteinA is 1 SC while for proteinB is 1,000 SC. We may not be sure whether the 1 SC difference of proteinA is true or from systematic-error because protein detection and quantitation are limited by the sensitivity of the instrumentation making the measurement (low spectral count proteins in label free proteomics is always suspect). As was confirmed in microarray[127], FC is no-good at handling low expression level features (genes or proteins) that are prevalent in biological data.

Methods

Label free LC-MS/MS Proteomics Data

Gregori data were obtained from 'msmsTests' R package[126,128]. Fecal proteome data were generated in a study of chapter 4, human fecal proteome study, via label free shotgun proteomics. The detail information of data generation is in method part of chapter 4.

Calculation of Cohen's effect sizes and statistics

All the statistical analysis, used in this study, was done with R 3.1.3. Cohen's effect sizes, fold change and p-values were obtained by 'selfea' R package we developed. Selfea is available at <http://cran.r-project.org/web/packages/selfea/>. Selfea calculates Cohen's f as described below:

$$f = \sqrt{\frac{\sum_{i=1}^k p_i (\mu_i - \mu)^2}{MSE}}$$

Where $p_i = N/n_i$, k = the number of groups, n_i = the number of observation in group i , N = total number of observations, μ_i = mean of group i , μ = grand mean over all the observations, and MSE = mean square error[129,130]. Cohen's w is calculated in selfea as described below:

$$w = \sqrt{\sum_{i=1}^m \frac{(p0_i - p1_i)^2}{p0_i}}$$

Where m = the number of cells, $p0_i$ = cell probability in i -th cell under null hypothesis, and $p1_i$ = cell probability in i -th cell under alternative hypothesis. We

used 'ES.w1' function of 'pwr' package[131] for Cohen's w calculation because we compare goodness of fit of two NB models[129,130,132]. Selfea calculates minimum Cohen's f using 'pwr.anova.test' function of pwr package with parameters (k = the number of groups, n = the common sample size in each group, sig.level = p-value threshold defined by an user (0.05 by default) and power = statistical power defined by an user (0.90 by default)). Cohen's w is calculated in selfea through 'pwr.chisq.test' function of pwr package with parameters (N = the total sample size, df = the degree of freedom, sig.level = p-value threshold defined by an user (0.05 by default) and power = statistical power defined by an user (0.90 by default)). Selfea calculates p-values from quasi-Poisson, negative binomial and Normal distribution. Since generalized linear model of basic R does not provide negative binomial distribution, 'glm.nb' function of MASS package[133] was used for making a negative binomial linear model. Generated linear models for quasi-Poisson, negative binomial and Normal distribution were compared against null linear model using 'anova' function. One-way ANOVA F-test is used to compare linear models based on quasi-Poisson and Normal distribution, while chi square likelihood ratio test is used for a negative binomial linear model. Since one-way ANOVA F-test is used for quasi-Poisson and Normal distribution linear models, Cohen's f is coupled. Negative binomial linear model is paired to Cohen's w due to chi-square likelihood ratio test. All the p-values were adjusted by FDR[110]. Fold change between experiment groups was calculated by selfea. Selfea returns maximum fold change among all the possible pairs as representative fold change.

Minimum statistical power is reported as representative statistical power in selfea package.

Benchmark data test and fecal proteome data test

We employed the benchmark data generated by Gregori et al[126] and fecal proteome data in order to test Cohen's effect sizes. The dataset, termed here as fecal proteomes, have 29 2D-LC-MS/MS runs from 11 fecal samples from healthy adult volunteers (unpublished). Selfea was used to find differently expressed proteins in 11 fecal samples. Gregori data consist of 19 label-free liquid chromatography-tandem mass spectrometry runs and have 4 groups that have a fixed loading amount of standard yeast lysates (500 ng) with different amounts of 48 human proteins (100, 200, 400 and 600 fmol). Differentially expressed proteins over 4 groups were found by using selfea. As is assumed by Gregori et al[128], we considered human proteins as true positives and yeast proteins as false positives. Using selfea, we got differentially expressed proteins of Gregori data and fecal proteome data through six methods; quasi-Poisson (FDR < 0.05) and Cohen's f (Power > 0.90), quasi-Poisson (FDR < 0.05) and 2 fold change, negative binomial (FDR < 0.05) and Cohen's w (Power > 0.90), negative binomial (FDR < 0.05) and 2 fold change, Normal ANOVA (FDR < 0.05) and Cohen's f (Power > 0.90), and Normal ANOVA (FDR < 0.05) and 2 fold change. With the significant proteins, we counted human proteins as true positives and yeast proteins as false positives. Percentage of false positives in a signature is the number of false positives divided by the total number of proteins

in a signature. In addition to NB, QP and NOR model, we calculated p-values using QProt to check if Cohen's ES can be successfully combined with other type method. QProt was downloaded from <http://sourceforge.net/projects/qprot> and installed on the Newton computational cluster at the University of Tennessee. Since QProt provides only two-sample test, we tested all the possible pairs of experiments of Gregori data. After collection of all the FDR-adjusted p-values from six two-sample tests, we picked minimum FDR as representative FDR. Input parameter of QProt was nburnin=2,000, niter=10,000, and normalized=1. Bar plots were drawn by Microsoft Excel. Scatter plots were drawn by 'ggplots2' package[108].

Monte Carlo simulation

We used 'rpois' function of R to generate random data based on Poisson distribution with mean expression level of each groups in Gregori data. Since Gregori data has four groups, we separately generated random data for each of four groups (λ = mean spectral counts within groups). From the randomly generated data, p-values, fold change and Cohen's effect sizes were computed by selfea. Six methods, used for the benchmark data test, were examined. We calculated minimum statistical power and counted human proteins as true positives and yeast proteins as false positives in significant proteins. FP rate is the number of false positives divided by the total number of yeast proteins in a random data. This simulation was repeated 1,000 times. Box-plots were drawn by 'ggplot2' package[108].

User guide for selfea package

R document of selfea is available at CRAN, R package repository, and its web addressed is <http://cran.r-project.org/web/packages/selfea/selfea.pdf>. This R document provides detail information about every function and data as well as examples.

Algorithm

To overcome limitations of FC, we developed an alternative approach based on Cohen's ES. Cohen's ES are designed to measure difference between groups and there are specific types for different hypothesis test methods, such as Cohen's d for Student's t-test[134]. Cohen's ES is, moreover, a parameter for statistical power calculation of a hypothesis test[134,135]. The higher Cohen's ES is, the better statistical power is. Statistical power is the probability of correctly rejecting the null-hypothesis when it is false. Therefore, higher statistical power can improve reliability of quantitative biological studies. We can obtain higher statistical power by increasing threshold of Cohen's ES filter.

First, our approach calculates Cohen's f and w for each comparison between experiment groups for every individual feature[129,135]. Cohen's f is for one-way ANOVA F-test using quasi-Poisson (QP) or Normal distribution (Normal). Cohen's w is linked to the Chi-square likelihood-ratio test (LR-test) for comparing goodness of fit of two generalized linear models based on negative binomial distribution (NB). Those three, QP, Normal and NB, are widely used distributions for treatment comparisons in biological data.

Next, the minimum Cohen's f and Cohen's w are calculated for desired statistical power, typically 90%, and alpha, typically 0.05, by using power analysis function[129,131]. Features having Cohen's f or w below the obtained minimum are filtered out. This gives a statistical basis to determine ES threshold by considering sample size, significance and statistical power. In addition to the statistical benefits, our approach can maintain the level of statistical power by filtering out low-powered hypotheses tests.

We incorporated these concepts in an R package named 'selfea', freely available on CRAN. To test if its performance was significantly better than traditional FC approach, we employed benchmark data, generated by Gregori et al[126], because it has clear true positive (TP) and false positive (FP) identifications. Gregori data consist of 19 label-free liquid chromatography-tandem mass spectrometry (LC-MS/MS) runs and have 4 groups that have a fixed loading amount of standard yeast lysates (500 ng) with different amounts of 39 purified standard human proteins (100, 200, 400 and 600 fmol). Theoretically, the human proteins should show significant quantitative differences over the 4 groups while yeast proteins should not. As is assumed by Gregori et al[128], we counted human proteins as TP and yeast proteins as FP in signatures of Gregori data. At FDR[110]-adjusted p-value (FDR) < 0.05, we compared Cohen's ES (> 0.90 power) to 2-FC (> 2-fold change) (Fig. 26). We chose 90% power for ES threshold because it is the level p-value requires for being reliable evidence[123]. We picked maximum FC as representative FC and minimum statistical power as representative statistical power.

Figure 26 Testing result with Gregori data and fecal proteome data. A) Statistical power, B) the number of true positives (human proteins) and C) the number of false positives (yeast proteins) in significant protein list of Gregori data benchmark test for 6 method sets, QP_F, QP_FC, NB_W, NB_FC, NOR_F and NOR_FC described in table 13. FDR thresholds for QP, NB and NOR are 0.05 and Cohen's ES (f and w) threshold is 90% statistical power. Scatter plot of 39 human (red) and 646 yeast (blue) proteins comparing D) Cohen's w and FC, E) Cohen's f and FC, F) QP model and Cohen's f , G) NB model and Cohen's w and H) NOR model and Cohen's f .

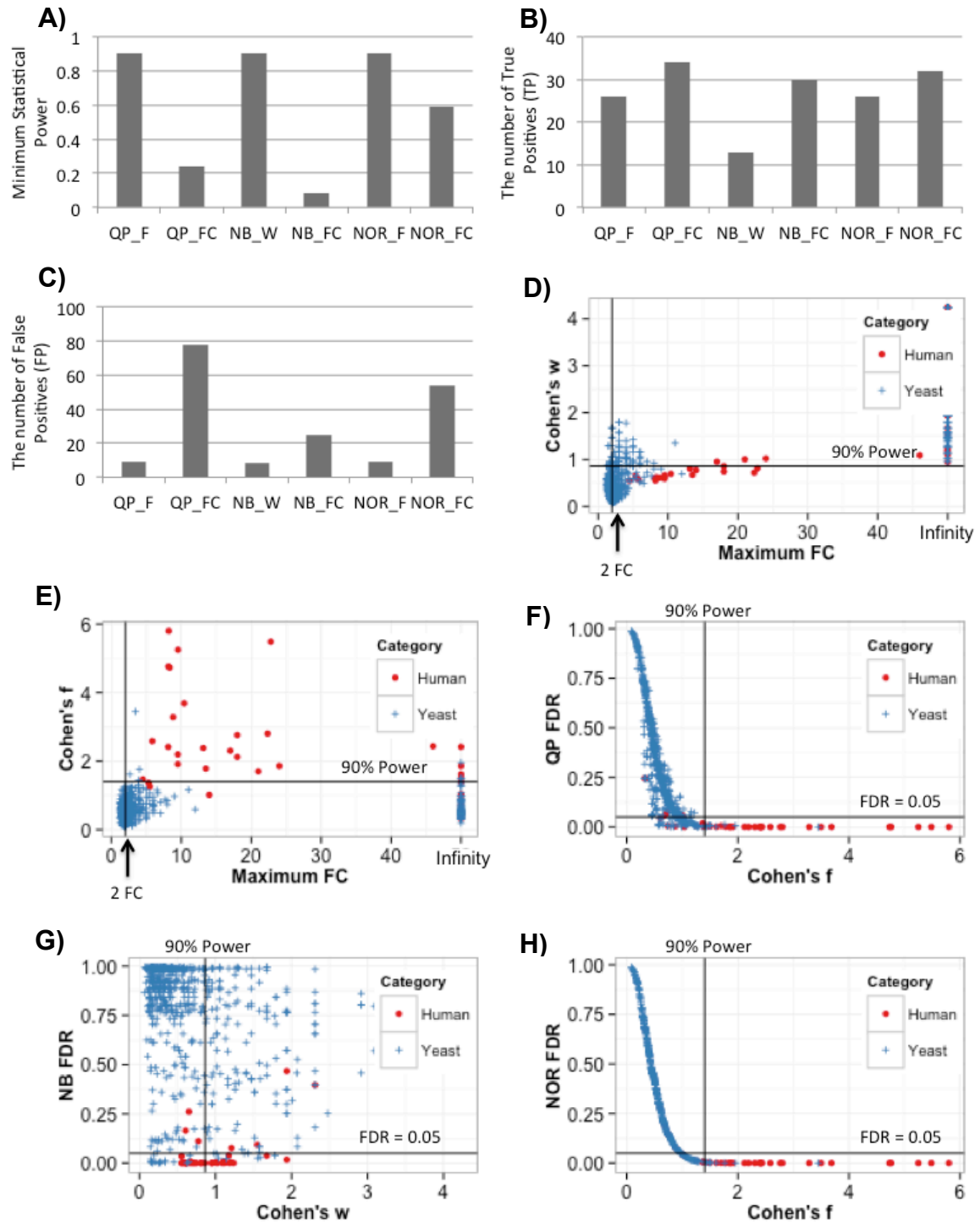


Figure 26 continued.

Benchmark test

Tests using Gregori data

We compared 2-FC and Cohen's ES for three models (QP, NB and NOR) using the benchmark data (Table 13). The benchmark data test showed that our approach has significantly better statistical power than 2-FC, 0.90 for Cohen's ES vs. 0.08~0.59 for 2-FC (Fig. 26A). Cohen's ES has similar number of true positives (TP) (Fig. 26B), while it has a lower number of false positives (FP) (Fig. 26C). This means that Cohen's ES lost fewer of TP while filtering out many more FP than 2-FC.

Protein scatter plots show that Cohen's ES filters out FP efficiently. Cohen's w (Fig. 26D) and Cohen's f (Fig. 26E) could remove more FP (blue dots) while excluding less number of TP (red dots). Especially, Cohen's f has superior selection power than Cohen's w and 2-FC (Fig. 26E). Also, Cohen's f showed that it could complement p-value by efficiently removing FP (Fig. 26F and 26H), while 2-FC removed a smaller number of FP (Fig. 27A-C). Relatively low performance of Cohen's w and NB (Fig. 26G) is speculated that these are not fit to the benchmark data. Nevertheless, Cohen's w showed better selection power than 2-FC (Fig. 26A-C).

Tests using fecal proteome data

We also compared the statistical methods with a more complex and real human dataset that is noisier than the well-organized benchmark data.

Table 13 Six methods used for comparing Cohen's ES and FC. Three probability models, widely used in biological data analysis, were employed for comparison of Cohen's ES with FC.

	<i>Quasi-Poisson model (QP) (< 0.05 FDR)</i>	<i>Negative binomial model (NB) (< 0.05 FDR)</i>	<i>Normal ANOVA model (NOR) (< 0.05 FDR)</i>
FC2 (> 2-fold change)	QP_FC2	NB_FC2	NOR_FC2
Cohen's f (> 90% power)	QP_F	-	NOR_F
Cohen's w (>90% power)	-	NB_W	-

Figure 27 Protein scatter plots comparing three models and FC. False positive yeast proteins (blue color cross) and true positive human proteins (red color dot) were distributed on the plain of A) FDR-adjusted p-value of QP model and FC, B) FDR-adjusted p-value of NB model and FC and C) FDR-adjusted p-value of Normal model and FC. Vertical line is 2-FC and horizontal line is 0.05 FDR. Red-colored left-bottom section, divided by vertical and horizontal lines, is filtered by 2-FC after probability filter ($FDR < 0.05$). Little or none of FPs (blue dots) is removed by 2-FC.

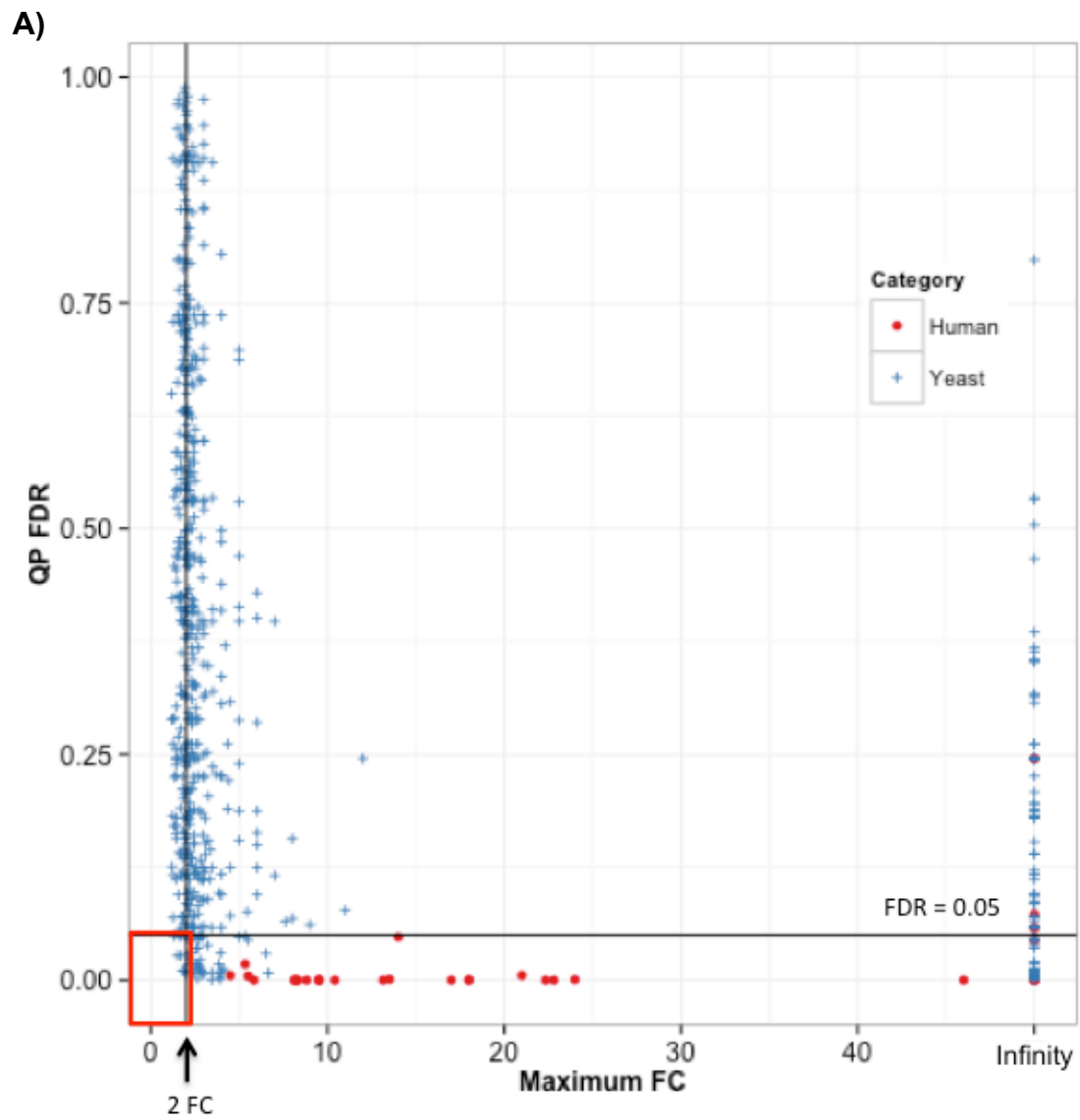


Figure 27 continued.

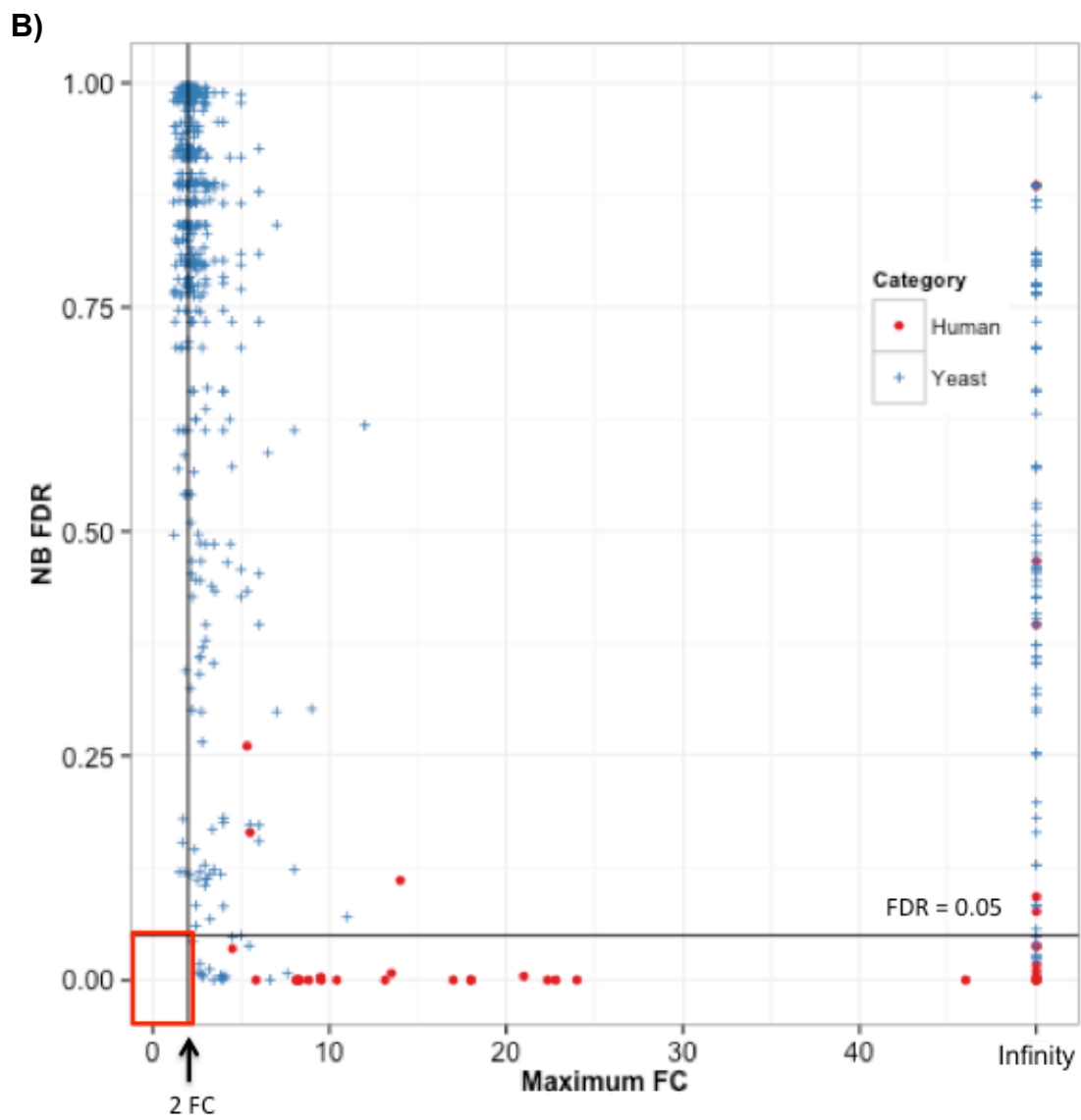


Figure 27 continued.

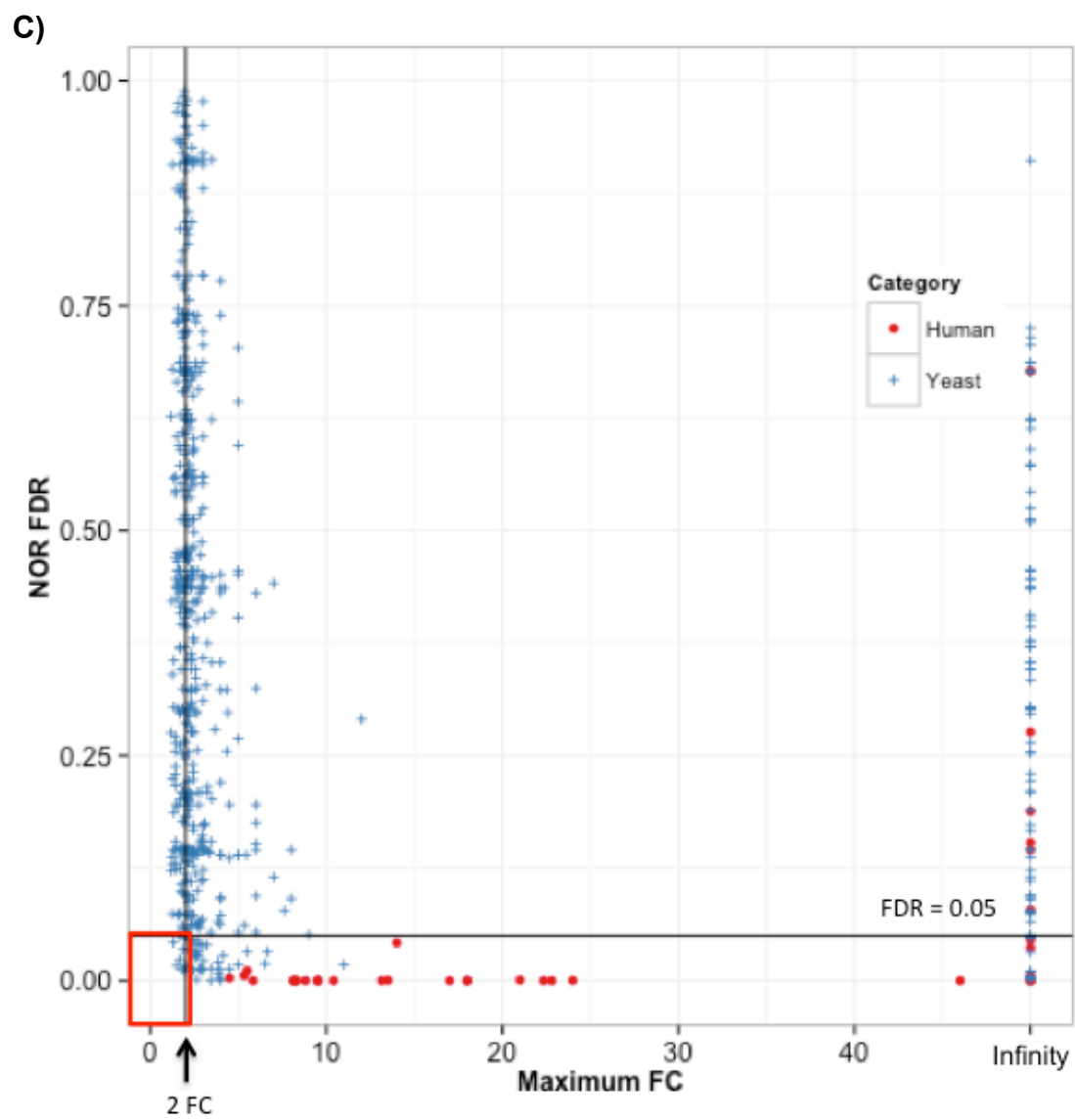


Figure 27 continued.

The dataset, termed here as fecal proteomes, have 29 2D-LC-MS/MS runs from 11 fecal samples from healthy adult volunteers (chapter 4). The fecal proteome is extremely complex having proteins coming from both human and bacterial origin, over a wide dynamic range, and high variability between individuals. Since we can't know TP and FP of fecal proteomes, we compared statistical power between Cohen's ES and 2-FC (Fig. 27). Similar to Gregori data, Cohen's ES had higher statistical power than 2-FC (0.90 for Cohen's ES vs. 0.16 ~ 0.70 for 2-FC).

Monte Carlo simulation

After the verifications of our method using the benchmark and actual experiment data, we explored our method via the Monte Carlo simulation (MC). We generated random data based on Poisson distribution with mean expression level of each groups in Gregori data. Then, we calculated statistics and counted TP and FP. These simulations were repeated 1,000 times in order to investigate how much our approach is effective. The simulation results are summarized and illustrated in Fig. 2. We could confirm that Cohen's ES have better power (0.90 for Cohen's ES vs. 0.06~0.55 for 2-FC) (Fig. 28A), reduced FP greatly (29~32 for Cohen's ES vs. 70~114 for 2-FC) (Fig. 28B) and lost fewer TP (20 for Cohen's ES vs. 27~32 for 2-FC) (Fig. 28C) when compared with 2-FC. FP rate is the probability of falsely rejecting the null hypothesis and is calculated by FP divided by total number of yeast proteins in Gregori data.

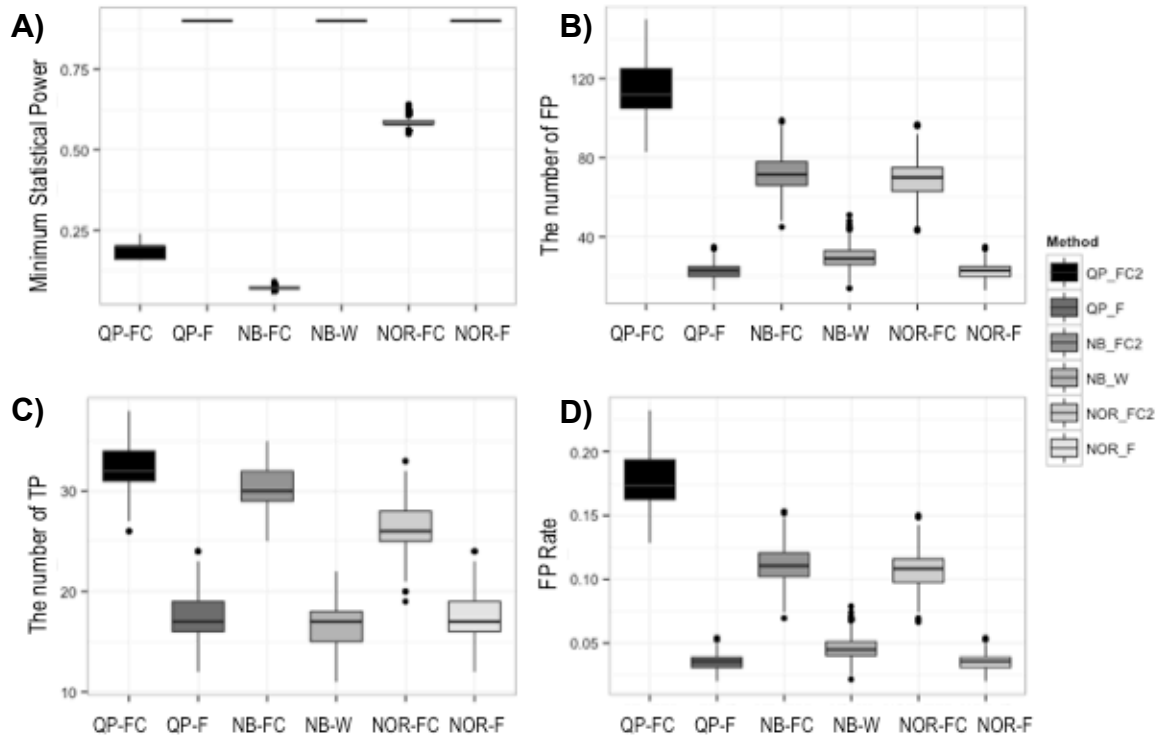


Figure 28 Boxplots of 1,000 times Monte Carlo simulation of Gregori data.

A) Statistical power, B) the number of false positives (yeast proteins), C) the number of true positives (human proteins) and D) FP Rate, $FP/(FP+TN)$, of 1,000 times Monte Carlo simulation, based on Gregori data, for 6 method sets, QP_F, QP_FC, NB_W, NB_FC, NOR_F and NOR_FC described in table 13. FDR thresholds for QP, NB and NOR are 0.05 and Cohen's ES (f and w) threshold is 90% statistical power. TN: True Negative (The number of total yeast proteins in Gregori data – FP).

Cohen's ES has a smaller FP rate than 2-FC (< 5% of Cohen's ES vs. > 11% of 2-FC), which means that our method is better at discerning FP than 2-FC (Fig. 28D). Similar to Gregori data, Cohen's ES had higher statistical power than 2-FC (0.90 for Cohen's ES vs. 0.16 ~ 0.70 for 2-FC).

Monte Carlo simulation

After the verifications of our method using the benchmark and actual experiment data, we explored our method via the Monte Carlo simulation (MC). We generated random data based on Poisson distribution with mean expression level of each groups in Gregori data. Then, we calculated statistics and counted TP and FP. These simulations were repeated 1,000 times in order to investigate how much our approach is effective. The simulation results are summarized and illustrated in Fig. 2. We could confirm that Cohen's ES have better power (0.90 for Cohen's ES vs. 0.06~0.55 for 2-FC) (Fig. 28A), reduced FP greatly (29~32 for Cohen's ES vs. 70~114 for 2-FC) (Fig. 28B) and lost fewer TP (20 for Cohen's ES vs. 27~32 for 2-FC) (Fig. 28C) when compared with 2-FC. FP rate is the probability of falsely rejecting the null hypothesis and is calculated by FP divided by total number of yeast proteins in Gregori data. Cohen's ES has a smaller FP rate than 2-FC (< 5% of Cohen's ES vs. > 11% of 2-FC), which means that our method is better at discerning FP than 2-FC (Fig. 28D).

Conclusion

P-values should not be the only piece of information used to discern reproducible findings in biological studies. Although effect size filters can complement p-value information, fold-change, widely used effect size in biological data analysis, is not sufficient. We suggest using Cohen's effect sizes at a desired statistical power. The benchmark test and MC simulation show that our method can reduce a number of FP while losing relatively small number of TP. This drop of FP is meaningful because many published research findings are concerned by potential false positives due to misused statistical approaches including low statistical power[124,136]. In addition to QP, NB and NOR, Cohen's ES can be associated with another type of probability-based method, such as QProt[137], and showed good selection power when compared to 2-FC (Fig. 27). We believe that Cohen's ES can help biologists to report reproducible discoveries. Even though selfea was tested only for label-free proteomics data, we believe that it can be applied to other type of biological data, such as RNA sequencing.

CHAPTER IV

APPLICATION CASE 1 – HUMAN FECAL PROTEOME STUDY

Introduction

Microbes inhabit every part of human body including the skin, the gut, mouth and the nose. Sometimes they cause illness, but most of the time, microorganisms commensally live with their human host organs. The number of microbes in symbiosis with human is estimated to 100 trillion, while we have 10 trillion cells, thus outnumbering human cells by 10 to 1. The microbes are not just living in human body, but they are interacting and playing a vital role in human health and diet [138]. In recent years the human gut microbiome, all the microbes living in human gastrointestinal tracts, has become an area of heavy in-depth research by a large number of labs across the planet. Many of research groups applied a wide range of analytical technologies for understanding microbiome's role in human health and how they interact with human host cells [139,140]. Also, gut microbiota are known to be associated with disease states, such as inflammatory bowel disease [140] and type-2 diabetes [141]. Since microbiome has unique microbial communities related to human health, the microbiome is an important target for metaproteomics [142].

Typically, metaproteome studies of gut microbiota have focused on fecal samples because feces are considered a direct proxy for the gastrointestinal tract. Furthermore, fecal material has a large amount of microbial proteins, convenience of sample collection interplay between the immune response of

human intestinal cells and microbiome [143]. Also, feces have abundant host proteins, so it enables us to explore biomarkers of disease and interactions between microbes and human cells. Therefore, fecal metaproteomics has a great potential for disease diagnostics, studying host-microbes interaction and microbial community cohabiting in gastrointestinal tracts.

Fecal metaproteomics has several unresolved problems that must be overcome before it can really be used in biomarker studies, disease state prediction, etc. First, we don't know the overall diversity of microbes inhabiting the fecal material. In other words, species composition of microbiome is not uncovered. This is important issue to PSM search. However, many studies have resolved this uncertainty by employing metagenome sequence directly from fecal material [142-144]. Another challenge is variation of individual microbiomes. Naturally, people have different microbiome due to diets, origin, drugs and genetic backgrounds. Probably, personal microbiome may be the most varying organ in entire human body, if the microbiome is considered as a 13th organ. We are not sure if there is normal microbiome shared by all the people. This large variation of the microbiome may make effective metaproteomics difficult.

The potential high variation of fecal proteomes will make the statistical analysis of resultant datasets difficult. Indeed, sample size is very limited for many biological studies due to finite budget and time; hence low statistical power is ineluctable destiny of biology researches conducted by small labs including fecal metaproteomics studies. At low-powered studies, high variance within biological replicates, caused by dissimilar microbiomes corresponding to different

environmental factors, impedes obtaining reproducible results via comparison between different individuals, disease states or diets. Thus the chance of false positive findings may be relatively high in fecal metaproteomics. This application study shows if there is large variation within fecal proteomes from healthy volunteers and how QFAM can be employed to analyze such a dataset.

Methods

Sample preparation and LC-MS/MS experiment

Proteins from all fecal samples were extracted by the direct cell lysis method (Fig. 29). Direct cell lysis method is preferred to collect host proteins, while indirect method is better for microbial proteins. Proteins of all the samples were digested by trypsin and analyzed via two-dimensional (2D) nano-LC MS/MS system with a split-phase column (SCX-RP) Velos Orbitrap Pro (Thermo Fisher Scientific) for feces and LTQ-Velos (Thermo-Fisher) for PBMC with 22 hour runs per sample as shown in Figure 29. The Orbitrap settings were as follows: 30K resolution on full scans in Orbitrap, data-dependent MS/MS in LTQ (Top 10 peaks were chosen), two micro-scans for both full and MS/MS scans, centroid data for all scans.

PSM search and protein identification

All MS/MS spectra were searched against our customized sequence database.

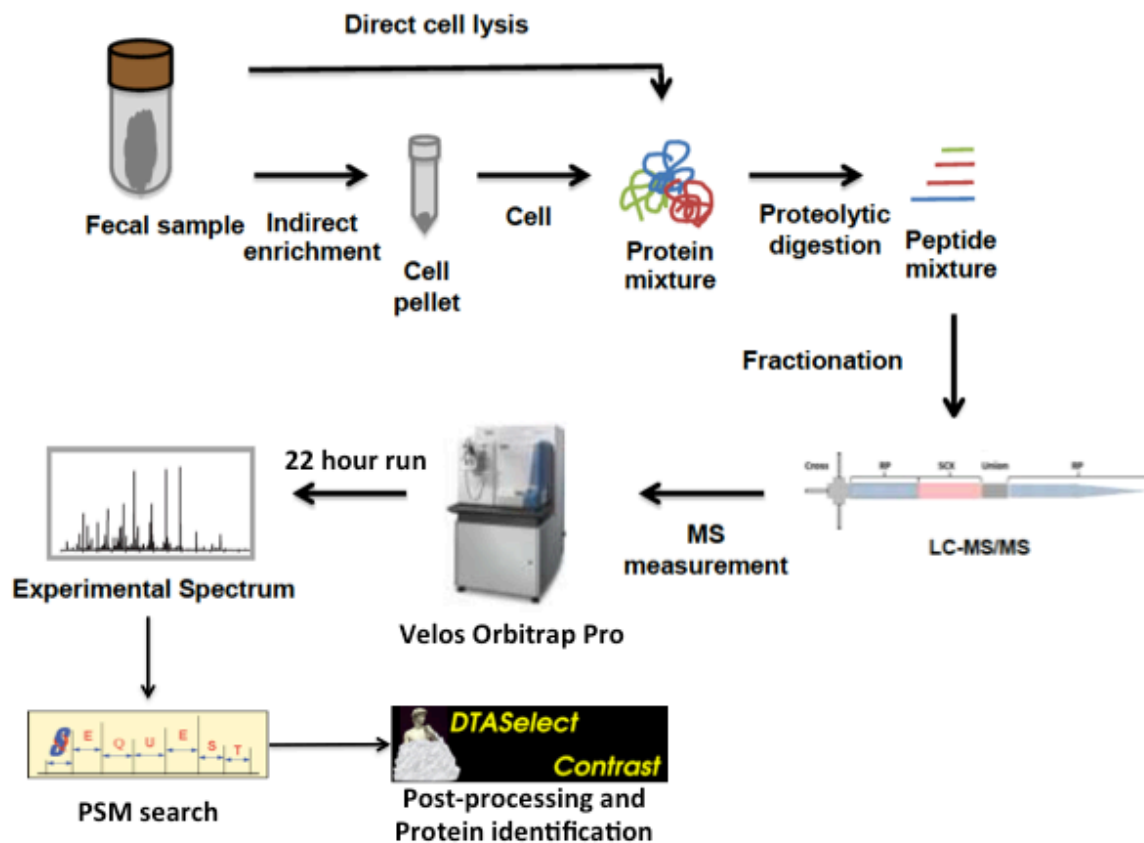


Figure 29 Experiment workflow of fecal metaproteome data generation.

This illustration was adapted from Xiong et al [143].

Our database consists of human protein sequences, microbial proteins isolated from human gut and common contaminants (i.e., trypsin and keratin; 36 protein sequences). We collected ~36,000 human protein sequences from NCBI RefSeq DB and microbial proteins databases from JGI IMG human microbiome project (<http://img.jgi.doe.gov/>) as is described in our previous paper [95]. For PBMC samples, we used human protein sequences only for the MS/MS search while microbiome proteins were included in the DB used for MS/MS search of fecal samples. All MS/MS individual runs were searched with the SEQUEST (v.27) algorithm [68] against our customized FASTA sequence database. Search parameters were followed to our prior study (<4 miscleavages, 3 Da mass tolerance window around the precursor ion mass and 0.5 Da for fragment ion masses). All SEQUEST output files were gathered and filtered using DTASelect (1.9) [86] at ≥ 2 peptides per protein and the following widely accepted parameters for all the MS runs: XCorr of at least 1.8, 2.5, 3.5 for +1, +2, and +3 charge states, respectively and a minimum ΔCN for 0.08. All the peptide spectrum matches (PSM) that could not satisfy a post-database search filter, $-10 \leq \text{ppm} \leq 10$, were excluded out to remove false positives as described in previous study [95]. In this study, only human proteins were analyzed, but we included microbial proteins in the FASTA database to remove false positives. We calculated spectral counts of human proteins via DTASelect. We removed redundant protein sequences in the result proteome using NCBI Entrez Gene [116] so that one gene symbol have one sequence. Spectral counts (SC) were normalized by scaling factor α_i for i -th protein described as below:

$$\alpha_i = \frac{\sum_j^N \sum_i^n SC_{ij}}{\sum_j^n SC_j}$$

Where N = the number MS runs, n = the number of proteins, SC is spectral counts of i-th protein in j-th MS run.

Statistical analyses

All the statistical analysis including hierarchical clustering were done by R version 3.1.0 and Python 2.7.3 in-house scripts. We used vegan, cluster, psych and gclus library of R to calculate distances between MS runs and draw dendrogram. To maximize clustering efficiency we calculated Cophenetic correlation coefficient and Gower similarity coefficient of several distance methods, such as single linkage, complete linkage, Unweighted pair-group method using arithmetic averages (UPGMA) and Ward's minimum variance matrixes. UPGMA matrix shows highest Cophenetic correlation coefficient (93.3%), so we used UPGMA for hierarchical clustering. Pearson correlation coefficients (PCC) between 29 MS runs were calculated by in-house R scripts. Shapiro-Wilk normality test was used to test our dataset if it is following Normality using shapiro.test function of R.

The proteomics results were also analyzed by using 'selfea' R package, described in chapter 3, providing quasi-Poisson Generalized Linear Modeling (QP) and Cohen's *f* [145]. We applied QP model to find significantly varied

proteins in 11 fecal proteomes. Poisson distribution is usually assumed to analyze frequency-based data like spectral counts [97]. However, Poisson distribution assumes that its parameter $\lambda = E(X) = \text{Var}(X)$. Because of the assumption, Poisson distribution cannot bear over-dispersion. Quasi maximum likelihood approach can alleviate Poisson test's less flexibility by employing variance function based on given data [97]. Our QP model is Spectral count ~ Fecal sample. P-value was calculated by QP model and adjusted by Benjamini-Hochberg method (FDR) [110]. Cohen's f was used as effect size filter. Cohen's f is ES linked with ANOVA F-test, which was used for QP model comparison. Important advantage of using Cohen's f is that we can filter out underpowered hypotheses tests. P-value cannot play its statistical evidence role at low statistical power (< 90%)[123,124]. Selfea provides minimum Cohen's f at 90% statistical power and reliable results using ES filter and FDR filter. Cohen's f is calculated as described below:

$$f = \sqrt{\frac{\sum_{i=1}^k p_i (\mu_i - \mu)^2}{MSE}}$$

Where, $p_i = N/n_i$, k = the number of groups, n_i = the number of observation in group i , N = total number of observations, μ_i = mean of group i , μ = grand mean over all the observations, and MSE = mean square error. The minimum Cohen's f at 90% statistical power was 1.3769 given sample size and p-value 0.05 threshold. We accepted proteins that have bigger Cohen's f than 1.3769 and less than 0.05 FDR for further analysis.

Functional analysis of fecal proteome was done by enrichment test and significance test. These analyses were done by in-house Python and R script. We annotated the dataset with GO terms [115] for human proteins and Cluster of orthologous genes (COG) terms [146] for microbial proteins. GO terms were downloaded from NCBI BioSystems database [106] via FTP. COG terms were assigned to fecal microbial proteins by using COGSoftware [120]. We followed steps recommended by COGSoftware to identify COG terms. With COG and GO terms, we conducted enrichment test using Fisher's exact test and significance test using selfea. Same filtering threshold, FDR < 0.05 and Cohen's $f > 1.3769$, were applied to find significant GO and COG terms.

Results

Data generation from 11 fecal and 5 PBMC samples

The 11 healthy volunteer fecal samples metaproteomic datasets were generated by nano 2D-LC-MS/MS on an OrbiVelos Pro (Thermo-Fisher). Of these 11 samples four were duplicated and others were triplicated, for a total of 29 MS run.. The datasets were searched against human NCBI RefSeq and 21 microbes isolates from human gut, downloaded from JGI IMG (<http://img.jgi.doe.gov/>), by using SEQUEST/DTASelect. From these 29 MS runs, we were able to identify 7,455 proteins, 2,102 human proteins and 5,353 microbiome proteins, after excluding out of redundant sequences using NCBI Entrez Gene DB (Table 14).

Table 14 Comparison of MS runs of fecal sample and PBMC. The number of identified proteins was counted for fecal human, microbiome and PBMC human data. Unshared proteins were uniquely identified in only one sample.

	<i>Feces</i>		<i>PBMC</i>
	<i>Human</i>	<i>Microbiome</i>	<i>Human</i>
Total number of proteins	2,102	5,353	2,556
The number of MS Runs	29	29	5
Average number of proteins detected/MS run	501	1,688	1,278
Average number of unshared proteins	115	213	158
% of unshared proteins	22.88%	12.61%	12.37%

These redundant protein sequences have same spectral counts (SC) due to short and redundant trypsin-digested peptides. SC of 7,455 proteins were normalized by scaled factor (α_i) that is average number of total SC per run (N) divided by sum of SC of run i (n_i) as is used in our previous study of Crohn disease [95]. This normalization makes 29 MS runs to have same total number of PSMs.

We also collected five peripheral blood mononuclear cell (PBMC) samples. PBMC is any blood cell that has a round nucleus, such as lymphocyte, a monocyte or a macrophage. We applied same pre-processing steps to five PBMC runs such as, normalization by total spectral counts and reduction of redundant protein sequences by NCBI Gene ID. Finally, 2,556 human proteins are identified from five PBMC replicates (Table 14). In terms of the number of identified human proteins, PBMC has two times more proteins per MS run than

feces, 501 in feces vs. 1,278 in PBMC, (Table 14). Unshared proteins are uniquely identified in only one sample. Percentage of unshared human proteins in fecal samples is almost two times that of fecal microbiome and PBMC human proteins (Table 14). More than 20% of human fecal proteins are uniquely identified from 11 healthy volunteers, which means that on average one fifth of human fecal proteins are not shared.

Comparison of 29 fecal and 5 PBMC MS runs

All the fecal samples were collected from healthy volunteers who do not have any of gastrointestinal disorders or other known disease states. To understand how their fecal proteomes are different, we first compared 29 MS runs in a broad sense using hierarchical clustering method (Fig. 30). Unweighted pair-group method using arithmetic averages (UPGMA) was used to calculate distance between MS runs because its Gower similarity coefficient is lower than other metrics, such as complete, average and single linkage metrics. The dendrogram of figure 30 shows that technical replicates of 11 fecal samples are clustered together (Fig. 30), which means technical replicates are not biased nor highly varied neither. Pearson correlation coefficients (PCC) between technical replicates confirm their similarity ranged from 0.75 to 0.98, which means closely correlated data (Fig. 31A).

In contrast to technical replicates' high similarity, all the individual donors of fecal proteomes are dissimilar with each other.

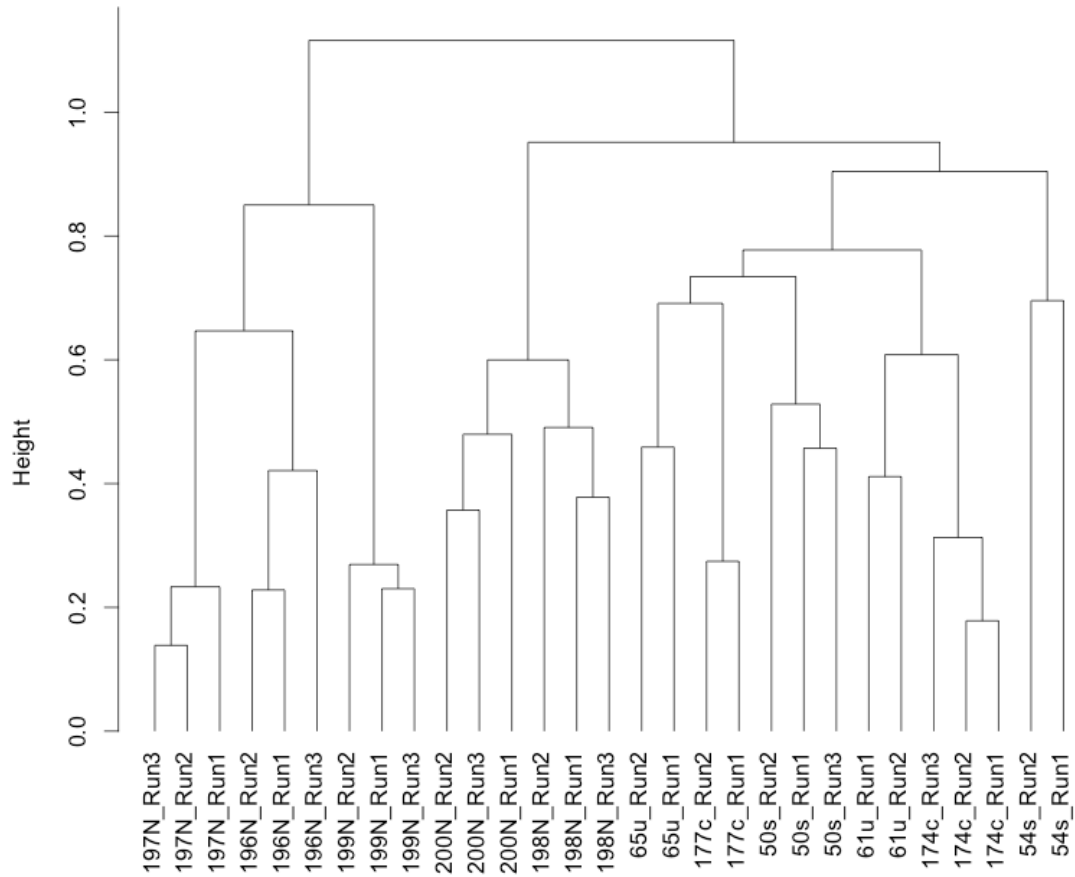


Figure 30 Hierarchical clustering of 29 MS runs of 11 fecal samples.

Technical replicates are closely clustered together, which means small variance within technical replicates.

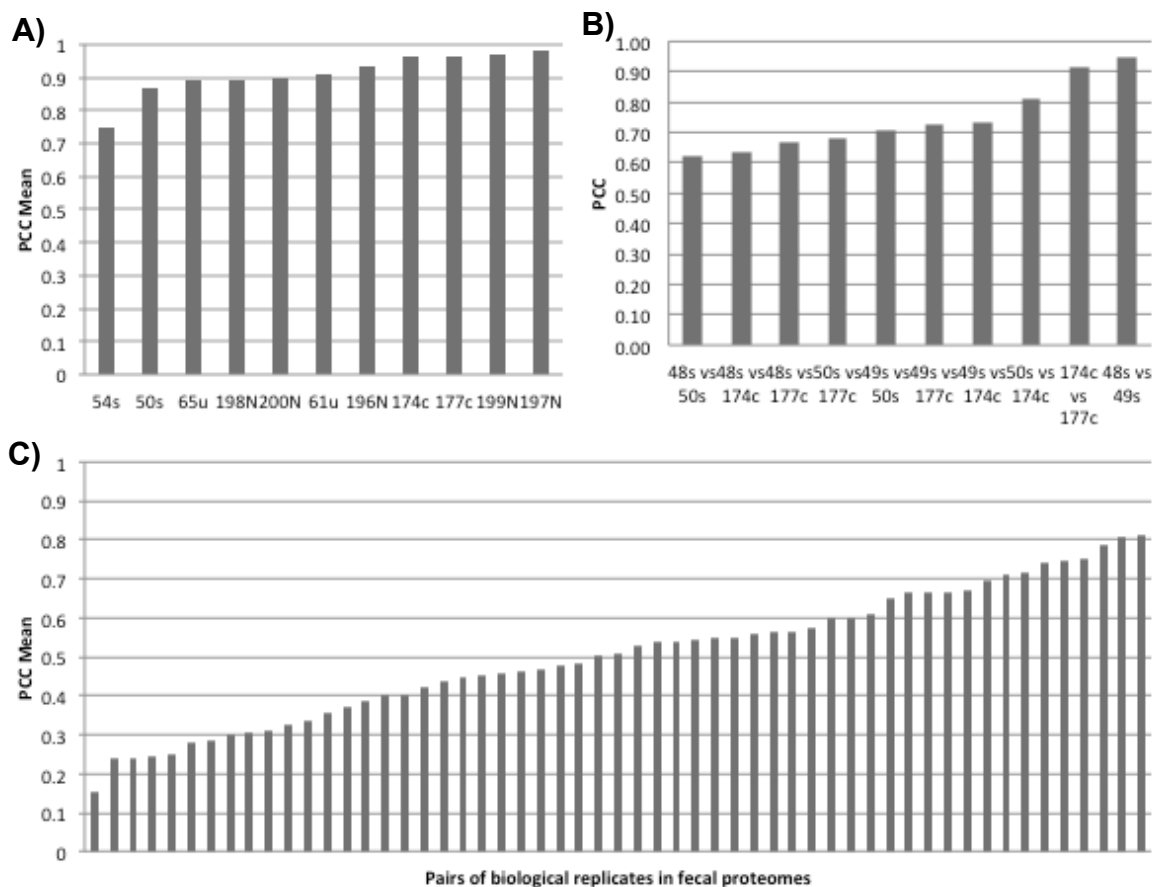


Figure 31 Correlation coefficients between fecal and PBMC MS runs.

Pearson correlation coefficients between A) technological replicates of fecal samples, B) biological replicates of PBMC samples and C) biological replicates of fecal samples.

In addition to hierarchical clustering result, we could confirm this dissimilarity from comparing PCC between biological replicate MS runs. PCC between biological replicates of feces are ranged from 0.15 to 0.81 (Fig. 31C). A difference between maximum and minimum PCC is 0.23 in technical replicates, but 0.66 in biological replicates. Moreover, minimum PCC in technical and in biological replicates was much different, 0.77 for technical replicates and 0.10 for biological replicates. Also, average PCC of biological replicates is 0.50 that is much lower than one of technical replicates (0.91). PCC in the biological replicates is relatively low and widely ranged. This shows the diversity between individual fecal proteomes. The highest PCC of biological replicates, 0.81, is similar to the lowest PCC of technical replicates, 0.77, showing dissimilarity of biological replicates and high similarity of technical replicates.

Comparing PCC of biological replicates between feces and PBMC proteome confirms that the heterogeneity of fecal proteome is especially high. PCCs between PBMC biological replicates are ranged from 0.62 to 0.95 (Fig. 31B). This range is much narrower and relatively higher than the fecal proteome (PCC: 0.15~0.81). The difference between minimum and maximum of PCC in PBMC is 0.33 that is half of the fecal proteome, 0.66. Also, minimum PCC of the fecal proteome, 0.15, is one fourth to one of PBMC, 0.62. All in all, we could see high variability in individual human fecal proteomes when compared to PBMC.

In addition to PCC comparisons, frequency of protein identifications within 11 fecal samples shows non-uniformity in protein identifications (Table 15). We counted how many times proteins were identified in 11 fecal samples.

Table 15 Protein identification frequency in fecal and PBMC proteomes.

<i>Protein Frequency</i>	<i>Fecal Human</i>		<i>PBMC Human</i>		<i>Fecal Microbiome</i>	
	<i># of proteins</i>	<i>% of frequency</i>	<i># of proteins</i>	<i>% of frequency</i>	<i># of proteins</i>	<i>% of frequency</i>
1	1,261	59.99%	947	37.05%	2,342	43.75%
2	287	13.65%	540	21.13%	738	13.79%
3	116	5.52%	342	13.38%	479	8.95%
4	86	4.09%	299	11.70%	315	5.88%
5	64	3.04%	428	16.74%	233	4.35%
6	32	1.52%	-	-	226	4.22%
7	30	1.43%	-	-	175	3.27%
8	32	1.52%	-	-	158	2.95%
9	37	1.76%	-	-	140	2.62%
10	54	2.57%	-	-	236	4.41%
11	103	4.90%	-	-	311	5.81%
Total	2,102	100.00%	2556	100.00%	5,353	100.00%

Surprisingly, of overall 7,455 proteins, the number of proteins with frequency 1 is 3,603, which means that almost half of total proteins are not redundantly detected in 11 healthy volunteers (Table 15). This percentage of unshared proteins is increased up to 60% in human proteins (Table 15). This 60% is much higher than one of PBMC, 37%. As protein frequency is increased, the number of proteins is drastically decreased in fecal samples, 60% to around 1%, when compared to PBMC samples, 37% to about 12% (Table 15). It is significant that almost 74% of human fecal proteins were discovered just once or twice in 11 fecal samples (Table 15). PCC comparison and identification frequency of fecal proteins shows its seriously high dissimilarity.

Human proteins in 11 fecal proteomes

Even though many of human fecal proteins are low abundant and unshared, about one hundred proteins were repeatedly and abundantly identified in 11 healthy volunteers (Table 15). Many of these high-spectral-count proteins are related to digestion, such as chymotrypsin like elastases (CELA3A and CELA3B), alpha-amylases (AMY1A, 1B, 1C and 2A), trypsins (PRSS1 and PRSS2) and chymotrypsinogen B (CTRB1 and CTRB2) (Table 16). Also, immune response related proteins are frequently spotted such as, IgG Fc binding protein, DBMT1 that is related to innate immunity [147,148], SERPINA3 that is related to inflammation [149] and PIGR that is polymeric immunoglobulin receptor precursor (Table 16). Polyubiquitin-C (UBC) is also in top-25 list and is related to innate immunity according to Gene Ontology DB (GO) [115]. Polyubiquitin B, isoform of UBC, is also discovered in all the fecal samples. Various myosin isotypes were found in fecal proteomes, such as myosin type 1, 2, 4, 6 and 8 (Table 16). Interestingly, transthyretin precursor (TTR) is copiously identified in 11 fecal samples. TTR is thyroid hormone-binding protein and is considered that it transports thyroxine from the bloodstream to the brain. TTR is found in various tissues including colon muscle according to Proteomics DB [31].

We searched top 25 abundant proteins of table 16 using Human Proteome Map (Fig. 32A) [150] and Proteomics DB (Fig. 32B) [31] to know where these came from. Human Proteome Map is a database that has proteins profiled from 17 adult tissues, 6 primary hematopoietic cells and 7 fetal tissues.

Table 16 Top 25 abundant human proteins in 11 fecal samples. “ * ”:

Immune response related proteins, “Frequency”: Identification frequency in 11 fecal samples of healthy volunteers and “Sum of SC”: total count of SC from 29 MS runs.

<i>Rank</i>	<i>Gene Symbol</i>	<i>Description</i>	<i>Frequency</i>	<i>Sum of SC</i>
1	CELA3A	Chymotrypsin-like elastase family member 3A preproprotein	11	7,562.19
2	*FCGBP	IgGfC-binding protein precursor	11	5,582.98
3	*DMBT1	Deleted in malignant brain tumors 1 protein isoform c precursor	11	4,793.61
4	SERPINA1	Alpha-1-antitrypsin precursor	11	4,089.48
5	AMY2A	Pancreatic alpha-amylase precursor	9	3,367.48
6	AMY1A	Alpha-amylase 1 precursor	9	3,225.87
7	AMY1B	Alpha-amylase 1 precursor	9	3,225.87
8	AMY1C	Alpha-amylase 1 precursor	9	3,225.87
9	CELA3B	Chymotrypsin-like elastase family member 3B preproprotein	11	3,176.00
10	AMY2B	Alpha-amylase 2B precursor	9	3,015.61
11	MYH2	Myosin-2	11	2,067.20
12	MYH1	Myosin-1	11	1,984.50
13	TTR	Transthyretin precursor	10	1,860.49
14	PRSS1	Trypsin-1 preproprotein	11	1,654.03
15	MYH4	Myosin-4	11	1,577.53
16	MYH8	Myosin-8	11	1,572.74
17	UBC	Polyubiquitin-C	11	1,568.63
18	PRSS2	Trypsin-2 preproprotein	11	1,178.51
19	MYH6	Myosin-6	11	1,134.14
20	MYH7	Myosin-7	11	1,126.16
21	CTRB2	Chymotrypsinogen B2 precursor	10	1,041.42
22	CTRB1	Chymotrypsinogen B precursor	10	997.56
23	*PIGR	Polymeric immunoglobulin receptor precursor	10	950.77
24	*IGLL5	Immunoglobulin lambda-like polypeptide 5	11	938.32
25	*SERPINA3	Alpha-1-antichymotrypsin precursor [Homo sapiens]	11	935.14

Figure 32 Expression location of top 25 abundant human fecal proteins.

Origin of fecal proteins were estimated by two databases, A) Proteomics DB and B) Human Proteome Map.

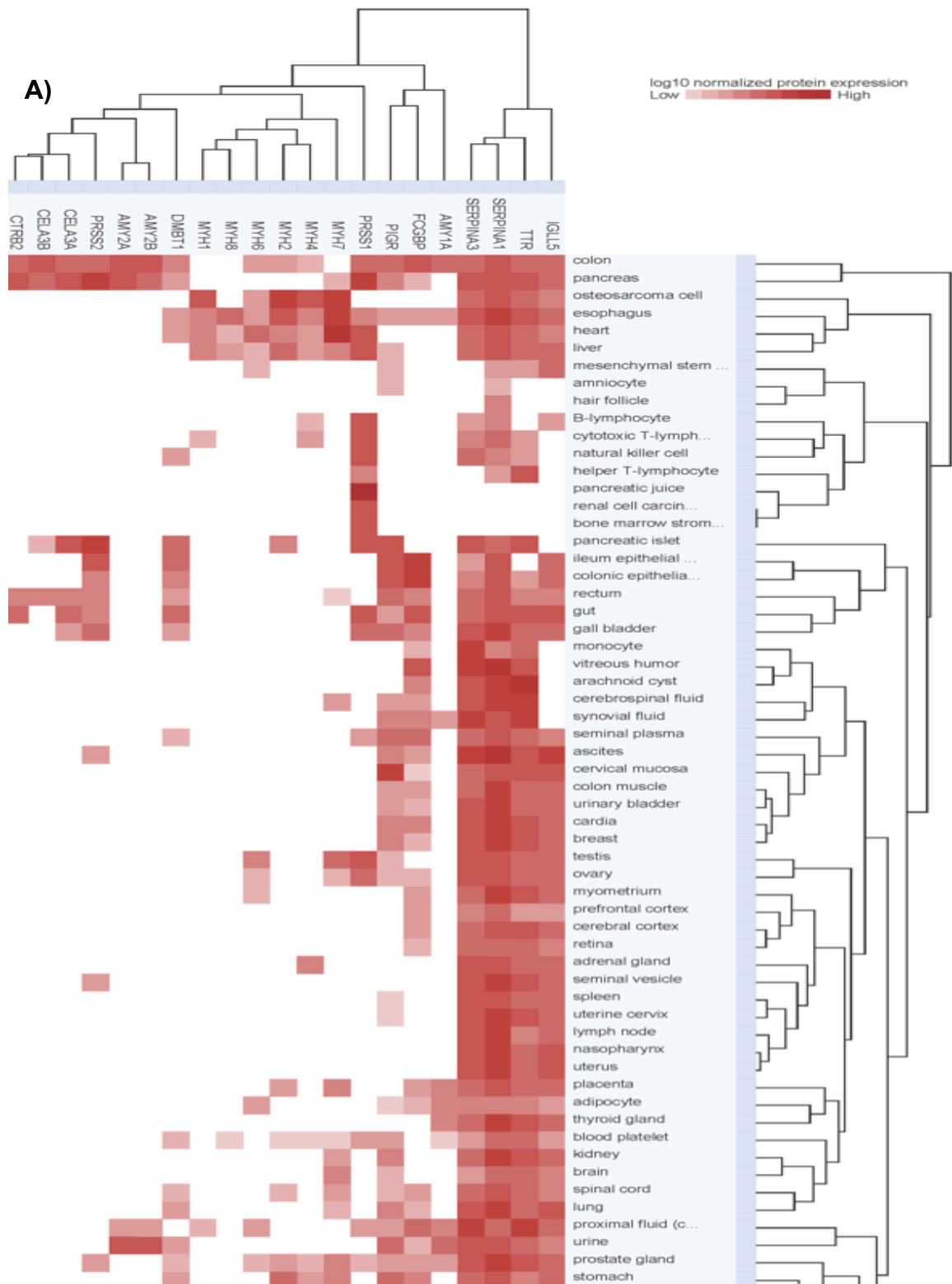


Figure 32 continued.

B)

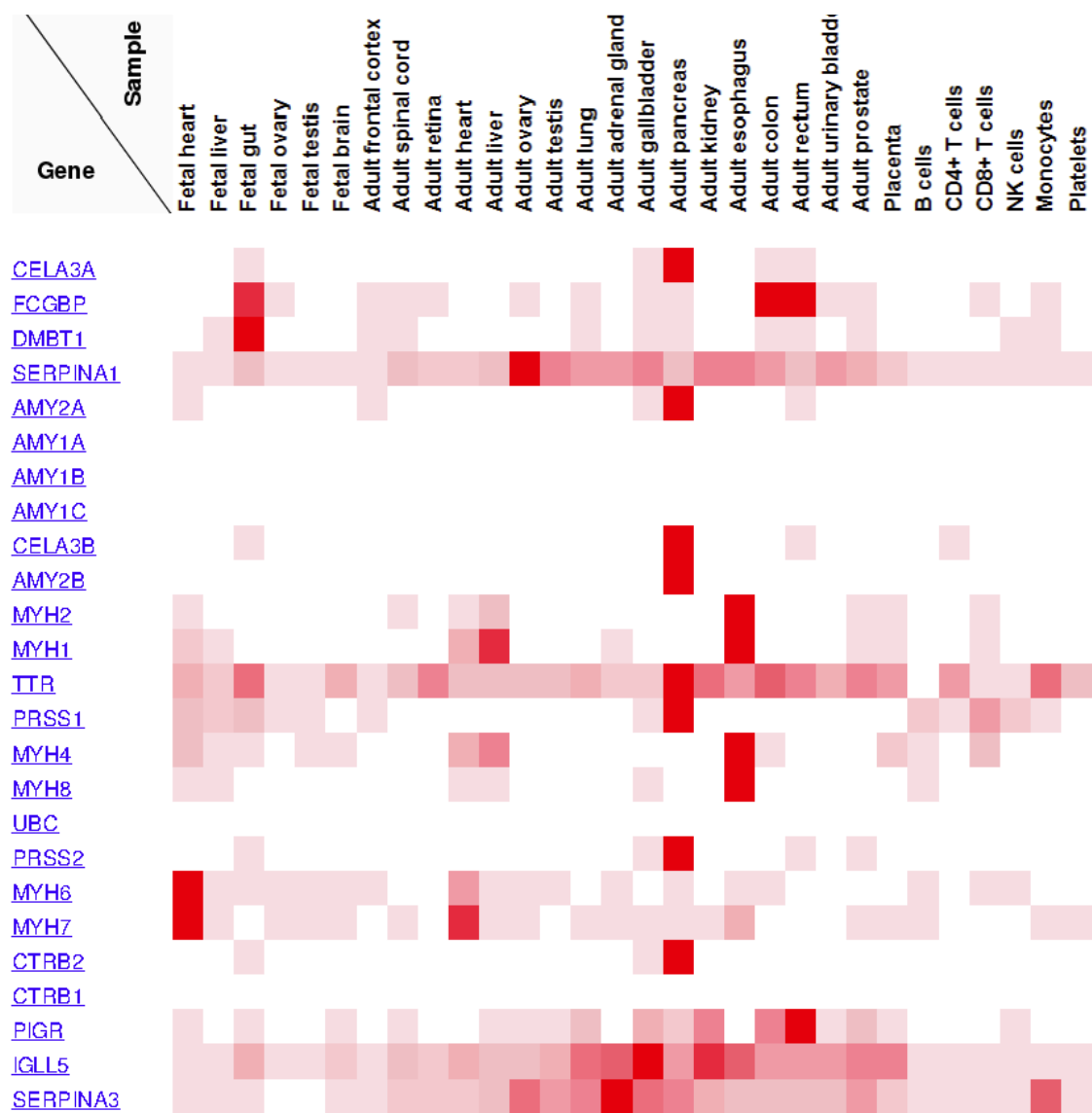


Figure 32 continued.

Proteomics DB is developed by Technische Universität München (TUM) and dedicated to deposit and share proteomics data to the scientific community. We compared our top-25 list to Proteomics DB through its provided analysis tool. The most frequently matched tissue in Proteomics DB is colon (22 proteins) and pancreas (19 proteins) (Fig. 32A), whereas is adult pancreas (16 proteins) and adult gallbladder (14 proteins) in Human Proteome Map (Fig. 32B).

All those tissues, frequently matched in both of Proteomics DB and Human Proteome Map are related with digestion. MYH1, MYH7 and MYH8 are not detected in colon tissue of Proteomics DB, but these are previously identified in esophagus among digestion related tissues. Therefore, we guess that MYH1, MYH7 and MYH8 come from esophagus unlike other isotypes of myosins. Many tissues involved in the digestion are found in the search result of Proteomics DB, such as colon muscle, saliva, salivary gland, rectum, pancreatic juice, pancreas, oral epithelium, esophagus and colon (Fig. 32B). Besides tissues on the digestion tract, immune response related cells are spotted in the result of Proteomics DB and Human Proteome Map (Fig. 32). Some of proteins in the top-25 list are reported that these are expressed in Natural killer cell, monocyte, helper T-cell, cytotoxic T-cell and B-lymphocyte according to previous studies of Proteomics DB (Fig. 32B). Human Proteome Map also shows similar result including B cells, CD4+ T cells, CD8+ T cells, NK cells and Monocytes (Fig. 32A).

On top of proteins related with digestion, proteins associated with immune response were also repeatedly identified in the 11 fecal samples. We expected that feces have abundant immune proteins because of innate and adaptive

immune responses against microbiome, pathogens or vaccines. Immune proteins were recognized by combining several annotation resources that are WikiGO [12], Immunome [151,152], Innate DB [153], 19 immune response GO terms [154] and immune response keyword search against protein description, such as immunoglobulin and IgA that were missed by other methods. Of 2,102 total human fecal proteins, 214 immune response proteins were discovered. We recognized frequently and abundantly identified 25 immune proteins in table 17. IgG_{FC}-binding protein precursor (FCGBP) is the most abundant protein in table 17. FCGBP is not directly related to immune response, but it is interacting with FC portion of IgG and indirectly involved in the mucosal immunity by making gel-like mucosa [155]. FCGBP also interacts with Mucin-2 (MUC2) [156] and is expressed at colon, colon epithelium, colon muscle and gut according to Proteomics DB (Fig. 32B). Some of inflammatory proteins, such as S100-A8, S100-A9, lactotransferrin (LTF) and Keratin type II cytoskeletal 1 (KRT1) were copiously detected in the fecal proteomes. LTF is antimicrobial protein by removing essential substrate required for bacterial growth and stimulate TLR4 signaling pathway [157,158]. KRT1 is also reported that it is involved in the inflammatory network in a skin by interacting with interleukin-18 [159]. S100A8 binds to S100A9 in order to be heterodimer named as Calprotectin that is secreted by Neutrophils and acts as important pre-inflammatory mediator in acute and chronic inflammation [160]. IGLL5 and IGJ are immunoglobulin subtypes, detected in the fecal proteome. IGLL5's function is not clarified while IGJ is known to interact with IGA and IGM to form polymers.

Table 17 Top 25 abundant immune response proteins in 11 fecal samples.

Frequency is identification frequency in 11 fecal samples of healthy volunteers.

“Sum of SC” is total count of SC from 29 MS runs. “Freq.” is identification frequency in 11 fecal samples. “Sources” is immune protein annotation source.

<i>Gene Symbol</i>	<i>Description</i>	<i>Freq.</i>	<i>Sum of SC</i>	<i>Sources</i>
FCGBP	IgGFC-binding protein precursor	11	5,582.98	Keyword
DMBT1	Deleted in malignant brain tumors 1 protein isoform c precursor	11	4,793.61	WikiGO, GO term, Innate DB
IGLL5	Immunoglobulin lambda-like polypeptide 5	11	950.77	Keyword
SERPINA3	Alpha-1-antichymotrypsin precursor	11	938.32	WikiGO
PIGR	Polymeric immunoglobulin receptor precursor	10	935.14	WikiGO, Keyword
IGJ	Immunoglobulin J chain	10	729.67	WikiGO, GO term, Immunome, Keyword
S100A9	Protein S100-A9	11	636.55	WikiGO, GO term, Innate DB
CEACAM5	Carcinoembryonic antigen-related cell adhesion molecule 5 preproprotein	11	414.82	WikiGO, Immunome
GP2	Pancreatic secretory granule membrane major glycoprotein GP2 isoform 4	10	414.06	GO term, Innate DB
CEACAM1	Carcinoembryonic antigen-related cell adhesion molecule 1 isoform 2 precursor	11	395.66	WikiGO, Innate DB, Immunome
LGALS3BP	Galectin-3-binding protein	11	369.90	WikiGO
DEFA1	Neutrophil defensin 1 preproprotein	11	293.57	WikiGO, GO term, Innate DB, Immunome

Table 17 continued.

<i>Gene Symbol</i>	<i>Description</i>	<i>Freq.</i>	<i>Sum of SC</i>	<i>Sources</i>
DEFA3	Neutrophil defensin 3 preproprotein	11	273.66	WikiGO, GO term, Innate DB, Immunome
LTF	Lactotransferrin isoform 1 precursor	7	227.05	WikiGO, GO term, Immunome
MPO	Myeloperoxidase	10	224.96	WikiGO, GO term, Immunome
LYZ	Lysozyme C precursor	9	214.56	WikiGO, Immunome
KRT1	Keratin, type II cytoskeletal 1	8	196.43	WikiGO, GO term
ALB	Serum albumin preproprotein	8	187.45	WikiGO
S100A8	Protein S100-A8	8	172.26	WikiGO, GO term, Innate DB, Immunome
CEACAM6	Carcinoembryonic antigen-related cell adhesion molecule 6 precursor	10	167.92	WikiGO, Immunome
DPP4	Dipeptidyl peptidase 4	6	150.79	WikiGO, GO term, Immunome
ANXA2	Annexin A2 isoform 1	9	142.79	Innate DB
ANPEP	Aminopeptidase N precursor	10	137.01	WikiGO, Immunome
DEFA5	Defensin-5 preproprotein	5	108.09	WikiGO, GO term, Innate DB, Immunome
CTSG	Cathepsin G preproprotein	8	101.60	WikiGO, GO term, Innate DB, Immunome

Even though IGA and IGM were not found in 11 fecal proteomes, we can speculate that there are IGA or IGM from IGJ identification. CEACAM5, CEACAM1 and CEACAM6 are another immunoglobulin superfamily called as CEA family. They are glycoprotein and their function is not clearly known, but CEACAM5 is reported that it is located on the cell surface and stimulates intracellular signaling by binding to adhesins of *Escherichia coli* [161].

Quasi-Poisson model construction using QFAM

After we confirmed variation of the fecal proteome by PCC and protein frequency, we constructed a linear regression model of the fecal proteins in order to find proteins significantly varied in 11 fecal proteomes. Since protein spectral counts of our proteome dataset deviate from Normality (Shapiro-Wilks normality test p-value = 2.2E-16) and are frequency-based data, we employed Poisson distribution via Generalized Linear Model (GLM) for hypothesis tests. Poisson distribution is usually assumed to analyze frequency-based data like spectral counts and shows better or equal performance for label-free proteomics data than other methods, such as Fisher's exact test, G-test, student's *t*-test [97,137,162]. However, Poisson distribution has less-flexibility to over-dispersion in terms of empirical fitting because of its assumption that its parameter $\lambda = E(X) = \text{Var}(X)$. GLM can handle non-normal response variables and Quasi maximum likelihood (QP) approach can alleviate Poisson test's less flexibility by employing variance function based on given data [97]. We used F-test to calculate p-values

that were adjusted by False Discovery Rate (FDR) method to correct multiple hypothesis testing [110].

In addition to QP model, we also calculated fold changes (FC) from all the possible pairs of 11 fecal samples and picked maximum-FC as representative FC of a protein because hypotheses test and fold change filtering can give us better selection power [126]. Interestingly, maximum-FC of all the 7,455 fecal proteins is over 2-fold change. Since many of fecal proteins are sparsely identified, many of proteins have infinite FC even though these are identified in only one fecal sample at low level of SC. Thus, we utilized another type of effect size, Cohen's f [145], to figure out how much different is reliable when considering sample size, significance threshold ($FDR < 0.05$) and desired statistical power (power $> 90\%$). Statistical power is a probability rejecting correctly null hypothesis when it is false. The certain level of power (90%) is required to correctly find significant proteins [123]. Cohen's f is designed for ANOVA's F-test and was used to keep our model to have enough statistical power, 90%, given samples size and significance level. Minimum Cohen's f for effect size filter was obtained by power analysis. The minimum Cohen's f for effect size filter is 1.3769 for 29 samples of 11 healthy volunteers for 90% power and 0.05 significance level. By combining QP and effect size filter, of total 7,455 fecal proteins, 1,610 proteins were categorized as differentially identified fecal proteins from 11 fecal proteomes (Fig. 33). Of significant 1,610 proteins, 355 proteins are from human and 1,255 proteins are from microbiome.

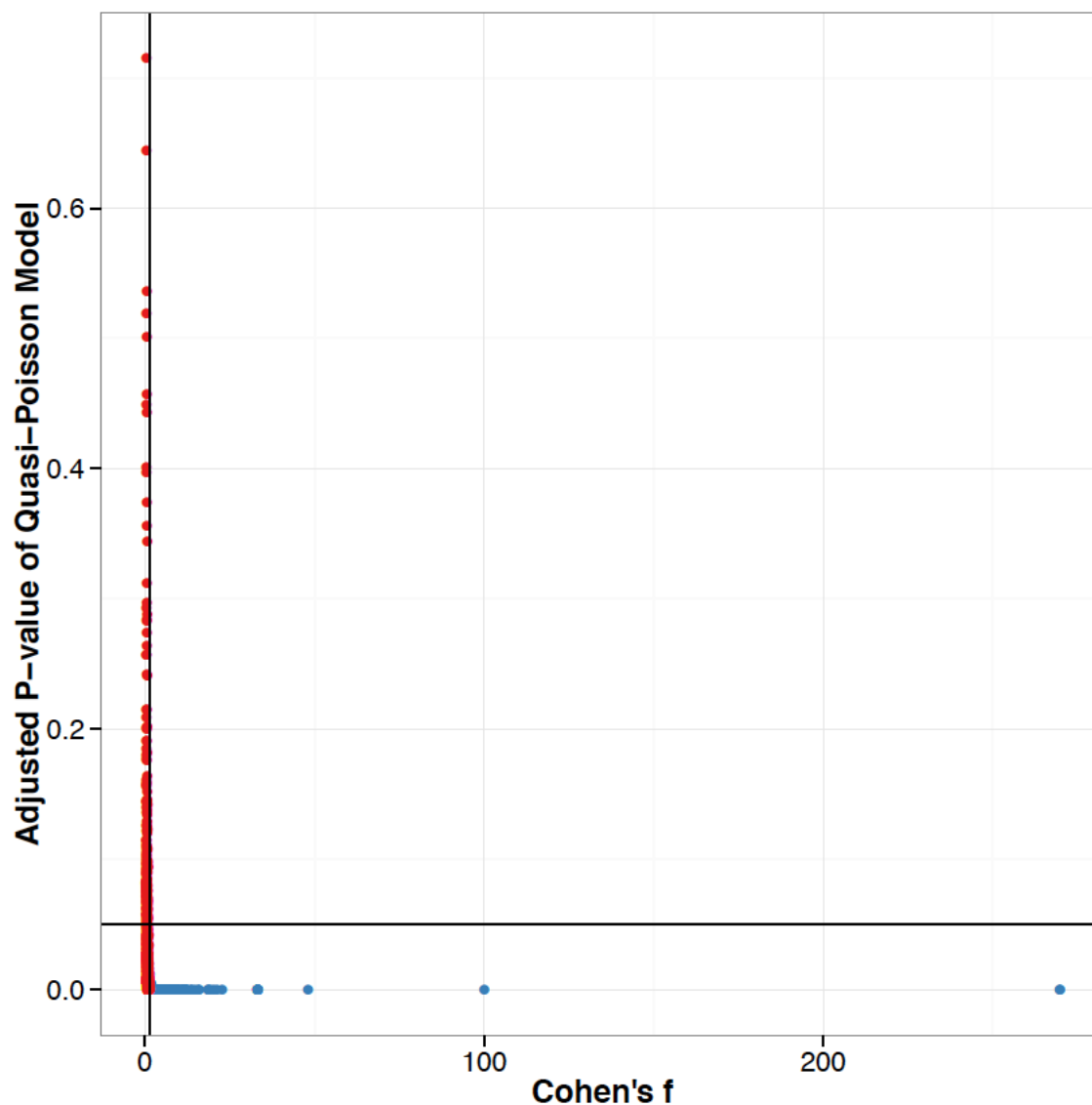


Figure 33 Protein scatter plot of human fecal proteome comparison. Blue dots are significantly categorized proteins, while red dots are proteins that could not passed two thresholds, $FDR < 0.05$ and Cohen's $f > 1.3769$.

More than 20% of total fecal proteins are significantly varied at 5% false positive level and 90% statistical power. This is not comparison between control and treatment group, but examining changes within biological replicates of healthy normal volunteers. Studies using fecal samples should consider high within-group variance. High variance within replicates affects on comparison between different experiment groups. To analyze highly varied data like fecal proteomes, many biological and technical replicates are required. Otherwise, strict threshold should be applied and reliable statistical methodology is required.

Functional analysis using QFAM

We analyzed fecal proteins in terms of functional terms, such as gene ontology (GO) [115] enrichment test for human proteins and cluster of ortholog groups (COG) [146] analysis for microbiome proteins, so that we understand metabolic functions of identified human fecal proteins. GO and COG terms are widely used to get to know functional features of proteomics data [163,164].

COG and GO terms were enriched over all the human fecal proteins by using Fisher's exact test. P-values of the enriched COG and GO terms were adjusted by FDR. Also, significance test was conducted by employing QP model and transferring SC of member proteins to COG and GO terms. Significantly enriched and varied COG and GO terms have less than 0.05 p-value (Fisher's exact test $FDR < 0.05$ and QP $FDR < 0.05$) and more than 1.3769 Cohen's f (statistical power $> 90\%$).

Table 18 Significantly varied GO terms in human fecal proteins. “GO ID”:

Gene ontology ID obtained from BioSystems DB, “FDR”: FDR-adjusted enrichment test, “Enrich. FDR”: FDR of enrichment test, “# of prots”: the number of member proteins and “QP FDR”: FDR calculated by QP model.

<i>GO ID</i>	<i>Description</i>	<i>Enrich. FDR</i>	<i># of prot</i>	<i>Cohen’ s f</i>	<i>QP FDR</i>
GO:0070062	Extracellular exosome	3.91E-98	691	4.4624	3.29E-10
GO:0065010	Extracellular membrane-bounded organelle	3.91E-98	691	4.4624	3.29E-10
GO:1903561	Extracellular vesicle	1.17E-97	691	4.4624	3.29E-10
GO:0043230	Extracellular organelle	1.17E-97	691	4.4624	3.29E-10
GO:0031988	Membrane-bounded vesicle	2.12E-87	765	4.7401	1.66E-10
GO:0031982	Vesicle	7.12E-84	770	4.7451	1.61E-10
GO:0044421	Extracellular region part	2.02E-76	771	5.0905	3.40E-11
GO:0005576	Extracellular region	3.67E-55	804	5.2393	1.67E-11
GO:0007586	Digestion	7.80E-07	39	3.8928	3.28E-11
GO:0002385	Mucosal immune response	7.08E-05	17	2.0228	1.94E-05
GO:0050832	Defense response to fungus	1.18E-03	13	4.1696	1.17E-08
GO:0009620	Response to fungus	8.80E-03	14	4.1802	1.03E-08
GO:0042742	Defense response to bacterium	3.15E-02	38	3.4573	4.33E-09
GO:0001906	Cell killing	3.39E-02	15	4.5767	2.52E-11
GO:0005856	Cytoskeleton	2.14E-40	418	4.2486	1.08E-09
GO:0071822	Protein complex subunit organization	8.08E-42	356	1.6116	1.44E-04
GO:0006936	Muscle contraction	1.91E-12	70	3.4826	2.61E-08
GO:0004022	Alcohol dehydrogenase (NAD) activity	4.42E-02	5	2.4551	1.16E-11

The result of GO term analysis includes interesting GO terms. Extracellular location related GO cellular component terms, extracellular vesicular exosome (GO:0070062), extracellular vesicle (GO:1903561), extracellular membrane-bounded organelle (GO:0065010), vesicle (GO:0031982), extracellular organelle (GO:0043230), extracellular region (GO:0005576) and extracellular region part (GO:0044421), were understandably enriched (Enrichment FDR $<1.07E-55$) from the significant human proteins (Table 18). This shows that many of the significant human fecal proteins are secreted into extracellular regions and excreted into feces.

Besides of extracellular location GO terms, digestion (GO:0007586) and mucosal immune response (GO:0002385) were enriched (Table 18). Mucosal immune response GO term was enriched from RYR1, TPM4, VCL, FLII, myosin proteins and actin proteins (FDR = $7.08E-05$). In addition to mucosal immunity, immune defense-related GO terms were enriched (FDR < 0.05), such as defense response to fungus (GO:0050832), response to fungus (GO:0009620), defense response to bacterium (GO:0042742) and cell killing (GO:0001906) (Table 18). Digestion related proteins, such as AMY2B, CTRB1, CTRB2 and CELA2A, are abundantly identified in the fecal proteomes (Table 17) and also included in the 355 significant protein list. Digestion proteins are required to obtain nutrition and energy for everybody, but abundance of those proteins is very dissimilar. For example, average spectral counts of digestion related proteins, assigned to digestion (GO:0007586), is 24.2904 in 174c while it is 289.1341 in 196N. The fold change between 174c and 196N is more than 10. Since cytoskeleton

proteins including keratins, tubulins, actins and myosins are significantly and differentially detected in the fecal proteomes, cytoskeleton (GO:0005856), protein complex subunit organization (GO:0071822) and muscle contraction (GO:0006936) are enriched with remarkable FDR ($FDR < 1.91E-12$) (Table 18).

We guess that these cytoskeletal proteins are excreted from the dead intestinal cells and immune cells, and their amount is different from person to person. Interestingly, alcohol dehydrogenase (GO: 0004022) was also differentially detected over healthy volunteers (Table 17). Its QP model FDR ($1.16E-11$) and Cohen's f (2.4551) shows how it was variously expressed from person to person.

Functional features of microbial proteins were analyzed by employing COG terms. We tested COG terms by enrichment test using Fisher's exact test and quantitative QP model with conferred SC from member protein to COG terms. Lots of ribosome related COG terms were frequently found in the COG analysis result, such as ribosomal protein L1 (COG00081), S2 (COG00052), S10 (COG00051), S11 (COG00100), L11 (COG00080), S12 (COG00048) and S19 (COG00185) at enrichment $FDR < 2.95E-09$ and QP $FDR < 0.0007$ (Table 19). Since ribosome is the most abundant protein in microorganism cells, many ribosome-related COGs were detected in the COG analysis. Different ribosome complex may mean different biomass of microbiome rather than expression level of ribosomal proteins. Energy-required (ATP or GTP) enzyme activity COG terms were also frequently enriched besides ribosomes (Table 19).

Table 19 Significant COG terms in microbial fecal proteins. “COG ID”: COG term ID assigned by COGSoftware, “FDR”: FDR-adjusted enrichment test, “Enrich. FDR”: FDR of enrichment test, “# of prots”: the number of member proteins and “QP FDR”: FDR calculated by QP model.

<i>COG ID</i>	<i>Description</i>	<i>Enrich. FDR</i>	<i># of prots</i>	<i>Cohen' s f</i>	<i>QP FDR</i>
COG00542	ATPases with chaperone activity; ATP-binding subunit	2.33E-13	36	3.1885	4.27E-07
COG00480	Translation elongation factors (GTPases)	5.32E-12	32	2.5506	9.50E-07
COG00050	GTPases - translation elongation factors	5.32E-12	26	1.5264	8.32E-05
COG00055	F0F1-type ATP synthase; beta subunit	1.42E-11	24	2.7013	1.14E-06
COG00465	ATP-dependent Zn proteases	8.86E-11	27	3.3691	5.58E-08
COG00100	Ribosomal protein S11	1.99E-10	21	2.0664	4.58E-06
COG00048	Ribosomal protein S12	1.99E-10	21	1.6682	0.000127452
COG00086	DNA-directed RNA polymerase; beta' subunit/160 kD subunit	1.99E-10	22	1.5654	0.000129515
COG00653	Preprotein translocase subunit SecA (ATPase; RNA helicase)	1.99E-10	23	1.4936	0.000254607
COG00459	Chaperonin GroEL (HSP60 family)	2.75E-10	22	2.9585	7.46E-08
COG00085	DNA-directed RNA polymerase; beta subunit/140 kD subunit	2.75E-10	22	1.9143	4.63E-06
COG01592	Rubryerythrin	7.43E-10	21	2.2773	1.10E-06
COG00057	Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase	7.43E-10	23	1.9770	1.11E-05
COG01629	Outer membrane receptor proteins; mostly Fe transport	7.43E-10	45	2.0563	2.90E-05

Table 19 continued.

<i>COG ID</i>	<i>Description</i>	<i>Enrich. FDR</i>	<i># of prots</i>	<i>Cohen' s f</i>	<i>QP FDR</i>
COG00052	Ribosomal protein S2	7.43E-10	20	1.5196	7.50E-05
COG00185	Ribosomal protein S19	7.43E-10	20	1.6309	0.000638
					967
COG00443	Molecular chaperone	8.88E-10	25	2.6864	7.01E-07
COG00080	Ribosomal protein L11	1.12E-09	20	2.0750	3.51E-06
COG00051	Ribosomal protein S10	1.12E-09	20	1.7237	4.86E-05
COG01454	Alcohol dehydrogenase; class IV	9.98E-06	21	1.9329	7.50E-05

ATPases with chaperone activity; ATP-binding subunit (COG00542), F0F1-type ATP synthase; beta subunit (COG00055), ATP-dependent Zn proteases (COG00465), Preprotein translocase subunit SecA (ATPase; RNA helicase) (COG00653), GTPases - translation elongation factors (COG00050) and Translation elongation factors (GTPases) (COG00480) were enriched at enrichment FDR < 1.99E-10 and QP FDR < 0.0003. Interestingly, alcohol dehydrogenate COG term (COG01454) was significantly and differentially enriched like GO analysis of human proteins at enrichment FDR = 9.98E-06 and QP FDR = 7.50E-05.

COG's broad functional terms were analyzed likewise to GO and COG terms (Table 20). We applied same thresholds (enrichment FDR < 0.05, QP, FDR < 0.05 and Cohen's f < 1.3769). As is confirmed by COG term analysis, translation, ribosomal structure and biogenesis (J) were significantly and differentially enriched (enrichment FDR: 7.00E-258 and QP FDR = 1.48E-10) (Fig. 34) (Table 20). Also, energy production and conversion (C), carbohydrate transport and metabolism (G), inorganic ion transport and metabolism (P), and amino acid transport and metabolism (E) were detected (Fig. 34) (Table 20). These are partially related to digestion and energy generation. These COG functions may be varied in different microbiome. Interestingly, cell motility and defense mechanisms (V) were found in the COG broad functional analysis result (enrichment FDR = 1.19E-07 and QP FDR = 2.95E-05) (Fig. 34) (Table 20). Microbiome is speculated that they may move to find nutrients and defend themselves from host immune system.

Table 20 Significant COG functional terms in microbial fecal proteins. “ID”:

COG functional term ID, FDR: FDR-adjusted enrichment test and # of prots: the number of member proteins.

<i>ID</i>	<i>Description</i>	<i>Enrich. FDR</i>	<i># of prots</i>	<i>Cohen' s f</i>	<i>QP FDR</i>
J	Translation; ribosomal structure and biogenesis	7.00E-258	1,060	4.1810	1.48E-10
S	Function unknown	1.44E-39	159	2.3167	3.51E-06
X	Not predicted	2.42E-37	205	2.3234	7.05E-07
C	Energy production and conversion	1.47E-34	476	2.2074	1.46E-06
G	Carbohydrate transport and metabolism	1.39E-31	626	3.1924	3.15E-09
O	Posttranslational modification; protein turnover; chaperones	1.88E-27	319	4.5972	1.48E-10
R	General function prediction only	1.10E-19	313	1.7545	3.43E-05
K	Transcription	6.68E-17	214	2.1878	2.77E-07
L	Replication; recombination and repair	2.51E-16	229	1.8058	2.76E-06
T	Signal transduction mechanisms	1.39E-09	94	1.5192	3.12E-04
V	Defense mechanisms	1.19E-07	70	2.1949	2.95E-05
M	Cell wall/membrane/envelope biogenesis	1.26E-05	226	3.7150	8.49E-09
E	Amino acid transport and metabolism	1.60E-05	471	2.2311	7.57E-07
N	Cell motility	5.57E-04	37	1.5247	9.60E-05
P	Inorganic ion transport and metabolism	3.44E-03	162	2.1932	1.45E-06

Figure 34 Line plot showing expression profile of 15 COG functions.

Spectrum counts of member proteins were transferred to COG broad functional terms. COG function name for corresponding symbol is in “ID” column of table 19. This plot was drawn by line plot module of QFAM.

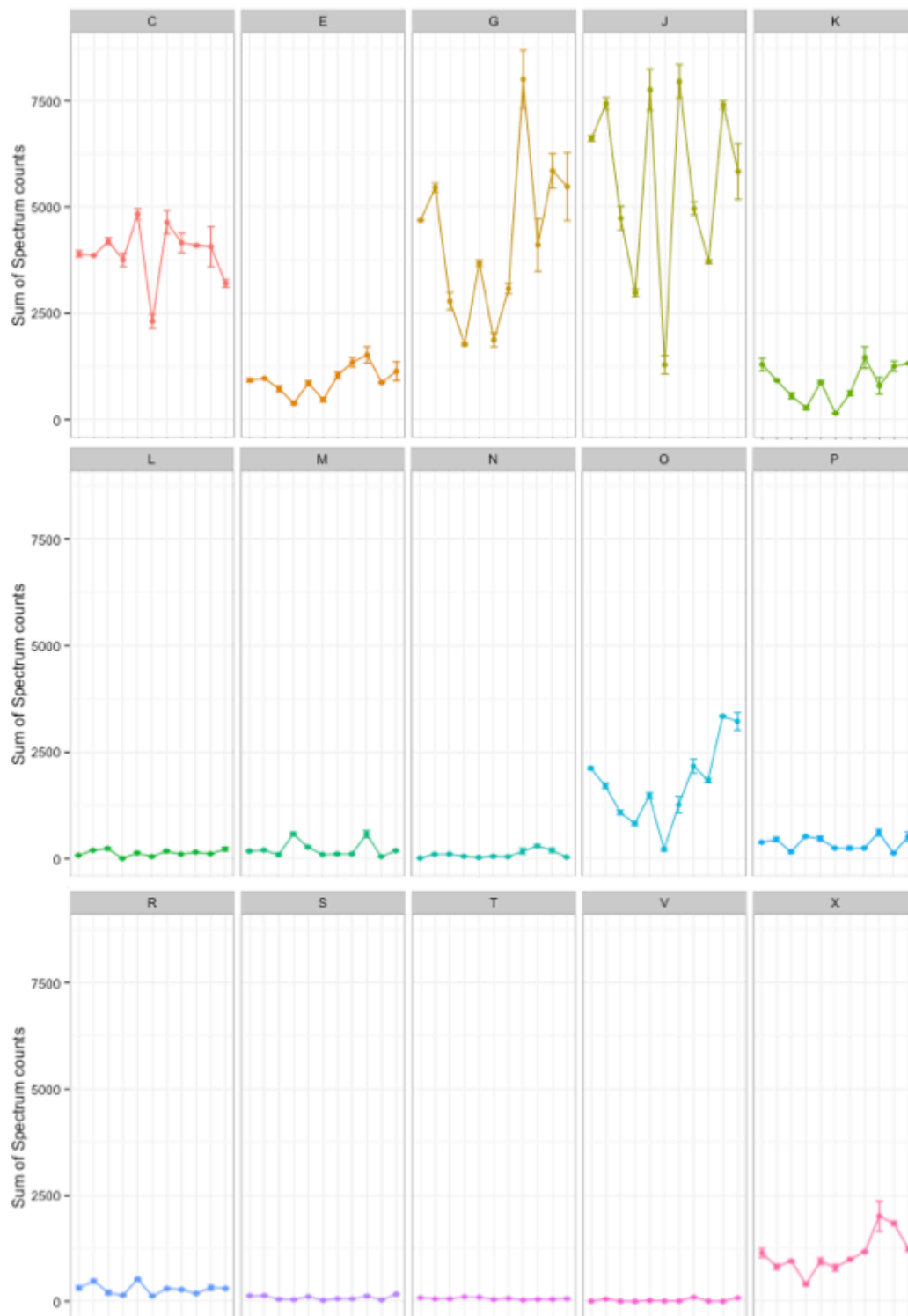


Figure 34 continued.

These functions also were dissimilar in different people, which mean that microbiome may be facing in different environment (human gut). Table 20 has statistics of significant COG broad functional terms whose expression profiles are described in figure 34. Figure 34 was produced by plot library of QFAM.

Conclusion

We could identify 2,102 fecal human proteins and 5,453 microbial proteins from feces of 11 healthy volunteers and confirmed their heterogeneity. This number of human fecal proteins is larger than previous studies, 234 proteins from three samples in Lichtman et al [17] and 446 proteins from two samples in our previous study [148]. We identified 501 proteins per a MS run that is larger number than total number of proteins from two or three samples in the previous studies. We confirmed that host proteins in feces are diverse in human as is shown in mice feces by Lichtman et al [17]. Host fecal protein's variance was disclosed by PCC (Fig. 31) and identification frequency (Table 15) comparisons with technical replicates and five PBMC proteomes. Figure 31 shows that fecal proteins are not uniformly identified in individual fecal samples. The QP model also reveals that expression levels of fecal proteomes are various. We could find 1,610 proteins that are differentially and significantly detected in 11 fecal proteomes via QP model and effect size filter. We utilized Cohen's f to enhance statistical power of our QP model. Functional enrichment analyses over the 1,610 significant proteins show that human significant proteins are generally located extracellular regions and are related with digestion, cytoskeleton and

protein metabolisms, and microbial significant proteins are related with translation, energy generation and protein metabolisms.

Many immune response proteins were also identified in our fecal proteome. Identification of immune response proteins in fecal proteomes is a natural outcome because the gastrointestinal tract is a direct interface of interaction between host immune system and microbiome including commensal flora and enteric pathogens. We found 214 immune proteins out of 11 fecal proteomes. Almost 40 immune proteins were detected in more than half of fecal samples (> 5 samples), while others are less-frequently identified. Toll-like receptor family protein (TLR) plays a key role in the innate immunity, so we expected TLR proteins in many fecal samples. However, only TLR6 was weakly identified in one fecal sample. TLR6 has a fundamental role in pathogen recognition and induction of innate immune response. We speculate that rare identification of TLRs is because of their low expression level in cells. We also expected to see immunoglobulins for innate immunity, such as IGA, IGD, IGE, IGG and IGM, but none of these are in our fecal proteomes. In place of above five immunoglobulins, IGJ are strongly and frequently identified. IGJ is involved in polymer form of IGA and IGM, so IGJ can be an indirect clue of existence of IGA and IGM. IGJ is also identified in previous studies, but TLR6 is found only in our study.

Human fecal proteins that were identified by previous studies were confirmed in our study. We could confirm that 196 proteins of 234 in Lichtman et al [17] and 243 proteins of 446 in our previous study were identified in our study.

The number of fecal proteins commonly identified in three studies including our fecal proteomes is 101. Of the commonly identified 101 proteins, 24 proteins were immune response proteins including IGJ and DMBT1. Digestion related proteins, such as, AMY2A and PRSS1, are also commonly found. Many of the simultaneously found proteins are detected with high SC in our study. Top 25 abundant fecal proteins in table 16 are found in at least one of previous studies.

CHAPTER V

APPLICATION CASE 2 – HUMAN DIET PROTEOME STUDY

Introduction

What we are eating undoubtedly has a great impact on the relative abundance and the types of our microbiome living in gastrointestinal tract [165]. For example, western diets, low in carbohydrates and high in simple sugar and fat, are associated with the decrease of *Bacteroidetes* in North Americans [166,167]. This change in gut microbiome is generally detected in feces [168]. For instance, certain type of bacteria are increased or decreased by short-term dietary changes, such as resistant starch (RS) diets and antibiotics, in the collected fecal samples. Antibiotics, effective against autochthonous species members, such as *Helicobacter pylori* and *Lactobacillus Reuteri*, create significant decrease of these microbes detected in the feces [168]. The microbial pattern changed by different RS amount in diets is likely associated with how long the changed diet is taken and its specific nutritional contents, such as more refined RS [169-171].

RS is one of starch or derivatives of starch and is not fully digested by human small intestine, while sugars and most starches are rapidly digested and absorbed as glucose. This undigested RS, dietary fiber, is diet-derived energy-source of microbiome living mostly in large intestine. Several scientific reports showed that high RS diet could result in abundance of specific bacteria in microbiome, such as *Ruminococcus bromii*, *Eubacterium rectale*, and

Bifidobacterium adolescentis [165,171-173]. These changes in microbiome are probably caused by substrate-specific preference of different microbes that can uptake and degrade RS efficiently. Generally, people inhabiting different places have dissimilar microbiome, but children in European Union and Burkina Faso showed similar response to high fiber diets, which *Bacteroidetes* and *Firmicutes* ratio are changed by different amount of dietary RS [174]. Potential increase of short chain fatty acid and butyrate-producing microbes and decrease of disease-causing bacteria enable us to associate RS diet with human health.

While the influence of resistant starch on the gut microbial community has been demonstrated, the precise functional and quantitative relationship between the gut microbiome and RS amount remains unclear. This chapter focuses on the metaproteomics apart of the integrative “omics” study including metagenomics, metabolomics and metaproteomics. Two different diets containing high and low amounts of RS were treated to 8 patients in crossover time course design. Feces after having specific diets were sampled and analyzed by 2D LC-MS/MS and QFAM.

Methods

Generation of Metaproteome data

The diet study design is summarized in Figure 35. Briefly, 8 patients who were insulin resistant (HOMA-IR > 50th percentile for sex), overweight or obese (BMI between 27 and 35 kg/m²), and met other eligibility criteria were enrolled in the study.

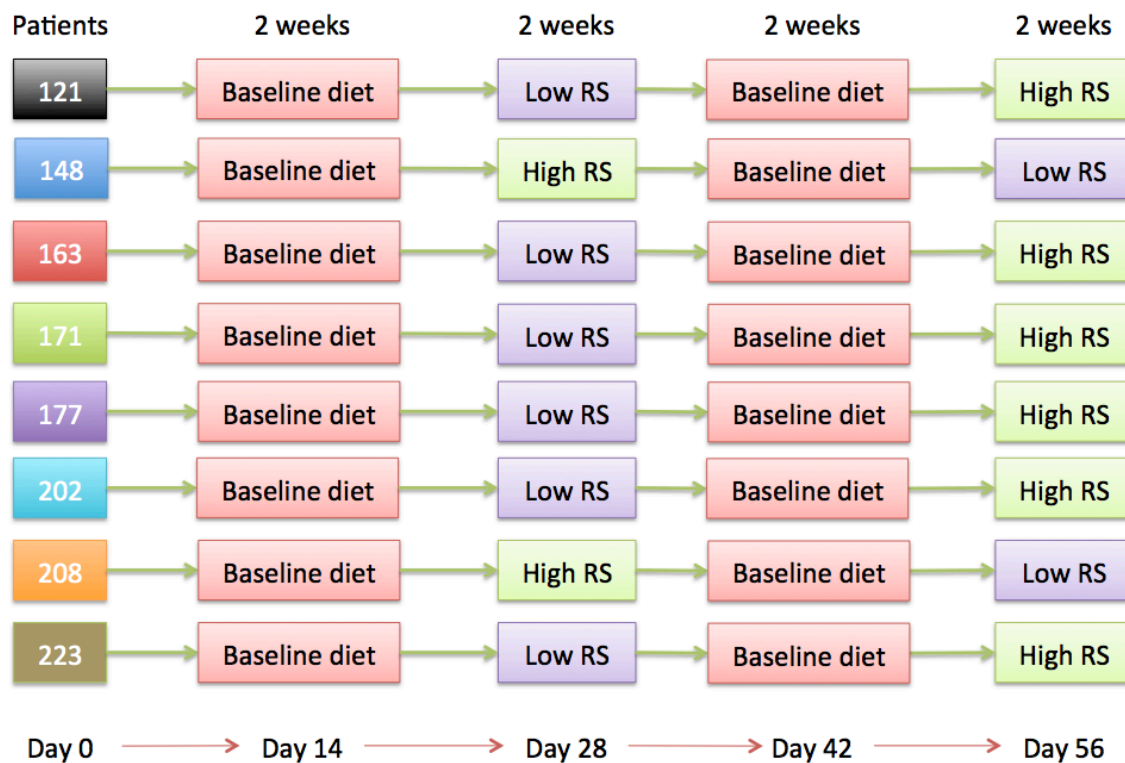


Figure 35 Experiment design of diet proteome study. Low RS has 2g amylose and High RS has 38g amylose. Every diet is treated for two weeks. Total 8 patients had different diets, as is scheduled in cross over design.

All subjects were otherwise healthy and not taking any antibiotics, lipid-reducing or blood-thinning medications, or hormones. The study protocols were approved by the Human Subjects Committee of both Children's Hospital Oakland Research Institute and Lawrence Berkeley National Laboratory. Written informed consent was obtained from each subject. Study participants were randomly assigned to a sequence of the two experimental diets: a low resistant starch (LRS) diet containing 2 grams of resistant starch and a high resistant starch (HRS) diet containing 38 grams of resistant starch. Each diet period was 2 weeks with 2 weeks baseline washout diet between diet periods (Fig. 35). In both diets, type 2 RS, a granular form of high amylose cornstarch, was used. Melojel of Ingredion Inc. was used for LRS diets, while Hi-maize 260 RS was used for the HRS diets. Following the initial baseline diet and each diet period, fecal samples were collected.

Sample preparation and LC-MS/MS experiment

Proteins of all fecal samples were extracted by indirect cell lysis method (Fig. 29) preferred for collection of microbial proteins. Proteins of all the samples were digested by trypsin and analyzed via 2D LC MS/MS with C18-SCX (Phenomenex) self-packed nano back column (3 cm C18, 3 cm SCX, 150 μ m i.d.) that serves as the first dimension of the 2D-LC system to capture peptides and wash away salts and a QExactive MS (Thermo Fisher). QExactive settings were as follows: the normalized collision energy for HCD was 28 eV, a full scan resolution of 70,000K from 400-1600 m/z, a HCD MS/MS resolution of 17,500

with an isolation width of 3 m/z, and the dynamic exclusion was set at 15 seconds. Peptides were not excluded based on charge state and 1 microscan for both full and MS/MS scans were acquired. All MS and MS/MS data were acquired in profile mode.

Quantitation and normalization of metaproteome data

All MS/MS spectra were searched against our customized sequence database. The database consists of human protein sequences, translated metagenome sequences, proteins of 34 human-gut-isolated microbial species, and common contaminants (i.e., trypsin and keratin; 36 protein sequences). We collected human protein sequences from NCBI RefSeq and microbial proteins from JGI human microbiome project (IMG-HMP) as is described in our previous paper [95]. All MS/MS individual runs were searched with the SEQUEST (v.27) algorithm against our customized FASTA sequence database. Search parameters were followed to our prior study [95] (<4 miss-cleavages, 3 Da mass tolerance window around the precursor ion mass and 0.5 Da for fragment ion masses). All SEQUEST output files were gathered and filtered using DTASelect (1.9) at ≥ 2 peptides per protein and the following widely accepted parameters for all the MS runs: XCorr of at least 1.8, 2.5, 3.5 for +1, +2, and +3 charge states, respectively and a minimum ΔCN for 0.08. All the peptide spectrum matches (PSM) that could not satisfy a post-database search filter, $-10 \leq \text{ppm} \leq 10$, were excluded out to remove false positives as described in our previous study [95]. Finally 57,397 proteins were quantitatively identified (human: 1,103 and

microbes: 56,294). Spectrum counts (SC) of protein is normalized as described below [95]:

$$Normalized\ SC_i = \frac{\sum_{j=1}^M \sum_{k=1}^N SC_k}{\sum_{k=1}^N SC_k} \times SC_i$$

Where, N = the number proteins and M = the number MS runs.

Quantitative analysis of metaproteome data

All the statistical analyses were fulfilled by using R 3.1.3 and Python 2.7.6. Since people have different gut microbiome and different genetic background, fecal proteomes are heterogeneous from person to person. Chapter 4 study showed that human fecal proteome is highly varied. Hence, we considered that personal variation is important to interpret our proteomics data in terms of diet type. Generalized Linear Mixed Model (GLMM) is the expanded model of generalized linear model (GLM) and able to include random effects in addition to the usual fixed effects. We set personal variation as random effect to explain different people's different fecal proteomes. Our model for the proteome analysis is $SC \sim \mu + \text{Diet type} + (1|\text{Person})$. Negative binomial distribution (NB) was employed as null distribution in GLMM instead of Normal distribution because spectral count is frequency-based data and NB distribution is generally used for count-type data analysis, such as read numbers of RNA sequencing. The constructed GLMM NB model was compared to null model through chi-square likelihood ratio test (LR test). All the P-values were adjusted by FDR in order to

reduce multiple comparison effect. Cut-off FDR is 0.0001. We used 'lmerTest' package to construct GLMM NB model [175].

On top of GLMM NB model, effect size (ES) filter was applied to enhance reproducibility. Combining p-value and ES filter can enhance reproducibility of proteomics data analysis [128]. As is showed in chapter 3, fold change is widely used ES, but Cohen's ES is better ES in terms of statistical power and selection power filtering false positives. We used 'selfea' R package, described in chapter 3, to calculate Cohen's w that is designed for chi-square test that was used for GLMM NB model comparison. Since a number of microbial proteins were significantly detected, we used very strict statistical filters (FDR-adjusted p-value (FDR) < 0.0001) to make the expected number of false positives is smaller than 1. When p-value threshold is 0.0001, minimum Cohen's w for our data was 1.0258 corresponding to 80% statistical power that is recommended level for reliable hypothesis test. Statistical power is the probability of correctly rejecting null hypothesis when it is false, so it is directly related statistics with reproducible result. The expected number of false positives is 0.7 at 0.0001 FDR and 80% power. The number of significantly detected human proteins is 94, while the number of microbial proteins is 7,018.

Functional analysis of metaproteome data

Metaproteome data was functionally analyzed by using COG for microbial proteins and KEGG pathway terms. COG and COG functional terms were examined by using COG-analysis pipeline of FAM. To assign COG ID to

microbial proteins, we used COG-software by following steps suggested by Kristensen et al [120]. For proteins, COG-software could not assign, PSI-BLAST search against COG sequences was used with cut off E-value 0.00000001. We could assign COG ID to 54,290 microbial proteins. To quantitatively analyze COG IDs, we transferred SC of member proteins, allocated to at least one COG ID, to COG ID. Same assignment was done to 25 broad COG functional terms. We applied GLMM NB model and Cohen's w to get significant COG terms and their broad functional terms. For COG analysis, 0.0001 FDR thresholds and 1.0258 Cohen's w were applied. We could examine significantly changed 400 COG terms and 4 COG functional terms across different diet types. Since many of significant COG terms were supported by only 1 or 2 member proteins, we filtered out COGs by the number of member proteins (member > 2). This rule was applied to COG functional terms. After filtering by minimum member proteins, remaining COGs were 166 terms and 2 functional terms. For COG functions, we applied 0.001 FDR of GLMM-NB model and 0.9051 Cohen's w (the number of expected false positive is 0.02).

FAM was used to analyze KEGG pathway for human proteins in diet metaproteome. SC transfer to KEGG pathways was done like COG analysis. Enrichment test was carried out by QAM module. FDR threshold for the enrichment test was 0.05. To quantitatively analyze KEGG pathways, we applied GLMM NB model and Cohen's w , 0.05 FDR thresholds and 0.6336 Cohen's w , so that the number of expected false positives is 1. We could identify 20 significant KEGG pathways.

Identification of translated-metagenome sequences

Undefined translated-metagenome sequences (metagenome proteins) were defined by BLAST search pipeline of QFAM against whole UniProt DB downloaded in Dec. 2014. BLAST search, using BLAST 2.2.24, was fulfilled in Newton cluster of the University of Tennessee (<https://newton.utk.edu/>). We divided 48,284 metagenome proteins into 48 groups and had run 48 BLAST jobs simultaneously onto 48 cores of Newton cluster. Sequence similarity was calculated for every alignment as described below:

$$Similarity = \frac{Aligned\ length}{Protein\ length} \times Percent\ identity$$

Where aligned length is value of 'alignment length' of BLAST result, protein length is the number of amino acid of a protein and percent identity is '% identity' of BLAST result. We assigned single UniProt ID to a metagenome protein by 95% sequence similarity without gap opens. At 95% similarity level, 38,543 metagenome proteins were successfully determined.

Results

Generation of diet metaproteome data

From 24 MS runs (8 patients * 3 diets), 57,398 proteins were determined. Of 57,398 proteins, 56,295 were from microbiome and 1,103 were host human proteins. This identification bias is caused by indirect protein extraction method (Fig. 29). Since we collected microbial pellets and removed supernatant, human proteins, floating in supernatant, were efficiently removed. Filtering out of host

proteins are required to get more microbial proteins from fecal samples. This diet proteome study targets understanding of changes in microbiome communities, so half number of human proteins was identified when compared to human fecal proteome study of chapter 4.

Much larger number of microbial proteins was detected in this study than human fecal proteome study (56k of diet proteome vs. 5k of human fecal proteome). This abundant microbial protein examination is due to two reasons. First, it is employing indirect protein collection method as is described in chapter 4. Another factor is the metagenome sequencing. As is explained in chapter 4, species composition of human feces is not known and is various from person to person. In this study, metagenome sequencing was employed and added to PSM search, while it wasn't in human fecal proteome study. Since metagenome is not annotated, metagenome characterization is required. Nevertheless, microbial protein identification is dramatically increased by metagenome sequences and indirect method. The number of metagenome proteins, translated metagenome coding genes, is 48,376 of 56,295 total microbial proteins. This is 86% of total microbial protein identifications, which means showing huge contribution of metagenome proteins to increase of microbial proteins.

Some species had significant changes in terms of the number of identified proteins (Table 21). The number of metagenome proteins was increased by RS amount (26,075 in LRS and 28,776 in HRS).

Table 21 The number of identified proteins in different species and in different diets. “SD” is standard deviation of numbers in three different diets.

<i>Species</i>	<i>Baseline</i>	<i>LRS</i>	<i>HRS</i>	<i>SD</i>
<i>Aeromonas hydrophila hydrophila</i> ATCC 7966	4	4	5	0.58
<i>Desulfovibrio</i> sp. 3_1_syn3	13	13	11	1.15
<i>Escherichia coli</i> CFT073	12	10	8	2.00
<i>Lactobacillus reuteri</i> CF48-3A	6	2	4	2.00
<i>Enterococcus faecalis</i> PC4.1	11	10	7	2.08
<i>Shigella flexneri</i> 2a 2457T	11	10	7	2.08
<i>Methanobrevibacter smithii</i> F1, DSM 2374	1	1	6	2.89
<i>Salmonella enterica enterica</i> sv Typhi Ty2	13	8	8	2.89
<i>Blautia hydrogenotrophicus</i> S5a33, DSM 10507	72	66	66	3.46
<i>Klebsiella</i> sp. 1_1_55	20	9	15	5.51
<i>Streptococcus</i> sp. 2_1_36FAA	21	14	8	6.51
<i>Coprococcus</i> sp. HPP0048	75	61	74	7.81
<i>Anaerostipes</i> sp. 3_2_56FAA	58	67	74	8.02
<i>Ruminococcus lactaris</i> ATCC 29176	87	103	89	8.72
<i>Prevotella melaninogenica</i> D18	64	82	71	9.07
<i>Clostridium</i> sp. 7_2_43FAA	45	31	50	9.85
<i>Dorea formicigenerans</i> ATCC 27755	93	104	84	10.02
<i>Coprococcus comes</i>	77	81	98	11.15
<i>Lachnospiraceae</i> bacterium 6_1_63FAA	90	89	115	14.73
<i>Alistipes putredinis</i> DSM 17216	130	157	161	16.86
<i>Dialister invisus</i> DSM 15470	23	47	57	17.47
<i>Roseburia intestinalis</i> L1-82	104	145	111	21.93
<i>Akkermansia muciniphila</i> ATCC BAA-835	14	11	58	26.31
<i>Eubacterium rectale</i> ATCC 33656	263	311	325	32.52
<i>Bifidobacterium adolescentis</i> ATCC 15703	41	38	102	36.12
<i>Ruminococcaceae</i> bacterium D16	75	147	126	37.03
<i>Bacteroides uniformis</i> CL03T00C23	393	314	367	40.26
<i>Bifidobacterium adolescentis</i> L2-32	44	38	111	40.53
<i>Bacteroides uniformis</i> CL03T12C37	392	312	366	40.81
<i>Prevotella copri</i> CB7, DSM 18205	529	633	523	61.85
<i>Homo sapiens</i>	893	723	760	89.40
<i>Faecalibacterium prausnitzii</i> A2-165	226	361	454	114.64
<i>Bacteroides vulgatus</i> ATCC 8482	777	571	785	121.31

Table 21 continued

<i>Species</i>	<i>Baseline</i>	<i>LRS</i>	<i>HRS</i>	<i>SD</i>
<i>Bacteroides sp. 1_1_6</i>	774	389	320	244.64
<i>Bacteroides sp. D2</i>	885	379	409	283.88
Metagenome	23,359	26,075	28,766	2,703.51

As is reported previously [174], *Bacteriodes* showed changes by RS amount (Table 20). *Bacteriodes vulgatus* ATCC 8482 had 571 proteins in low RS, however its number of proteins is increased in high RS up to 785 (Table 21). Interestingly, all the *Bacteriodes* species, *Bacteriodes* sp. D2, *Bacteriodes* sp. 1_1_6 and *Bacteriodes vulgatus* ATCC 8482, were remarkably decreased in low RS when compared to baseline diet (Table 21). These changes are relatively large, considering changes in other species. The number of human proteins is not quite changed by different diets (Table 21). Abundance of many species in fecal samples was not much changed except *Bacteroides* and Metagenome (Table 21).

Construction of generalized linear mixed model

Generalized linear mixed model (GLMM) was constructed for comparing proteomes between different diets. This model considers that SC can be varied by diet type in different patients. We assumed that eight patients could have different fecal proteome profiles, even though they are basically same biological replicates that have same disease and healthy background. GLMM has patients as random effect and diet type as fixed effect. Negative binomial distribution (NB) was employed as null distribution in GLMM instead of Normal distribution because spectral count is frequency-based data and NB distribution is generally used for count-type data analysis, such as read numbers of RNA sequencing. NB is widely used distribution in RNA sequencing data comparison [176,177], while has not been used in proteomics data analysis. NB is robust to over-

dispersion when compared to Poisson that is also basically designed distribution for count type data. Additionally, NB is uniquely determined by two parameters, mean and variance. However, the number of replicates in biological studies is often not enough to estimate mean and variance reliably. Therefore, tools based on NB for RNA sequencing data comparison, such as DESeq[177] and edgeR[176], have their own variance functions adjusted for RNA sequencing data. The constructed GLMM-NB model was compared to null model through chi-square likelihood ratio test (LR test). All the p-values were adjusted by FDR in order to reduce multiple comparison effect. On top of GLMM-NB model, Cohen's w was applied because GLMM model used NB distribution.

Using GLMM-NB and Cohen's w we found 7,113 significant proteins, 94 from human and 7,018 from microbiome and metagenome, that are differently detected in different diet types. These significant proteins were picked by 0.0001 FDR and 1.0258 Cohen's w threshold (Fig. 36). We increased p-value threshold to reduce the number of expected false positives. The expected false positive in our results is less than 1. Protein scatter plot in figure 36 shows how many proteins we filtered out from total proteins. Since fecal proteomes are highly varied by patients as is shown in chapter 4, we applied rigorous thresholds to exclude out potential false positives. Cohen's w was calculated by selfea R package used in chapter 3 and 4.

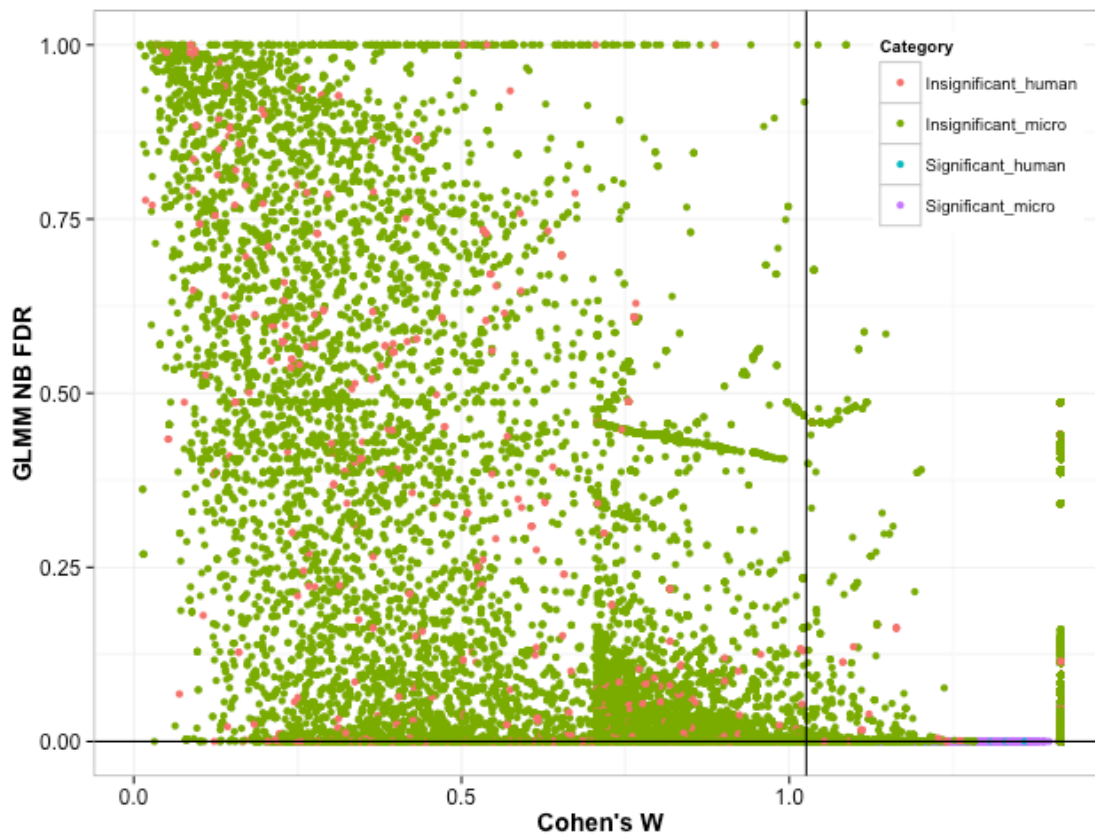


Figure 36 Protein scatter plot of diet proteome comparison. Green dots mean insignificant microbial proteins, and red dots are insignificant human proteins. Significant proteins are colored as purple (significant microbial) and blue (significant human). Since we used rigorous thresholds to make the expected number of false positive below 1, almost all the visible spots were excluded by two lines (FDR threshold (= 0.0001) horizontal line and Cohen's w (= 1.0258) vertical line).

Functional analysis using FAM

Functional analysis of diet proteomes was conducted by using FAM. For this study, we used KEGG pathway analysis for human proteins and COG terms for microbial proteins. Even though relatively small number of human proteins was identified in the diet proteome, we analyzed their functional characteristics. Since we are interested in microbial changes caused by different RS amount, COG analysis of QFAM helped us to get insights about changes in microbiome.

First, we analyzed human proteins in the diet proteome using KEGG pathways. FAM was employed to get reliable results using enrichment test and significance test. Enrichment test was carried out via Fisher's exact test. For quantitative comparison GLMM NB model and Cohen's w were applied. Also, SC transferring from member proteins to KEGG pathways was done for human proteins of diet metaproteome. We downloaded up-to-date NCBI BioSystems DB at the moment of analysis. All the p-values from two tests were adjusted by FDR. FDR thresholds of two tests were 0.05 and minimum Cohen's w was 0.6336 for 80% statistical power that is generally recommended level.

With these probability and ES filters, we could get 20 significant KEGG pathways (Fig. 37). HIF-1 signaling pathway, biosynthesis of amino acids, galactose metabolism, central carbon metabolism in cancer, proteasome and immunoproteasome were enriched almost only in LRS. In other diets, HRS and baseline, these were not expressed. Since proteasome and immunoproteasome are related to defense against microbial proteins, we speculate that innate immune system was turned on against increased species in LRS diets.

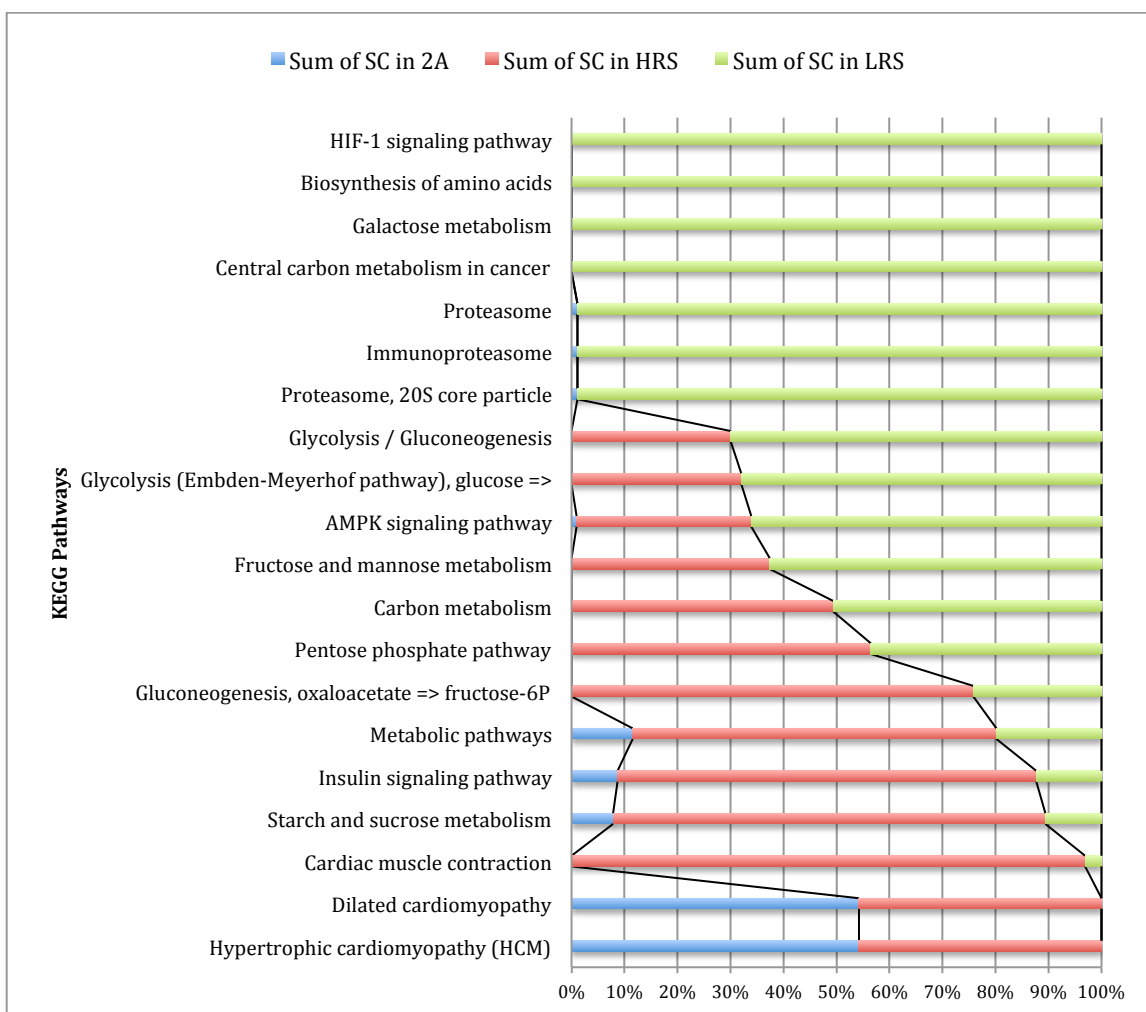


Figure 37 Stacked bar graphs of accumulated spectral counts of 20 significant KEGG pathways. Y-axis is KEGG pathway and X-axis is percentage of accumulated spectral counts of 2A (baseline), HRS and LRS samples. Blue-colored bar is sum of spectral counts in baseline diet. Red-colored bar means expression level (sum of spectral counts) in HRS. Green bars are from sum of expression level in LRS.

Interestingly, digestion-related pathways were enriched in both of LRS and HRS, such as glycolysis, fructose and mannose metabolism, carbon metabolism, pentose phosphate pathway and gluconeogenesis. This may be human body response to get nutrients from changed diets (baseline to LRS or HRS). Insulin signaling pathway and starch and sucrose metabolism were enriched in HRS. These means that uptake of glucose was not efficient in HRS because high amount of RS in HRS cannot be digested and absorbed by small intestine. Therefore, these were expressed higher than other diet types.

COG analysis was achieved by employing FAM. For this study we employed FAM's quantitative test only because of time. Since diet metaproteomics used metagenome, more than several million sequences are in protein DB. To conduct the enrichment test, we have to run PSI-BLAST of more than million sequences, which means that it takes too much long time about a month in Newton. This is critical disadvantage of QFAM. QFAM analyzes highly complicated and very large metaproteomics data, so it requires high level of computing power. Unfortunately, Newton could not afford to run this type of big data in a reasonable time. Therefore, we used arbitrary threshold instead of the enrichment test.

We transferred SC of member proteins to COG terms and constructed GLMM NB model for COG terms. Also, Cohen's w for COG terms was computed for quantitative analysis using QAM. P-values were adjusted by FDR. FDR threshold of GLMM NB model was 0.001 and minimum Cohen's w for 80% statistical power was 0.9051. On top of GLMM NB model, we applied arbitrary

threshold for COG terms by the number of member proteins (> 2). With these thresholds, we could have 166 significant COG terms (Table 22). Also, heat-map of significant 166 COGs was drawn by using QFAM's R graphic library (Fig. 38). Energy generation and digestion related COG terms were enriched in baseline diet, such as COG03808, COG01775 and COG01748 (Table 22 and Fig. 38). Also, ribosomal COGs were significantly enriched, such as COG00261 and COG00238 (Table 22). We speculate that microbiome is growing with baseline diet, considering significantly enriched translation and DNA replication COG terms (Table 22).

COG terms analysis in LRS and HRS diets give us insight speculating changes caused by RS addition. In LRS diet, microbes in human gastrointestinal track are not quite different from baseline. Energy generation (COG00716, COG01013, COG01951, COG00554 and COG01805) is activated and digesting nutrients, such as amino acids, carbohydrates, lipid metabolism, are turned on similarly to baseline (Table 22 and Fig. 38). However, microbes seem to adapt changed nutrients by expressing appropriate proteins and post-translational modifications (COG00060, COG00254, COG00124, COG00621, COG00525 and COG01438 for transcription and translation and COG00760, COG00484, COG00691, COG01066 and COG00740 for post-translational modifications). Interestingly, cell motility COG terms are started being expressed from LRS diets (COG01360 and COG01749), which mean that microbes were moving to somewhere (Table 22). Also, cell wall biogenesis is differentially enriched in LRS when compared to baseline (Table 22 and Fig. 38).

Table 22 166 significant COG terms in different diets. COG Func: COG broad function single digit symbol. Related diets (Rel. Diets) are categorized by expression level in three diets (Base: Baseline, LRS: low RS and HRS: high RS).

<i>COG ID</i>	<i># of Prot</i>	<i>Description</i>	<i>COG Func</i>	<i>Rel. Diets</i>
COG03808	158	Inorganic pyrophosphatase	C	Base
COG01748	13	Saccharopine dehydrogenase and related proteins	E	Base
COG01775	12	Benzoyl-CoA reductase/2-hydroxyglutaryl-CoA dehydratase subunit; BcrC/BadD/HgdB	E	Base
COG00040	7	ATP phosphoribosyltransferase	E	Base
COG00006	4	Xaa-Pro aminopeptidase	E	Base
COG01027	3	Aspartate ammonia-lyase	E	Base
COG00176	9	Transaldolase	G	Base
COG02160	8	L-arabinose isomerase	G	Base
COG03693	5	Beta-1;4-xylanase	G	Base
COG01830	4	DhnA-type fructose-1;6-bisphosphate aldolase and related enzymes	G	Base
COG02893	4	Phosphotransferase system; mannose/fructose-specific component IIA	G	Base
COG01482	3	Phosphomannose isomerase	G	Base
COG03345	3	Alpha-galactosidase	G	Base
COG02022	3	Uncharacterized enzyme of thiazole biosynthesis	H	Base
COG00447	3	Dihydroxynaphthoic acid synthase	H	Base
COG01260	4	Myo-inositol-1-phosphate synthase	I	Base
COG01022	3	Long-chain acyl-CoA synthetases (AMP-forming)	I	Base
COG02030	3	Acyl dehydratase	I	Base
COG00261	35	Ribosomal protein L21	J	Base
COG00238	25	Ribosomal protein S18	J	Base
COG00188	21	Type IIA topoisomerase (DNA gyrase/topo II; topoisomerase IV); A subunit	L	Base
COG01918	21	Fe ²⁺ transport system protein A	P	Base
COG00523	21	Putative GTPases (G3E family)	R	Base
COG02406	10	Protein distantly related to bacterial ferritins	R	Base
COG03937	5	Uncharacterized conserved protein	S	Base
COG01217	25	Predicted membrane GTPase involved in stress response	T	Base

Table 22 continued.

<i>COG ID</i>	<i># of Prot</i>	<i>Description</i>	<i>COG Func</i>	<i>Rel. Diets</i>
COG01702	6	Phosphate starvation-inducible protein PhoH; predicted ATPase	T	Base
COG03706	5	Response regulator containing a CheY-like receiver domain and a GGDEF domain	T	Base
COG00317	9	Guanosine polyphosphate pyrophosphohydrolases/synthetases	TK	Base
COG00342	28	Preprotein translocase subunit SecD	U	LRS
COG00716	51	Flavodoxins	C	LRS
COG01013	20	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases; beta subunit	C	LRS
COG01951	5	Tartrate dehydratase alpha subunit/Fumarate hydratase class I; N-terminal domain	C	LRS
COG00554	3	Glycerol kinase	C	LRS
COG01805	3	Na ⁺ -transporting NADH:ubiquinone oxidoreductase; subunit NqrB	C	LRS
COG00347	31	Nitrogen regulatory protein PII	E	LRS
COG00436	10	Aspartate/tyrosine/aromatic aminotransferase	E	LRS
COG00076	3	Glutamate decarboxylase and related PLP-dependent proteins	E	LRS
COG00175	4	3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase and related enzymes	EH	LRS
COG00444	4	ABC-type dipeptide/oligopeptide/nickel transport system; ATPase component	EP	LRS
COG01102	3	Cytidylate kinase	F	LRS
COG00395	10	ABC-type sugar transport system; permease component	G	LRS
COG03534	5	Alpha-L-arabinofuranosidase	G	LRS
COG02376	4	Dihydroxyacetone kinase	G	LRS
COG01440	4	Phosphotransferase system cellobiose-specific component IIB	G	LRS
COG03669	4	Alpha-L-fucosidase	G	LRS
COG01904	4	Glucuronate isomerase	G	LRS
COG03717	3	5-keto 4-deoxyuronate isomerase	G	LRS

Table 22 continued.

<i>COG ID</i>	<i># of Prot</i>	<i>Description</i>	<i>COG Func</i>	<i>Rel. Diets</i>
COG00413	12	Ketopantoate hydroxymethyltransferase	H	LRS
COG01429	7	Cobalamin biosynthesis protein CobN and related Mg-chelatases	H	LRS
COG00452	4	Phosphopantothenoylcysteine synthetase/decarboxylase	H	LRS
COG00161	4	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	H	LRS
COG00007	3	Uroporphyrinogen-III methylase	H	LRS
COG00214	3	Pyridoxine biosynthesis enzyme	H	LRS
COG01024	15	Enoyl-CoA hydratase/carnithine racemase	I	LRS
COG02185	13	Methylmalonyl-CoA mutase; C-terminal domain/subunit (cobalamin-binding)	I	LRS
COG01788	6	Acyl CoA:acetate/3-ketoacid CoA transferase; alpha subunit	I	LRS
COG00332	3	3-oxoacyl-[acyl-carrier-protein] synthase III	I	LRS
COG00060	15	Isoleucyl-tRNA synthetase	J	LRS
COG00254	7	Ribosomal protein L31	J	LRS
COG00124	5	Histidyl-tRNA synthetase	J	LRS
COG00621	4	2-methylthioadenine synthetase	J	LRS
COG00525	3	Valyl-tRNA synthetase	J	LRS
COG01438	3	Arginine repressor	K	LRS
COG02885	25	Outer membrane protein and related peptidoglycan-associated (lipo)proteins	M	LRS
COG00860	3	N-acetylmuramoyl-L-alanine amidase	M	LRS
COG00773	3	UDP-N-acetylmuramate-alanine ligase	M	LRS
COG02089	3	Sialic acid synthase	M	LRS
COG01360	12	Flagellar motor protein	N	LRS
COG01749	9	Flagellar hook protein FlgE	N	LRS
COG00760	5	Parvulin-like peptidyl-prolyl isomerase	O	LRS
COG00484	4	DnaJ-class molecular chaperone with C-terminal Zn finger domain	O	LRS
COG00691	3	tmRNA-binding protein	O	LRS
COG01066	3	Predicted ATP-dependent serine protease	O	LRS
COG00740	7	Protease subunit of ATP-dependent Clp proteases	OU	LRS
COG01116	5	ABC-type nitrate/sulfonate/bicarbonate transport system; ATPase component	P	LRS

Table 22 continued.

<i>COG ID</i>	<i># of Prot</i>	<i>Description</i>	<i>COG Func</i>	<i>Rel. Diets</i>
COG01160	6	Predicted GTPases	R	LRS
COG00312	6	Predicted Zn-dependent proteases and their inactivated homologs	R	LRS
COG02872	3	Predicted metal-dependent hydrolases related to alanyl-tRNA synthetase HxxxH domain	R	LRS
COG04716	26	Myosin-crossreactive antigen	S	LRS
COG04198	19	Uncharacterized conserved protein	S	LRS
COG04260	8	Putative virion core protein (lumpy skin disease virus)	S	LRS
COG00393	6	Uncharacterized conserved protein	S	LRS
COG04475	5	Uncharacterized protein conserved in bacteria	S	LRS
COG02013	3	Uncharacterized conserved protein	S	LRS
COG02197	3	Response regulator containing a CheY-like receiver domain and an HTH DNA-binding domain	TK	HRS
COG00039	26	Malate/lactate dehydrogenases	C	HRS
COG01456	8	CO dehydrogenase/acetyl-CoA synthase gamma subunit (corrinoid Fe-S protein)	C	HRS
COG03426	7	Butyrate kinase	C	HRS
COG00778	6	Nitroreductase	C	HRS
COG03288	6	NAD/NADP transhydrogenase alpha subunit	C	HRS
COG00479	6	Succinate dehydrogenase/fumarate reductase; Fe-S protein subunit	C	HRS
COG00711	4	F0F1-type ATP synthase; subunit b	C	HRS
COG00371	3	Glycerol dehydrogenase and related enzymes	C	HRS
COG00031	20	Cysteine synthase	E	HRS
COG03643	13	Glutamate formiminotransferase	E	HRS
COG02873	13	O-acetylhomoserine sulfhydrylase	E	HRS
COG00527	9	Aspartokinases	E	HRS
COG00747	7	ABC-type dipeptide transport system; periplasmic component	E	HRS
COG00140	6	Phosphoribosyl-ATP pyrophosphohydrolase	E	HRS
COG01362	5	Aspartyl aminopeptidase	E	HRS

Table 22 continued.

<i>COG ID</i>	<i># of Prot</i>	<i>Description</i>	<i>COG Func</i>	<i>Rel. Diets</i>
COG00410	4	ABC-type branched-chain amino acid transport systems; ATPase component	E	HRS
COG00160	3	4-aminobutyrate aminotransferase and related aminotransferases	E	HRS
COG00106	3	Phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide (ProFAR) isomerase	E	HRS
COG00252	3	L-asparaginase/archaeal Glu-tRNA ^{Gln} amidotransferase subunit D	EJ	HRS
COG00145	3	N-methylhydantoinase A/acetone carboxylase; beta subunit	EQ	HRS
COG03185	10	4-hydroxyphenylpyruvate dioxygenase and related hemolysins	ER	HRS
COG00274	8	Deoxyribose-phosphate aldolase	F	HRS
COG00504	8	CTP synthase (UTP-ammonia lyase)	F	HRS
COG00035	7	Uracil phosphoribosyltransferase	F	HRS
COG00027	6	Formate-dependent phosphoribosylglycinamide formyltransferase (GAR transformylase)	F	HRS
COG03363	6	Archaeal IMP cyclohydrolase	F	HRS
COG00150	5	Phosphoribosylaminoimidazole (AIR) synthetase	F	HRS
COG00152	4	Phosphoribosylaminoimidazolesuccinocarboxamide (SAICAR) synthase	F	HRS
COG00167	4	Dihydroorotate dehydrogenase	F	HRS
COG00284	3	Orotidine-5'-phosphate decarboxylase	F	HRS
COG01109	14	Phosphomannomutase	G	HRS
COG01080	12	Phosphoenolpyruvate-protein kinase (PTS system EI component in bacteria)	G	HRS
COG02190	6	Phosphotransferase system IIA components	G	HRS
COG03957	6	Phosphoketolase	G	HRS
COG01820	5	N-acetylglucosamine-6-phosphate deacetylase	G	HRS
COG01523	4	Type II secretory pathway; pullulanase PulA and related glycosidases	G	HRS
COG00698	4	Ribose 5-phosphate isomerase RpiB	G	HRS

Table 22 continued.

<i>COG ID</i>	<i># of Prot</i>	<i>Description</i>	<i>COG Func</i>	<i>Related Diets</i>
COG01638	4	TRAP-type C4-dicarboxylate transport system; periplasmic component	G	HRS
COG00054	6	Riboflavin synthase beta-chain	H	HRS
COG00511	5	Biotin carboxyl carrier protein	I	HRS
COG00245	3	2C-methyl-D-erythritol 2;4-cyclodiphosphate synthase	I	HRS
COG00304	4	3-oxoacyl-(acyl-carrier-protein) synthase	IQ	HRS
COG00233	17	Ribosome recycling factor	J	HRS
COG01534	5	Predicted RNA-binding protein containing KH domain; possibly ribosomal protein	J	HRS
COG00255	4	Ribosomal protein L29	J	HRS
COG01190	3	Lysyl-tRNA synthetase (class II)	J	HRS
COG00568	28	DNA-directed RNA polymerase; sigma subunit (sigma70/sigma32)	K	HRS
COG01158	18	Transcription termination factor	K	HRS
COG00781	3	Transcription termination factor	K	HRS
COG04775	7	Outer membrane protein/protective antigen OMA87	M	HRS
COG04932	5	Predicted outer membrane protein	M	HRS
COG00438	5	Glycosyltransferase	M	HRS
COG01209	4	dTDP-glucose pyrophosphorylase	M	HRS
COG00793	3	Periplasmic protease	M	HRS
COG00643	19	Chemotaxis protein histidine kinase and related kinases	NT	HRS
COG00840	9	Methyl-accepting chemotaxis protein	NT	HRS
COG01450	4	Type II secretory pathway; component PulD	NU	HRS
COG00576	5	Molecular chaperone GrpE (heat shock protein)	O	HRS
COG02077	5	Peroxiredoxin	O	HRS
COG00735	6	Fe ²⁺ /Zn ²⁺ uptake regulation proteins	P	HRS
COG01117	6	ABC-type phosphate transport system; ATPase component	P	HRS
COG00004	4	Ammonia permease	P	HRS
COG00226	4	ABC-type phosphate transport system; periplasmic component	P	HRS
COG00704	3	Phosphate uptake regulator	P	HRS
COG02368	3	Aromatic ring hydroxylase	Q	HRS

Table 22 continued.

<i>COG ID</i>	<i># of Prot</i>	<i>Description</i>	<i>COG Func</i>	<i>Rel. Diets</i>
COG00714	17	MoxR-like ATPases	R	HRS
COG01837	8	Predicted RNA-binding protein (contains KH domain)	R	HRS
COG01461	8	Predicted kinase related to dihydroxyacetone kinase	R	HRS
COG01878	4	Predicted metal-dependent hydrolase	R	HRS
COG01159	3	GTPase	R	HRS
COG04864	16	Uncharacterized protein conserved in bacteria	S	HRS
COG02268	13	Uncharacterized protein conserved in bacteria	S	HRS
COG00217	11	Uncharacterized conserved protein	S	HRS
COG01624	6	Uncharacterized conserved protein	S	HRS
COG01917	5	Uncharacterized conserved protein; contains double-stranded beta-helix domain	S	HRS
COG01302	4	Uncharacterized protein conserved in bacteria	S	HRS
COG02859	4	Uncharacterized protein conserved in bacteria	S	HRS
COG04702	4	Uncharacterized conserved protein	S	HRS
COG03976	3	Uncharacterized protein conserved in bacteria	S	HRS
COG01875	6	Predicted ATPase related to phosphate starvation-inducible protein PhoH	T	HRS

Figure 38 Heat-map of significant 166 COG terms. We standardized sum of transferred SC of each of COG terms. Deeper steel-blue color means higher expression level in diet metaproteome. COG terms are clustered by their expression pattern and broad functional term.

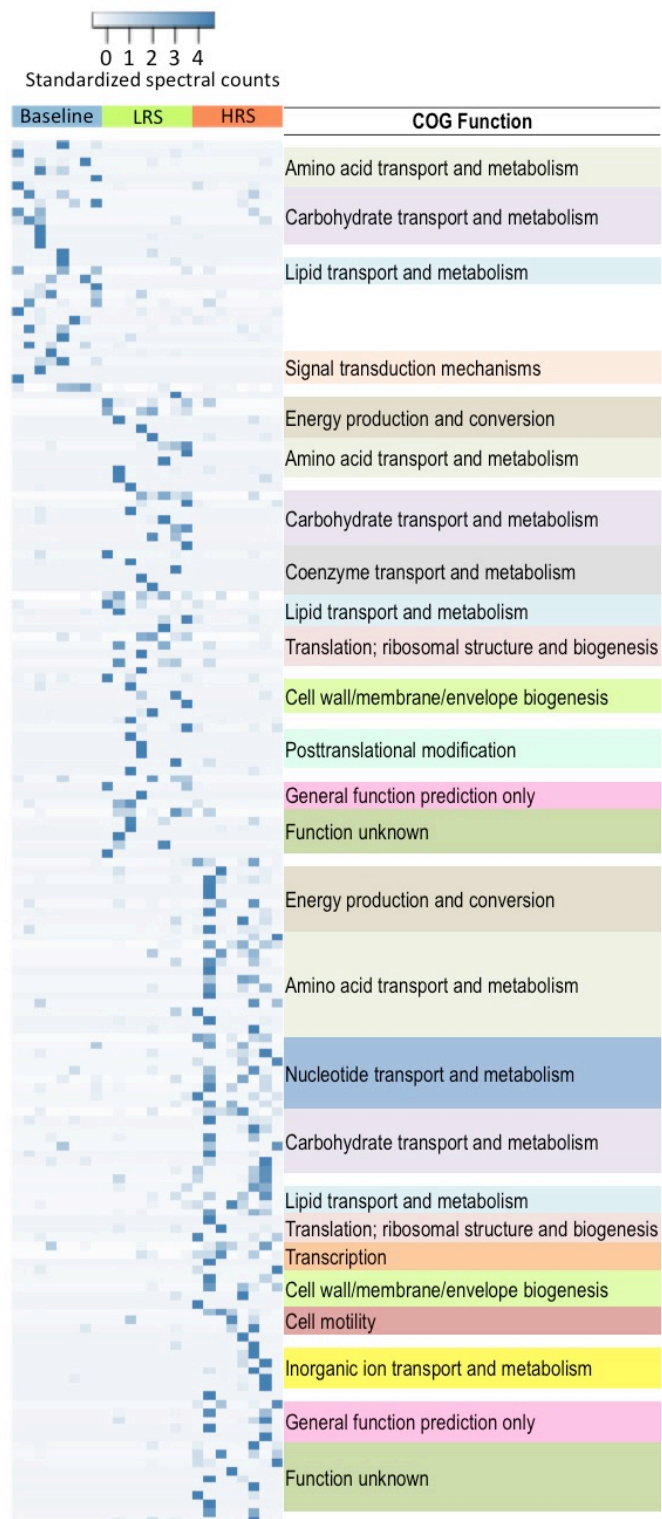


Figure 38 continued

In HRS, energy production and uptake of nutrients were similarly expressed to LRS and baseline. However, cell wall and envelope biogenesis is turned on so that microbes may need to defense themselves from increase of competing or host immune system (COG04775, COG04932, COG00438, COG01209 and COG00793). Likewise to LRS, microbes are trying to express and modify appropriate proteins in order to adapt to the changed environment, high amount of RS (COG00233, COG01534, COG00255 and COG01190 for translation, COG00568, COG01158 and COG00781 for transcription and COG00576 and COG02077 for post-translational modification). Interestingly, HRS also has differentially enriched cell motility COG terms like LRS. This may mean that microbes were moving toward new energy source (COG00643, COG00840 and COG01450).

Conclusion

We could identify 57,398 proteins from 24 MS runs (8 patients * 3 diets). Of 57,398 proteins, 56,295 were from microbiome and 1,103 were host human proteins. This seriously unequal protein identification is caused by employing indirect methods and metagenome. Those two could dramatically increase the number of identified microbial proteins that are main targets of this study.

With this successful microbial protein identification, we constructed GLMM-NB model to compare diet metaproteomes. Since the experiment design is based on cross over, we selected GLMM model by setting patients as random effect and diet types as fixed effect. High variation of fecal proteome, even within

same group was also considered while making GLMM model. NB is employed as null distribution of GLMM model because NB is effective to analyze count type data, such as spectrum counting of the diet metaproteome. On top of GLMM NB model, Cohen's w was combined to enhance statistical power of hypothesis tests. This GLMM NB model and Cohen's w could provide reliable statistical comparisons and were used for finding significantly changed proteins and functional analysis of the diet metaproteome. We could find 7,113 significantly changed proteins (94 from human and 7,018 from microbiome and metagenome). These significant proteins were picked by 0.0001 FDR and 1.0258 Cohen's w

Functional analysis of the diet metaproteome provided understanding of systematic changes in human and microbial community. KEGG pathway analysis showed that functional changes in host proteome are immune response and glucose uptake. COG analysis illustrated that microbial community also had changes in energy generation, gene and protein expression and cell motility. Considering those alterations, we speculate that specific bacteria grow as RS amount in diet is increased because they can use RS as energy source in large intestine. Then, population growth of specific bacteria stimulates human immune response including adaptive and innate immunities. These changes in large intestine, enhanced host defense and nutrient change, result in expression of self-defense and cell motility of existing microbes in the microbiome.

All in all, diet metaproteome is highly complicated and very large and complicated data set. However, QFAM could appropriately analyze the

generated metaproteome and could give exciting insights about functional changes in human host and microbiome. In spite of high variance within biological replicates, GLMM NB and Cohen's w could rigorously compare and give interesting and reliable analysis results.

CHAPTER VI

Conclusions

Overview

The ability to analyze metaproteome datasets quantitatively and functionally, with reliability is an essential requirement in metaproteomics research. Although metaproteomics has emerged as a promising technique for deep and wide characterization of the protein complement of microbial communities, bioinformatic analysis strategies for metaproteomics are currently inadequate. In terms of processing speed and sequence-based functional analysis, required attributes for processing highly complicated and redundant number of metaproteome sequences, MPA and GalaxyP, mentioned in chapter 1, are not optimal pipelines for metaproteomics. In contrast to MPA and GalaxyP, QFAM enables multiprocessing up to 48 CPUs of Newton computational cluster and COG analysis for poorly annotated metagenome and microbial metaproteome sequences. Moreover, QFAM utilize Cohen's effect sizes to filter out low-powered hypothesis tests at desired statistical power level. This quantitative analysis module is applied for all the analysis steps in QFAM, such as QAM, FAM for BioSystems and FAM for COG analysis. The details of QFAM are explained in chapter 2. Chapter 3 illustrated how 'selfea', utilized for quantitative analysis in QAM, could enhance reliability of hypothesis tests in quantitative proteomics data analysis. Selfea can filter out hypothesis tests that have low statistical power in order to keep certain level of statistical power. P-

value cannot be reliable evidence at low statistical power. In the benchmark data test, real noisy data test and Monte Carlo simulation test, selfea could filter out false positives efficiently, while losing few true positives at high statistical power.

We also demonstrated that QFAM is useful to analyze metaproteome data. Chapter 4 illustrated how QFAM is applied for human fecal metaproteome analysis. This human fecal proteome study compared 29 MS runs from 11 healthy volunteers. Human fecal proteome is highly varying between healthy individuals. Its variance within group was shown in figure 31 (0.75 to 0.98 PCC for technical replicates and 0.15 to 0.81 PCC for biological replicates). This variation was confirmed by 90% statistical power and 0.05 p-value of QFAM. From BioSystems of QFAM, we could reveal that digestion related proteins and immune response proteins are differently detected in human fecal proteomes (Table 18). COG analysis showed that almost all the functions including energy generation and defense mechanism were variously enriched in microbial fecal proteomes (Table 20).

Also, QFAM was employed to analyze diet metaproteome data described in chapter 5. Fecal metaproteomes were collected from 8 patients at three different time points. At each time points patients were treated by three different diets, baseline, low resistant starch (RS) (2 mg) and high RS diet (38 mg). Since RS amount in diet is known to be associated with changes in microbiome, this study is designed to investigate quantitative and functional relationship between RS and the gut microbiome. However, fecal proteomes are very dissimilar as was confirmed in chapter 4. Hence, stringent thresholds were applied for

selection of significant features in order to make the expected number of false positives less than 1, and thus results more reliable. QFAM analyzed diet proteomes by constructing GLMM NB model and employing Cohen's w . GLMM NB model could consider personal fecal proteome variation based on negative binomial distribution. Proteins, KEGG pathways, COG terms were compared at 80% statistical power and 0.0001 FDR-adjusted p-value. QFAM's functional and quantitative analysis revealed that glucose uptake, digestion related pathways and immune responses were differentially detected in human host proteomes (Fig. 37). Through COG analysis, QFAM discovered differentially expressed microbial functions, such as cell wall construction, cell motility, energy generation and protein modification (Fig. 38). Based on these findings using QFAM, we speculated that specific bacteria grow as RS amount in diet is increased because they can use RS as energy source in large intestine. Then, population growth of specific bacteria stimulates human immune response including adaptive and innate immunities. These changes in large intestine, enhanced host defense and nutrient change, result in expression of self-defense and cell motility of existing microbes in the microbiome. From two application studies, we could confirm QFAM is practically useful to biologists by providing functional and quantitative analysis.

Current status of QFAM and remaining challenges

QFAM can be employed to normal single species proteomics for quantitative and functional analysis. Since quantitative comparison of proteomes

has been done by using Student's *t*-test or ANOVA based on Normal distribution. However, spectral counting does not follow Normal distribution, so Student's *t*-test and Normal ANOVA are not recommended for proteomics data analysis (Fig. 14). Moreover, Student's *t*-test has serious disadvantage in repeated tests using same dataset. Type I error of Student's *t*-test is linearly increased, while ANOVA can keep Type I error at certain level with multiple tests [98]. Even though QProt is developed to overcome error rate problem of Student's *t*-test, it only enables two-sample comparison. As was explained in chapter 3, *p*-value is only valid at high statistical power that is not achieved in most of biology studies due to limited money and time. Therefore, using Cohen's ES with *p*-value can be applied to general proteomics and other type of data besides metaproteomics.

Despite the availability of many functional analysis tools, FAM is unique due to employing two statistical tests, enrichment test and significance test. Enrichment test using Fisher's exact test is widely accepted methodology to enrich functional terms from signatures. However, enrichment test has not been combined with significance test, to date and to our knowledge. Moreover, FAM's significance test uses selfea providing reliable feature selection. Therefore, FAM can be distinguished from many pre-existing tools by two tests. Usefulness of these two tests was demonstrated by two application studies in chapter 4 and 5.

Even though QFAM can use 48 CPUs, maximum number of allowed CPUs in Newton cluster, for unlimited time in Newton cluster, it is not enough to analyze huge metaproteomics data. For example, while searching PSMs of permafrost metaproteome, it took two weeks for one MS run using 48 CPUs.

Permafrost has 900 megabytes protein database. Since MyriMatch is DB search engine, its search speed is more dependent on the size of DB. Each MS/MS spectra is compared with every theoretically generated peptides of the DB after precursor mass filter. Narrowing mass tolerance window and more precise parent mass measurement can reduce the number of comparisons for one MS/MS spectrum. However, MS instrument with high mass accuracy and precision can be very expensive and narrowing mass tolerance window may result in missing true targets.

De novo sequencing based search algorithm, such as MS-GF+, can reduce PSM search time. Search speed of MS-GF+ is dependent on generation of possible *de novo* sequences rather than DB size. Even though MS-GF+ has not been used for metaproteomics study, if it can complete PSM search even for 1 gigabytes protein DB within reasonable time, it can replace DB search engine efficiently. Moreover, MS-GF+ has reliable score system that is confirmed by several proteomics studies [178]. Therefore, we believe that MS-GF+ can be good replacement of widely used DB search engines, such as SEQUEST and Mascot for metaproteomics studies.

The increase of the available CPUs is the best way to reduce PSM search with QFAM. Although performance test of MyriMatch with 1,000 cores is not confirmed yet, MyriMatch is expected to get faster with larger number of CPUs as was tested up to 48 in Newton cluster (Table 4, 5 and 6). QFAM has Python job controller and can run PSM search and analysis processes using multiple CPUs. Therefore, increasing the number of CPUs in either Newton or in sun grid engine

cluster doesn't matter to QFAM. QFAM's current job controller is designed only for sun grid engine, thus if QFAM is installed to different batch-queuing system; it requires simple update to recognize new job processes.

Increased number of computational cores can speed up COG analysis as well. COG annotation is one of two most time consuming job in QFAM along with the PSM search. In order to calculate enrichment test of COG terms, QFAM tries to annotate all the available protein sequences in protein DB. In human fecal proteome study of chapter 4, the total number of proteins in DB was 100,000. However, the number of proteins in diet metaproteome DB was several millions. This is why we couldn't implement COG enrichment test for diet metaproteome study. Another study, Permafrost metaproteome, had two times more number of proteins than diet proteome study. Therefore, high computational power is necessary for QFAM to operate at optimal levels for all types of metaproteome studies.

QFAM needs to have more options for protein identification. Currently, QFAM has only one option, MyriMatch and IDPicker, for PSM search. Even though MyriMatch and IDPicker are useful and accepted tools for protein identification, users need more options for protein identification. Basically, different algorithms in PSM search generate different protein identification. These identifications can either be combined for greater proteome depth or take only those identifications that each algorithm finds for higher confidence. Other PSM search engines are available, such as X!Tandem [78] and wrapper for proteomics [79]. MS-GF+ would be good option for metaproteomics because it is

not dependent on DB size. QFAM may need to provide various search engines, working with multiple CPUs, to users, such as X!Tandem and MS-GF+ to complement MyriMatch/IDPicker.

QFAM is hard to access by untrained end user because it does not have graphical user interface (GUI), while GalaxyP and MPA provides excellent GUI interface for easy use. GUI should be able to communicate between user and analysis server including transferring large metaproteomics data to the server and receiving analysis results from the server. However, file transfer is limited in Newton cluster for security purpose. Globus (<https://www.globus.org/>) or Dropbox (<https://www.dropbox.com/>) can be possible solution because these provide easy and fast data sharing. Data transfer is recently getting easier by aforementioned tools.

Metaproteomics is promising technique for discovering mechanisms of unknown microbial communities and is a rapidly growing field available to microbial ecologist. However, metaproteomics data is large and highly complicated, so the current largest challenge and bottleneck is the bioinformatics data processing. QFAM is developed to analyze this huge and intricate data quantitatively and functionally at enhanced speed and reliability. Two application studies in chapter 4 and 5 prove usefulness of QFAM for metaproteome data analysis. We believe that QFAM can be an excellent tool for metaproteome experts to improve their analyses of microbial communities.

LIST OF REFERENCES

1. Pieper R, Huang S-T, Suh M-J (2014) Proteomics and Metaproteomics. Encyclopedia of Metagenomics: Springer. pp. 1-11.
2. Hosp F, Vossfeldt H, Heinig M, Vasiljevic D, Arumughan A, et al. (2015) Quantitative Interaction Proteomics of Neurodegenerative Disease Proteins. Cell reports 11: 1134-1146.
3. Lee M-Y, Huang C-H, Kuo C-J, Lin C-LS, Lai W-T, et al. (2015) Clinical Proteomics Identifies Urinary CD14 as a Potential Biomarker for Diagnosis of Stable Coronary Artery Disease. PloS one 10: e0117169.
4. Schnell G, Boeuf A, Westermann B, Jaulhac B, Lipsker D, et al. (2015) Discovery and targeted proteomics on cutaneous biopsies infected by *Borrelia* to investigate Lyme disease. Molecular & Cellular Proteomics 14: 1254-1264.
5. Aerni SJ, Sirota M (2014) A Bioinformatics Guide for Molecular Biologists: Cold Spring Harbor Laboratory Press.
6. Bagg A, Neilands J (1987) Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in *Escherichia coli*. Biochemistry 26: 5471-5477.
7. Ogorzalek Loo RR, Cavalcoli JD, VanBogelen RA, Mitchell C, Loo JA, et al. (2001) Virtual 2-D Gel Electrophoresis: Visualization and Analysis of the *E. coli* Proteome by Mass Spectrometry. Analytical chemistry 73: 4063-4070.
8. Dill BD, Young JC, Carey PA, VerBerkmoes NC (2010) Metaproteomics: techniques and applications: Caister Academic Press: Norfolk, UK.
9. Wilmes P, Bond PL (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environmental Microbiology 6: 911-920.
10. Tanca A, Palomba A, Pisanu S, Deligios M, Fraumene C, et al. (2014) A straightforward and efficient analytical pipeline for metaproteome characterization. Microbiome 2: 1-16.
11. Ram RJ, VerBerkmoes NC, Thelen MP, Tyson GW, Baker BJ, et al. (2005) Community proteomics of a natural microbial biofilm. Science 308: 1915-1920.
12. Lacerda CM, Choe LH, Reardon KF (2007) Metaproteomic analysis of a bacterial community response to cadmium exposure. Journal of proteome research 6: 1145-1152.
13. Powell MJ, Sutton JN, Del Castillo CE, Timperman AT (2005) Marine proteomics: generation of sequence tags for dissolved proteins in seawater using tandem mass spectrometry. Marine chemistry 95: 183-198.
14. Schulze WX, Gleixner G, Kaiser K, Guggenberger G, Mann M, et al. (2005) A proteomic fingerprint of dissolved organic carbon and of soil particles. Oecologia 142: 335-343.
15. Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, et al. (2015) Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature.

16. Xiong W, Abraham P, Li Z, Pan C, Hettich RL (2015) Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics*.
17. Lichtman JS, Marcobal A, Sonnenburg JL, Elias JE (2013) Host-centric proteomics of stool: a novel strategy focused on intestinal responses to the gut microbiota. *Molecular & Cellular Proteomics* 12: 3310-3318.
18. Wu CC, Yates JR (2003) The application of mass spectrometry to membrane proteomics. *Nature biotechnology* 21: 262-267.
19. Washburn MP, Wolters D, Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology* 19: 242-247.
20. Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annual review of biomedical engineering* 11: 49-79.
21. VerBerkmoes NC, Denef VJ, Hettich RL, Banfield JF (2009) Systems biology: functional analysis of natural microbial consortia using community proteomics. *Nature Reviews Microbiology* 7: 196-205.
22. Kelleher NL (2004) Peer reviewed: Top-down proteomics. *Analytical chemistry* 76: 196 A-203 A.
23. VerBerkmoes NC, Bundy JL, Hauser L, Asano KG, Razumovskaya J, et al. (2002) Integrating “top-down” and “bottom-up” mass spectrometric approaches for proteomic analysis of *Shewanella oneidensis*. *Journal of proteome research* 1: 239-252.
24. Strader MB, VerBerkmoes NC, Tabb DL, Connelly HM, Barton JW, et al. (2004) Characterization of the 70S Ribosome from *Rhodopseudomonas palustris* Using an Integrated “Top-Down” and “Bottom-Up” Mass Spectrometric Approach. *Journal of proteome research* 3: 965-978.
25. Piehowski PD, Petyuk VA, Orton DJ, Xie F, Moore RJ, et al. (2013) Sources of technical variability in quantitative LC–MS proteomics: human brain tissue sample analysis. *Journal of proteome research* 12: 2128-2137.
26. Cary SC, Fierer N (2014) The importance of sample archiving in microbial ecology. *Nature Reviews Microbiology* 12: 789-790.
27. Ranjard L, Lejon DP, Mougél C, Schehrer L, Merdinoglu D, et al. (2003) Sampling strategy in molecular microbial ecology: influence of soil sample size on DNA fingerprinting analysis of fungal and bacterial communities. *Environmental Microbiology* 5: 1111-1120.
28. Goodwin RJ, Lang AM, Allingham H, Borén M, Pitt AR (2010) Stopping the clock on proteomic degradation by heat treatment at the point of tissue excision. *Proteomics* 10: 1751-1761.
29. Hettich RL, Pan C, Chourey K, Giannone RJ (2013) Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Analytical chemistry* 85: 4203-4214.
30. Wiśniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. *Nature methods* 6: 359-362.

31. Sharma R, Dill BD, Chourey K, Shah M, VerBerkmoes NC, et al. (2012) Coupling a detergent lysis/cleanup methodology with intact protein fractionation for enhanced proteome characterization. *Journal of proteome research* 11: 6008-6018.
32. Tang Y, Underwood A, Gielbert A, Woodward MJ, Petrovska L (2014) Metaproteomics analysis reveals the adaptation process for the chicken gut microbiota. *Applied and environmental microbiology* 80: 478-485.
33. Pan C, Banfield JF (2014) Quantitative metaproteomics: functional insights into microbial communities. *Environmental Microbiology: Springer*. pp. 231-240.
34. Zhao M, Zhang D-l, Su X-q, Duan S-m, Wan J-q, et al. (2015) An Integrated Metagenomics/Metaproteomics Investigation of the Microbial Communities and Enzymes in Solid-state Fermentation of Pu-erh tea. *Scientific reports* 5.
35. Mann M, Hendrickson RC, Pandey A (2001) Analysis of proteins and proteomes by mass spectrometry. *Annual review of biochemistry* 70: 437-473.
36. Saveliev S, Engel L, Strauss E, Jones R, Rosenblatt M (2007) The Advantages to Using Arg-C, Elastase, Thermolysin and Pepsin for Protein Analysis. pp. Available from: <http://www.promega.com/resources/pubhub/the-advantages-to-using-arg-c-elastase-thermolysin-and-pepsin-for-protein-analysis/>.
37. Wu C, Tran JC, Zamdborg L, Durbin KR, Li M, et al. (2012) A protease for middle-down proteomics. *Nature methods* 9: 822-824.
38. Cannon J, Lohnes K, Wynne C, Wang Y, Edwards N, et al. (2010) High-throughput middle-down analysis using an orbitrap. *Journal of proteome research* 9: 3886-3890.
39. Zhou H, Ning Z, E. Starr A, Abu-Farha M, Figeys D (2011) Advancements in top-down proteomics. *Analytical chemistry* 84: 720-734.
40. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* 26: 1367-1372.
41. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246: 64-71.
42. Kebarle P, Tang L (1993) From ions in solution to ions in the gas phase-the mechanism of electrospray mass spectrometry. *Analytical Chemistry* 65: 972A-986A.
43. Quenzer T, Robinson J, Bolanos B, Milgram E, Greig M. Automated accurate mass analysis using FTICR mass spectrometry; 2002.
44. Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, et al. (2013) Multiplexed MS/MS for improved data-independent acquisition. *Nature methods* 10: 744-746.
45. Michalski A, Cox J, Mann M (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is

- inaccessible to data-dependent LC- MS/MS. *Journal of proteome research* 10: 1785-1793.
46. Egertson JD, MacLean B, Johnson R, Xuan Y, MacCoss MJ (2015) Multiplexed peptide analysis using data-independent acquisition and Skyline. *Nature protocols* 10: 887-903.
 47. Purvine S, Yi EC, Goodlett DR (2003) Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* 3: 847-850.
 48. Myung S, Lee YJ, Moon MH, Taraszka J, Sowell R, et al. (2003) Development of high-sensitivity ion trap ion mobility spectrometry time-of-flight techniques: a high-throughput nano-LC-IMS-TOF separation of peptides arising from a *Drosophila* protein extract. *Analytical chemistry* 75: 5137-5145.
 49. Venable JD, Dong M-Q, Wohlschlegel J, Dillin A, Yates JR (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature methods* 1: 39-45.
 50. Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, et al. (2009) Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Analytical chemistry* 81: 6481-6488.
 51. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, et al. (2010) Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26: 966-968.
 52. Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, et al. (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature biotechnology* 32: 219-223.
 53. Distler U, Kuharev J, Navarro P, Levin Y, Schild H, et al. (2014) Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nature methods* 11: 167-170.
 54. Doerr A (2015) DIA mass spectrometry. *Nature Methods* 12: 35-35.
 55. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry* 389: 1017-1031.
 56. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, et al. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics* 1: 376-386.
 57. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, et al. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology* 17: 994-999.
 58. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, et al. (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics* 3: 1154-1169.
 59. Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, et al. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry* 75: 1895-1904.

60. Belnap CP, Pan C, Denev VJ, Samatova NF, Hettich RL, et al. (2011) Quantitative proteomic analyses of the response of acidophilic microbial communities to different pH conditions. *The ISME journal* 5: 1152-1161.
61. Belnap CP, Pan C, VerBerkmoes NC, Power ME, Samatova NF, et al. (2010) Cultivation and quantitative proteomic analyses of acidophilic microbial communities. *The ISME journal* 4: 520-530.
62. Pan C, Fischer CR, Hyatt D, Bowen BP, Hettich RL, et al. (2011) Quantitative tracking of isotope flows in proteomes of microbial communities. *Molecular & Cellular Proteomics* 10: M110. 006049.
63. Taubert M, Jehmlich N, Vogt C, Richnow HH, Schmidt F, et al. (2011) Time resolved protein-based stable isotope probing (Protein-SIP) analysis allows quantification of induced proteins in substrate shift experiments. *Proteomics* 11: 2265-2274.
64. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 35: D61-D65.
65. Consortium U (2008) The universal protein resource (UniProt). *Nucleic acids research* 36: D190-D195.
66. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, et al. (2012) IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic acids research* 40: D115-D122.
67. Cantarel BL, Erickson AR, VerBerkmoes NC, Erickson BK, Carey PA, et al. (2011) Strategies for metagenomic-guided whole-community proteomics of complex microbial environments. *PLoS One* 6: e27173.
68. Eng JK, McCormack AL, Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5: 976-989.
69. Ma B, Zhang K, Hendrie C, Liang C, Li M, et al. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry* 17: 2337-2342.
70. Frank A, Pevzner P (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry* 77: 964-973.
71. Seidler J, Zinn N, Boehm ME, Lehmann WD (2010) De novo sequencing of peptides by MS/MS. *Proteomics* 10: 634-649.
72. Cottrell JS, London U (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *electrophoresis* 20: 3551-3567.
73. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20: 1466-1467.
74. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. (2004) Open mass spectrometry search algorithm. *Journal of proteome research* 3: 958-964.
75. Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of proteome research* 6: 654-661.

76. Kim S, Pevzner PA (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature communications* 5.
77. Kim S, Gupta N, Bandeira N, Pevzner PA (2009) Spectral dictionaries integrating de novo peptide sequencing with database search of tandem mass spectra. *Molecular & Cellular Proteomics* 8: 53-69.
78. Bjornson RD, Carriero NJ, Colangelo C, Shifman M, Cheung K-H, et al. (2007) X!! Tandem, an improved method for running X! tandem in parallel on collections of commodity computers. *The Journal of Proteome Research* 7: 293-299.
79. Mohammed Y, Mostovenko E, Henneman AA, Marissen RJ, Deelder AM, et al. (2012) Cloud parallel processing of tandem mass spectrometry based proteomics data. *Journal of proteome research* 11: 5101-5108.
80. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods* 4: 207-214.
81. Gupta N, Bandeira N, Keich U, Pevzner PA (2011) Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry* 22: 1111-1120.
82. Wilkins MJ, VerBerkmoes NC, Williams KH, Callister SJ, Mouser PJ, et al. (2009) Proteogenomic monitoring of *Geobacter* physiology during stimulated uranium bioremediation. *Applied and environmental microbiology* 75: 6591-6599.
83. Craig R, Cortens J, Fenyo D, Beavis RC (2006) Using annotated peptide mass spectrum libraries for protein identification. *Journal of proteome research* 5: 1843-1849.
84. Yen C-Y, Houel S, Ahn NG, Old WM (2011) Spectrum-to-spectrum searching using a proteome-wide spectral library. *Molecular & Cellular Proteomics* 10: M111. 007666.
85. Dasari S, Chambers MC, Martinez MA, Carpenter KL, Ham A-JL, et al. (2012) Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *Journal of proteome research* 11: 1686-1695.
86. Tabb DL, McDonald WH, Yates JR (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *Journal of proteome research* 1: 21-26.
87. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry* 74: 5383-5392.
88. Searle BC (2010) Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* 10: 1265-1269.
89. Meyer-Arendt K, Old WM, Houel S, Renganathan K, Eichelberger B, et al. (2011) IsoformResolver: A peptide-centric algorithm for protein inference. *Journal of proteome research* 10: 3060-3075.

90. Zhang B, Chambers MC, Tabb DL (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *Journal of proteome research* 6: 3549-3557.
91. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature methods* 4: 923-925.
92. Mondav R, Woodcroft BJ, Kim E-H, McCalley CK, Hodgkins SB, et al. (2014) Discovery of a novel methanogen prevalent in thawing permafrost. *Nature communications* 5.
93. Zhang Y, Wen Z, Washburn MP, Florens L (2010) Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Analytical chemistry* 82: 2272-2281.
94. Zybailov B, Mosley AL, Sardi ME, Coleman MK, Florens L, et al. (2006) Statistical Analysis of Membrane Proteome Expression Changes in *Saccharomyces cerevisiae*. *Journal of proteome research* 5: 2339-2347.
95. Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, et al. (2012) Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PloS one* 7: e49138.
96. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, et al. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics* 4: 1265-1272.
97. Li M, Gray W, Zhang H, Chung CH, Billheimer D, et al. (2010) Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *Journal of proteome research* 9: 4295-4305.
98. Ltd Lr (2013) 1-way ANOVA lecture note. pp. <https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-2.php>.
99. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11: R86.
100. Muth T, Behne A, Heyer R, Kohrs F, Berndorf D, et al. (2015) The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *Journal of proteome research* 14: 1557-1565.
101. Park CY, Klammer AA, Käll L, MacCoss MJ, Noble WS (2008) Rapid and accurate peptide identification from tandem mass spectra. *Journal of proteome research* 7: 3022-3027.
102. Tanner S, Shu H, Frank A, Wang L-C, Zandi E, et al. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical chemistry* 77: 4626-4639.
103. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28: 27-30.
104. Webb EC (1992) Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and

- Molecular Biology on the Nomenclature and Classification of Enzymes: Academic Press.
105. Federhen S (2012) The NCBI taxonomy database. *Nucleic acids research* 40: D136-D143.
 106. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, et al. (2009) The NCBI biosystems database. *Nucleic acids research*: gkp858.
 107. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278: 631-637.
 108. Wickham H (2009) *ggplot2: elegant graphics for data analysis*: Springer Science & Business Media.
 109. Gross ML, Chen G, Pramanik B (2011) *Protein and peptide mass spectrometry in drug discovery*: John Wiley & Sons.
 110. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*: 289-300.
 111. Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 8: e1002375.
 112. Caspi R, Altman T, Billington R, Dreher K, Foerster H, et al. (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research* 42: D459-D471.
 113. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, et al. (2014) The Reactome pathway knowledgebase. *Nucleic acids research* 42: D472-D477.
 114. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) PID: the pathway interaction database. *Nucleic acids research* 37: D674-D679.
 115. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. *Nature genetics* 25: 25-29.
 116. Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* 33: D54-D58.
 117. Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic acids research* 29: 137-140.
 118. Van Der Walt S, Colbert SC, Varoquaux G (2011) The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering* 13: 22-30.
 119. Jones E, Oliphant T, Peterson P (2014) *{SciPy}: Open source scientific tools for {Python}*.
 120. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, et al. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* 26: 1481-1487.
 121. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, et al. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic acids research* 30: 281-283.

122. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25: 3389-3402.
123. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB (2015) The fickle P value generates irreproducible results. *Nature methods* 12: 179-185.
124. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14: 365-376.
125. Sullivan GM, Feinn R (2012) Using effect size-or why the P value is not enough. *Journal of graduate medical education* 4: 279-282.
126. Gregori J, Villarreal L, Sánchez A, Baselga J, Villanueva J (2013) An effect size filter improves the reproducibility in spectral counting-based comparative proteomics. *Journal of Proteomics* 95: 55-65.
127. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 98: 5116-5121.
128. Gregori J, Sanchez A, Villanueva J (2013) msmsTests. Bioconductor. pp. msmsTests, <http://www.bioconductor.org/packages/release/bioc/html/msmsTests.html>.
129. Kabacoff RI (2014) Power Analysis. Quick-R. pp. Power Analysis, <http://www.statmethods.net/stats/power.html>.
130. Cohen J (2013) Statistical power analysis for the behavioral sciences: Academic press.
131. Stephane Champely CE, Peter Dalgaard, Jeffrey Gill, Helios De Rosario (2015) R package 'pwr'. CRAN. pp. R package 'pwr' ver. 1.1-2, power analysis functions along the lines of Jacob Cohen.
132. Wikipedia (2014) Effect size. Wikipedia. pp. http://en.wikipedia.org/wiki/Effect_size.
133. Venables WN, Ripley BD (2002) Modern applied statistics with S: Springer Science & Business Media.
134. Cohen J (1990) Things I have learned (so far). *American psychologist* 45: 1304.
135. Cohen J (1992) A power primer. *Psychological bulletin* 112: 155.
136. Ioannidis JP (2005) Why most published research findings are false. *PLoS medicine* 2: e124.
137. Choi H, Fermin D, Nesvizhskii AI (2008) Significance analysis of spectral count data in label-free shotgun proteomics. *Molecular & cellular proteomics* 7: 2373-2385.
138. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, et al. (2012) Gut microbiota composition correlates with diet and health in the elderly. *Nature* 488: 178-184.
139. Pacheco AR, Curtis MM, Ritchie JM, Munera D, Waldor MK, et al. (2012) Fucose sensing regulates bacterial intestinal colonization. *Nature* 492: 113-117.

140. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, et al. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491: 119-124.
141. Qin J, Li Y, Cai Z, Li S, Zhu J, et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490: 55-60.
142. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, et al. (2009) Shotgun metaproteomics of the human distal gut microbiota. *The ISME journal* 3: 179-189.
143. Xiong W, Abraham PE, Li Z, Pan C, Hettich RL (2015) Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics*.
144. Fouts DE, Pieper R, Szpakowski S, Pohl H, Knoblich S, et al. (2012) Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J Transl Med* 10: 174.
145. Cohen J (1992) Statistical power analysis. *Current directions in psychological science*: 98-101.
146. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic acids research* 29: 22-28.
147. Madsen J, Mollenhauer J, Holmskov U (2010) Review: Gp-340/DMBT1 in mucosal innate immunity. *Innate immunity* 16: 160-167.
148. Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, et al. (2008) Shotgun metaproteomics of the human distal gut microbiota. *The ISME journal* 3: 179-189.
149. Sun Y, Wright H, Janciauskiene S (2002) Glioma cell activation by Alzheimer's peptide A β 1-42, α 1-antichymotrypsin, and their mixture. *Cellular and Molecular Life Sciences CMLS* 59: 1734-1743.
150. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. (2014) A draft map of the human proteome. *Nature* 509: 575-581.
151. Ortutay C, Vihinen M (2006) Immunome: a reference set of genes and proteins for systems biology of the human immune system. *Cellular immunology* 244: 87-89.
152. Ortutay C, Siermala M, Vihinen M (2007) Molecular characterization of the immune system: emergence of proteins, processes, and domains. *Immunogenetics* 59: 333-348.
153. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, et al. (2008) InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Molecular systems biology* 4.
154. Diehl AD, Lee JA, Scheuermann RH, Blake JA (2007) Ontology development for biological systems: immunology. *Bioinformatics* 23: 913-915.

155. Harada N, Iijima S, Kobayashi K, Yoshida T, Brown WR, et al. (1997) Human IgG₁ binding protein (FcγBP) in colonic epithelial cells exhibits mucin-like structure. *Journal of Biological Chemistry* 272: 15232-15241.
156. Johansson ME, Thomsson KA, Hansson GC (2009) Proteomic analyses of the two mucus layers of the colon barrier reveal that their main component, the Muc2 mucin, is strongly bound to the Fcγbp protein. *Journal of proteome research* 8: 3549-3557.
157. Velliyagounder K, Kaplan JB, Furgang D, Legarda D, Diamond G, et al. (2003) One of two human lactoferrin variants exhibits increased antibacterial and transcriptional activation activities and is associated with localized juvenile periodontitis. *Infection and immunity* 71: 6141-6147.
158. Ando K, Hasegawa K, Shindo Ki, Furusawa T, Fujino T, et al. (2010) Human lactoferrin activates NF-κB through the Toll-like receptor 4 pathway while it interferes with the lipopolysaccharide-stimulated TLR4 signaling. *FEBS journal* 277: 2051-2066.
159. Roth W, Kumar V, Beer H-D, Richter M, Wohlenberg C, et al. (2012) Keratin 1 maintains skin integrity and participates in an inflammatory network in skin through interleukin-18. *Journal of cell science* 125: 5269-5279.
160. Gebhardt C, Németh J, Angel P, Hess J (2006) S100A8 and S100A9 in inflammation and cancer. *Biochemical pharmacology* 72: 1622-1631.
161. Taheri M, Saragovi U, Fuks A, Makkerh J, Mort J, et al. (2000) Self recognition in the Ig superfamily Identification of precise subdomains in carcinoembryonic antigen required for intercellular adhesion. *Journal of Biological Chemistry* 275: 26935-26943.
162. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, et al. (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. *Journal of proteome research* 5: 2909-2918.
163. Yin Y, Yu G, Chen Y, Jiang S, Wang M, et al. (2012) Genome-wide transcriptome and proteome analysis on different developmental stages of *Cordyceps militaris*. *PloS one* 7: e51853.
164. Zeng J, Liu Y, Liu W, Liu X, Liu F, et al. (2013) Integration of transcriptome, proteome and metabolism data reveals the alkaloids biosynthesis in *Macleaya cordata* and *Macleaya microcarpa*. *PloS one* 8: e53409.
165. Walker AW, Ince J, Duncan SH, Webster LM, Holtrop G, et al. (2011) Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *The ISME journal* 5: 220-230.
166. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* 444: 1022-1023.
167. Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI (2008) Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell host & microbe* 3: 213-223.
168. Walter J, Ley R (2011) The human gut microbiome: ecology and recent evolutionary changes. *Annual review of microbiology* 65: 411-429.

169. Englyst HN, Kingman S, Cummings J (1992) Classification and measurement of nutritionally important starch fractions. *European journal of clinical nutrition* 46: S33-50.
170. Duncan SH, Belenguer A, Holtrop G, Johnstone AM, Flint HJ, et al. (2007) Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Applied and environmental microbiology* 73: 1073-1078.
171. Martínez I, Kim J, Duffy PR, Schlegel VL, Walter J (2010) Resistant starches types 2 and 4 have differential effects on the composition of the fecal microbiota in human subjects. *PloS one* 5: e15046-e15046.
172. Abell GC, Cooke CM, Bennett CN, Conlon MA, McOrist AL (2008) Phylotypes related to *Ruminococcus bromii* are abundant in the large bowel of humans and increase in response to a diet high in resistant starch. *FEMS microbiology ecology* 66: 505-515.
173. Leitch E, Walker AW, Duncan SH, Holtrop G, Flint HJ (2007) Selective colonization of insoluble substrates by human faecal bacteria. *Environmental microbiology* 9: 667-679.
174. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, et al. (2010) Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences* 107: 14691-14696.
175. Kuznetsova A, Brockhoff PB, Christensen RHB (2013) lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.
176. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139-140.
177. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome biology* 11: R106.
178. Granholm V, Kim S, Navarro JC, Sjölund E, Smith RD, et al. (2013) Fast and accurate database searches with MS-GF+ Percolator. *Journal of proteome research* 13: 890-897.

VITA

Lang Ho Lee was born in Seoul, Republic of South Korea. He graduated from Hanseo High School in 1998. Then, he entered to Handong Global University at Pohang city of Republic of South Korea and graduated with B.S. in 2003. Following the completion of his B.S. in 2003, he enrolled in the graduate school of Biochemistry at Yonsei University in 2003 and graduated with M.S. in 2005. After graduation with M.S., he worked at industry and academy as professional bioinformatics scientist. He moved to the United States of America in 2010 and enrolled in Genome Science and Technology (GST) at the University of Tennessee, where he studied biological mass spectrometry and bioinformatics for quantitative proteomics analysis. He expects to receive his Ph.D. in August of 2015. With Ph.D., he minored Statistics at the University of Tennessee.