Journal of Biomolecular Techniques • Volume 34(3); 2023 Sep

The Association of Biomolecular Resource Facilities Proteome Informatics Research Group Study on Metaproteomics (iPRG-2020)

Pratik D. Jagtap[1] Michael R. Hoopmann[2] Benjamin A. Neely[3] Antony Harvey[4] Lukas Käll[5] Yasset Perez-Riverol[6] Milky K. Abajorga[7] Julie A. Thomas[8] Susan T. Weintraub[9] Magnus Palmblad[10]

[1]University of Minnesota, Minneapolis, Minnesota 55455, USA,
[2]Institute for Systems Biology, Seattle, Washington 98109, USA,
[3]National Institute of Standards and Technology, Charleston, South Carolina 29412, USA,
[4]Protein Metrics LLC, Chandler, Texas 75758, USA,
[5]Royal Institute of Technology, 114 28 Stockholm, Sweden,
[6]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom,
[7]UMass Chan Medical School, Worcester, Massachusetts 01655, USA,
[8]Rochester Institute of Technology, Rochester, New York 14623, USA,

[9]University of Texas Health Science Center at San Antonio, Texas 78229, USA,
[10]Center for Proteomics and Metabolomics, Leiden University Medical Center, 2000 RC Leiden, The Netherlands

Association of Biomolecular Resource Facilities

Published on: Aug 07, 2023

DOI: https://doi.org/10.7171/3fc1f5fe.a058bad4

ABSTRACT

Metaproteomics research using mass spectrometry data has emerged as a powerful strategy to understand the mechanisms underlying microbiome dynamics and the interaction of microbiomes with their immediate environment. Recent advances in sample preparation, data acquisition, and bioinformatics workflows have greatly contributed to progress in this field. In 2020, the Association of Biomolecular Research Facilities Proteome Informatics Research Group launched a collaborative study to assess the bioinformatics options available for metaproteomics research. The study was conducted in 2 phases. In the first phase, participants were provided with mass spectrometry data files and were asked to identify the taxonomic composition and relative taxa abundances in the samples without supplying any protein sequence databases. The most challenging question asked of the participants was to postulate the nature of any biological phenomena that may have taken place in the samples, such as interactions among taxonomic species. In the second phase, participants were provided a protein sequence database composed of the species present in the sample and were asked to answer the same set of questions as for phase 1. In this report, we summarize the data processing methods and tools used by participants, including database searching and software tools used for taxonomic and functional analysis. This study provides insights into the status of metaproteomics bioinformatics in participating laboratories and core facilities.

0:00 / 32:41

Listen to this article

ADDRESS CORRESPONDENCE TO: Pratik D. Jagtap, 321 Church Street SE, University of Minnesota, Minneapolis, Minnesota 55455 (Phone: 612-816-4232; Email: pjagtap@umn.edu).

ADDRESS CORRESPONDENCE TO: Susan T. Weintraub, 7703 Floyd Curl Drive, University of Texas Health Science Center, San Antonio, Texas 78229 (Phone: 210-567-4043; Email: weintraub@uthscsa.edu).

ADDRESS CORRESPONDENCE TO: Magnus Palmblad, 2300 RC Leiden, Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands (Phone: +31(0)71 526 6969; Email: n.m.palmblad@lumc.nl).

Conflict of Interest: None of the individuals and financial support funding associations have a conflict of interest.

All mass spectrometry datasets are available via ProteomeXchange with identifier PXD034795.

Keywords: metaproteomics, microbiome, bioinformatics, taxonomy, mass spectrometry

INTRODUCTION

Mass spectrometry–based metaproteomics provides valuable insight into microbiome composition and function by identifying and quantifying the proteins expressed by microbiota. The field has seen a steady growth in the last few years.[1],[2] While notable progress has been made in the optimization of sample preparation and data acquisition methods, bioinformatics analysis has remained particularly challenging.[3],[4] For example, metaproteomics samples often have a high level of microbial diversity, and when the host is also present, the microbial content may be proportionally very low.[5] Once the mass spectrometry data are acquired, the spectra need to be matched against protein sequences. However, when the composition is very diverse, it is necessary to search against large protein sequence databases, which can lead to low peptide identification sensitivity and/or detection of high numbers of false positives.[6] Moreover, the protein sequence databases need to be appropriately annotated to permit protein inference as well as functional pathway and taxonomic analyses.[2] Researchers can address this challenge by generating sample-specific databases using matched metagenomic data.[7] However, these tasks require substantial computational resources at both the software and hardware infrastructure levels. Moreover, mapping microbial peptides to proteins is further complicated by the fact that many of the peptides are shared by homologous proteins across taxonomic groups. The relative abundance of the taxonomic species, dynamic range of their protein expression, and biological variability add further challenges in biological interpretation.[8]

To address issues related to the detection of microbial peptides and proteins, iterative approaches have been used to search metaproteomics datasets against large public repository databases.[9],[10],[11],[12],[13],[14] Other advancements for database search strategies include the following: database reduction by processes such as the sectioning of the large protein sequences database,[4],[12] the use of specialized database structure and search,[13] and de novo search methods.[15],[16] More recently, the availability of matched metagenome data has made it possible to generate customized search databases for optimal search results.[1],[7],[17] In peptide-centric approaches, researchers have used various software tools to detect taxonomy[18],[19],[20] and function.[21],[22],[23],[24] Bioinformatics approaches have also been developed for the quantitative assessment of changes in taxonomy and function.[25],[26],[27],[28],[29]

The Association of Biomolecular Resource Facilities (ABRF) Proteome Informatics Research Group (iPRG) considered it important to assess the state of metaproteomics bioinformatical analysis and decided to conduct a 2-phase research study on this topic in 2020. In addition to standard mechanisms for the announcement of ABRF studies, we proactively contacted metaproteomics researchers across the world and invited them to participate by submitting their analysis of our bacteriophage infection dataset. In the first phase, minimal information about the dataset was provided, while in the second phase, we provided a search database that would aid in the metaproteomics analysis. We asked participants to respond with the following information: (a) a detailed explanation of their data processing workflow, including protein sequence database(s) and strategy used for peptide-spectral matching; (b) taxonomic composition (along with details on how this was

determined); and (c) any biological functions or phenomena observed in the data. See Figure 1 for an overview of the study design.

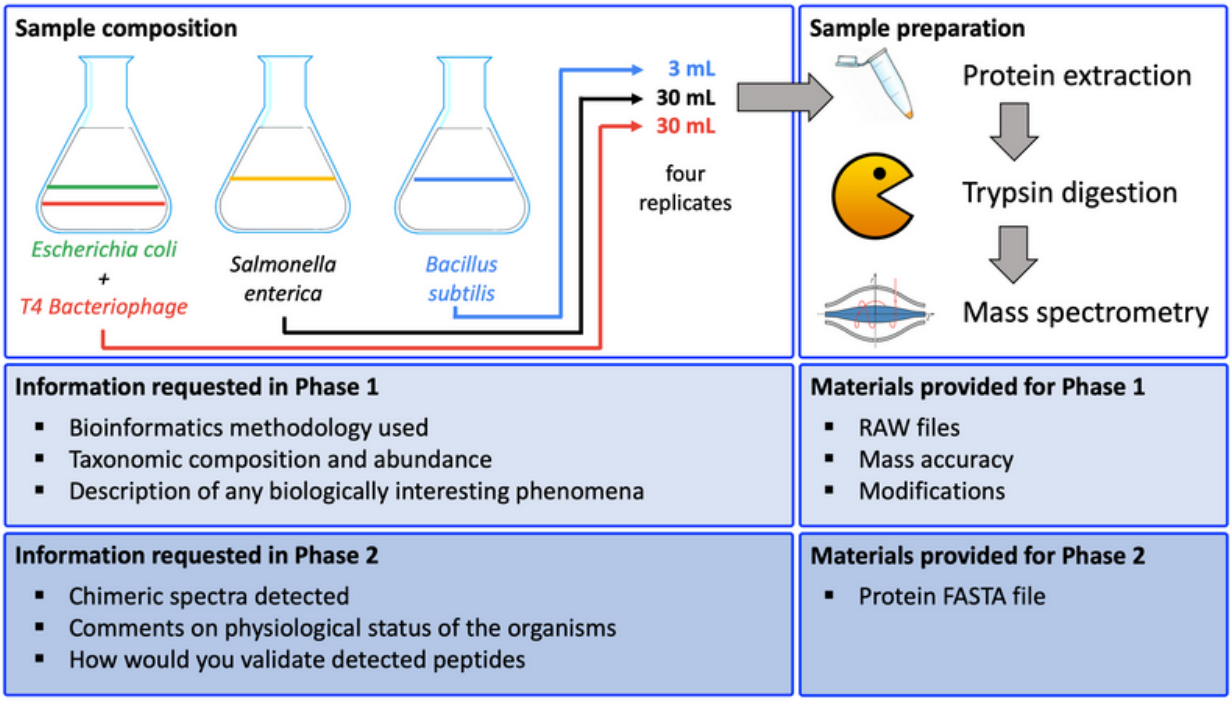

Figure 1

Overview of the bacteriophage infection experimental design. Four biological replicates of T4 bacteriophage–infected *E. coli* were subjected to protein extraction and trypsin digestion followed by Liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis. The experimental samples also contained *S. enterica* and *B. subtilis*, which are not hosts for T4 bacteriophage. The RAW files generated by the mass spectrometry analyses were provided to participants who were then tasked to apply bioinformatics methods of their choosing and report on the taxonomic composition and biological significance of the dataset. A protein sequence database was not made available for Phase 1 of study, but it was provided to participants for Phase 2. The second phase also requested answers to questions that were not asked in the first phase regarding detection of chimeric spectra, comparisons with existing data in public databases, and validation of peptide identifications that supported the conclusions.

MATERIALS AND METHODS

Sample preparation

Four separate liquid cultures were prepared for each of the following bacteria: *Escherichia coli* strain B^E, *Salmonella enterica* strain UB-0015, and *Bacillus subtilis* strain 168. Each culture was derived from a different single colony that was inoculated into 15 mL of Luria Broth (LB) (Miller) supplemented with 0.2% nutrient broth (LB+N) and incubated with shaking (150 rpm) overnight at 34 °C. The following morning, the 12

cultures were separately subcultured (1:75 dilution) into 30 mL LB+N broth and incubated at 34 °C with shaking. An additional subculture of *E. coli* was included to permit the monitoring of optical density. At optical density at 600 nm wavelength (OD_{600}) 0.9, the 5 *E. coli* cultures were infected with bacteriophage T4 at a multiplicity of infection of ~3, and the infection was allowed to proceed for 20 minutes. One culture of each of the 3 different bacteria was used to generate 4 biological mixtures, each containing 3 mL of *B. subtilis*, 30 mL of *S. enterica*, and 30 mL of T4-infected *E. coli*. The cells were then rapidly harvested by centrifugation (7900 *g*, 23 °C, 3 minutes) and the supernatants decanted. The cell pellets were subjected to 2 freeze–thaw cycles, and each was resuspended in 1.4 mL of buffer (50 mM Tris-Cl [pH 7.5], 50 mM NaCl, and 1 mM $MgCl_2$) that was supplemented with 1x BugBuster reagent (MilliporeSigma) and 4 µL of Lysonase Bioprocessing Reagent (MilliporeSigma). The mixtures were incubated at 23 °C for 10 minutes with vortexing and then stored at -80 °C.

Mass spectrometry sample preparation and data acquisition

Four biological replicates of the mixture described above of proteins from T4 bacteriophage and its host *E. coli* along with non-host species *S. enterica* and *B. subtilis* were used for mass spectrometry (MS) analysis. After thawing, samples were mixed with a buffer containing 5% sodium dodecyl sulfate (SDS) /50 mM triethylammonium bicarbonate (TEAB) in the presence of protease and phosphatase inhibitors (Halt; Thermo Scientific) and nuclease (Pierce Universal Nuclease for Cell Lysis; Thermo Scientific). Aliquots corresponding to 100 µg protein (EZQ Protein Quantitation Kit; Thermo Scientific) were reduced with tris(2-carboxyethyl) phosphine hydrochloride, alkylated in the dark with iodoacetamide, and applied to S-Traps (mini; Protifi) for tryptic digestion (sequencing grade; Promega) in 50 mM TEAB. Peptides were eluted from the S-Traps with 0.2% formic acid in 50% aqueous acetonitrile and quantified using Pierce Quantitative Fluorometric Peptide Assay (Thermo Scientific). A 1-µg sample of each digest was analyzed by capillary LC-MS/MS on a Thermo Scientific Orbitrap Fusion Lumos mass spectrometer. On-line High-performance liquid chromatography (HPLC) separation was accomplished with a Thermo Scientific/Dionex RSLC NANO HPLC system: column, PicoFrit (New Objective; 75 µm i.d.) packed to 15 cm with C18 adsorbent (Vydac; 218MS 5 µm, 300 Å); mobile phase A, 0.5% acetic acid (HAc)/0.005% trifluoroacetic acid (TFA); mobile phase B, 90% acetonitrile/0.5% HAc/0.005% TFA; gradient 3 to 42% B in 30 minutes; flow rate, 0.4 µL/minutes. Precursor ions were acquired in the Orbitrap in centroid mode (scan range, *m/z* 300-1500; resolution, 120000); data-dependent, higher-energy, collision-induced dissociation spectra of ions in the precursor scan were acquired at the same time in the ion trap ("top speed"; threshold to trigger MS2, 50000; quadrupole isolation, 0.7; charge states, 2+ to 5+; dynamic exclusion, 30 seconds; normalized collision energy, 30%).

Study phases

Phase 1 (no protein database provided)

The "iPRG-2020 Proteome Informatics Research Group Study on Metaproteomics" was first announced in April 2020 via the iPRG website (see announcement here) and social media (Google Forms and Twitter). In addition, potential participants were contacted via email. In this phase, no information was provided about the composition of the samples (such as the number of species present or the domains that were represented). Raw LC-MS/MS data files that had been acquired on a Thermo Fisher Scientific Orbitrap Fusion Lumos instrument were made available to participants along with specifics about data acquisition. Participants were asked to provide the following:

> 1.1. A detailed explanation of the steps used for the analysis, including why the specific sequence databases were selected, how they were assembled, how spectra were matched/assigned, and how the taxa were identified.
>
> 1.2. A list (as a text file or table) of the taxa identified in the sample.
>
> 1.3. A relative abundance of the different species (such as numbers of peptide-spectrum matches [PSMs], distinct peptides, or proteins).
>
> 1.4. A description of any biologically interesting phenomena observed (such as biological pathways, functional groups, or proteins).

The results were to be submitted via email to the study anonymizer by the end of November 2020. The anonymizer did not share the identities of the participants with anyone, including other members of the iPRG.

Phase 2 (protein database provided)

In the second phase, participants were given some clues about the composition of the sample and were asked to use a provided FAST-all text-based format (FASTA) protein sequence database to answer a set of additional questions. The sequence database contained all species present in the sample and 3 additional species—*Citrobacter freundii*, *Clostridium butyricum*, and *Salmonella bongori*—having varied evolutionary distance from its closest relative among the species in the study sample, and, consequently, varying degrees of overlap of shared peptide sequences. Participants were not told which of the 7 species should have been detectable in the sample. In this second phase, the participants were asked to provide the following:

> 2.1. A detailed description of how you performed the analysis after being provided with the FASTA sequence databases covering the species present in the sample.
>
> 2.2. A list of the species or taxa identified in the sample, along with metrics of their relative abundances (including number of PSMs, distinct peptides or proteins, and/or their confidences).

> 2.3. A description of any biologically interesting phenomena you can observe (such as biological pathways, functional groups, or proteins).

Participants were also asked the following 3 “bonus” questions:

> 2.4. Did you find chimeric tandem mass spectra in the dataset? If so, how did you report the PSM, and how did you decide between the multiple options?
>
> 2.5. There are public datasets for some of the species that are in these samples. Comparing the proteins identified in these samples with those in the public resources, what can you infer about the physiology/state of the organisms in the study samples?
>
> 2.6. For peptides corresponding to important taxonomy or functions in a metaproteomics study, how would you validate them?

The results were to be submitted via email to the study anonymizer by the end of January 2021.

The data files were subsequently deposited in the ProteomeXchange Consortium via the PRIDE[30] partner repository with the dataset identifier PXD034795 and DOI 10.6019/PXD034795.

RESULTS

Submissions were received from 9 participants in the first phase and 8 participants in the second phase; 7 groups participated in both phases. Please see https://osf.io/pze9x for Phase 1 submissions and https://osf.io/w6msx for Phase 2 submissions.

Search engines, databases, and processing strategies

No information was initially provided about sample complexity or taxonomic composition. As such, participants used a wide variety of search engines for the study (Table 1[31],[32],[33],[34],[35],[36]) protein sequence databases (Table 2) and informatics tools for taxonomy and functional analysis (Table 3[37],[38],[39],[40],[41]), as shown below.

Table 1

Search engines used by participants.

Search engine	Citation	Number of participants
COMPIL 2.0	[13]	1
Mascot	[31]	2
MetaProteomeAnalyzer	[20]	1

MS2 Rescore	[32]	1
MS-Fragger	[33]	1
SearchGUI/PeptideShaker	[34],[35]	4
X! Tandem	[36]	1

Table 2

Protein sequence databases used by participants.

Database	Number of sequences	Number of participants
NCBInr	2.9e8	1
Swiss-Prot	5.0e5	3
UniProt Pan-proteome	5.2e7	1
UniRef50	3.9e7	1
UniREF100	1.1e8	1
Wellcome Sanger Institute MetaHIT 3.3	2.0e6	1

Table 3

Informatics tools used by participants.

Category/tool	Citation	Number of participants
Taxonomy		
Prophane	[20]	1
Unipept	[18]	6
Functional analysis		
BlastKOALA	[37]	1
EggNOG Mapper	[22]	1
GhostKOALA	[37]	1

Unipept	[18]	4
Quantitative analysis		
FlashLFQ	[38]	1
NSAF	[39]	1
Spectral counts and relative biomass estimation	[40],[41]	1

Taxonomic composition analysis in both phases of the study

The sample was made up by approximately 48% T4 bacteriophage-infected *E. coli*, 48% *S. enterica*, and 4% *B. subtilis*. A correct ranking of the taxa in order of abundance should, therefore, have *E. coli* and *S. enterica* above *B. subtilis*. Only 2 groups (Figure 2A and Supplementary Table 1) reported the correct species abundance order in the samples. In Phase 1, participants used various metrics to estimate taxonomic abundance, including the number of PSMs, number of peptides, number of proteins, peptide intensities, relative abundance of detected taxa, or the relative cell biomass associated with taxonomic units. Of the 9 participants in Phase 1, all detected *E. coli* in the sample, and all but one found evidence for *S. enterica*. Only 5 participants identified *B. subtilis* in the sample, while 6 participants reported the presence of Enterobacteria phage T4.

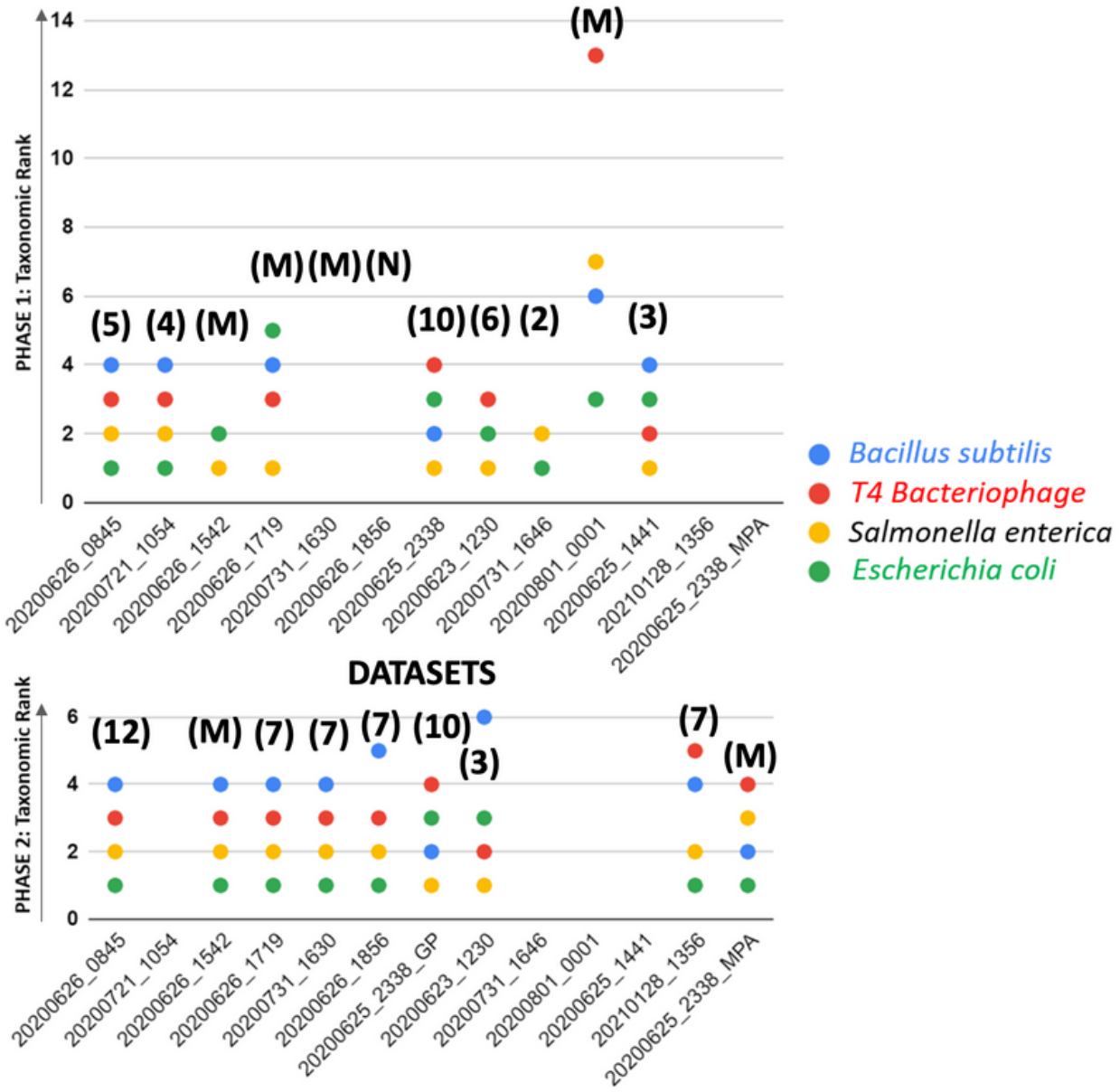

Figure 2

Taxonomy analysis Phase 1 and Phase 2. Rank order of relative abundance among species reported by the participants in Phase 1 (top graph) and Phase 2 (bottom graph). *E. coli* (green), *S. enterica (orange),* T4 bacteriophage (red), and *B. subtilis (*blue) have been represented along with the reported ranking on the *y*-axis and participant identifier on the *x*-axis. The numbers in parentheses indicate the number of species reported by each participant in both the phases (M = multiple species; N = No species reported).

Phase 2 participants were provided with a database containing the protein sequences of the 4 organisms in the samples plus sequences for 3 related organisms that were not present in the samples (see Supplementary Figure 1). In Phase 2, all 8 participants reported the presence of *E. coli*, *S. enterica*, *B. subtilis*, and Enterobacteria phage T4 in the samples. Four of the groups (Figure 2B) determined the correct order of abundance. One

participant (20200626_0845) specifically reported that there was insufficient mass spectrometry evidence for any of the 3 additional species that were included in the provided database but not actually present in the sample as decoys.

Functional analysis and biological interpretation in both the phases of study

For functional analysis in Phase 1, participants reported Gene Ontology terms (taxon specific and nonspecific), heatmap analyses, and Sankey diagrams. Since information was not initially provided about the species present in the samples, we expected to receive a range of observations and conclusions about the dataset. One participant (20200626_0845) submitted a comprehensive functional interpretation in which detected proteins were matched to Kyoto Encyclopedia of Genes and Genomes Ontology (KO) entries, and then KO entries with normalized spectral matches were used for principal component analysis (PCA). PCA separated *E. coli* and *S. enterica*, with 2 primary differences being that (a) OmpC (receptor for T4 bacteriophage) is more abundant in *S. enterica*, and (b) OmpF (receptor for T2 bacteriophage) is more abundant in *E. coli* (Figure 3). Based on this observation (along with the detection of T4 bacteriophage), the participants hypothesized that the interaction of the bacteriophage T4 with the OmpC of *E. coli* impacted the expression of OmpC. Their logic was that *E. coli* might overproduce OmpF to compensate for OmpC functional loss. Since the effect of OmpC expression is specific to *E. coli*, the participants further speculated that bacteriophage T4 specifically interacted with *E. coli*.

Figure 3

Functional analysis. Representative functional analyses as submitted by participant 20200626_0845. Main loadings for the first principal component of a PCA analysis between KO assignments for the 4 samples restricted to *E. coli* and *S. enterica*.

Other reported conclusions for Phase 1 of the study were the following: (a) the samples were generated by infection of *E. coli* and *Salmonella* with T4 bacteriophage (20200625_1441); (b) the T4 bacteriophage selectively infected *E. coli* (20200623_1230); (c) the samples contained *E. coli* and *S. enterica*. (20200731_1646); (d) speculation that the samples were obtained from human gut following foodborne salmonellosis (20200626_1542); and (e) organism-specific functions were detected (20200721_1054), but the participant did not report anything about taxonomy.

For the second phase, the participant who had correctly ascertained that the bacteriophage T4 had specifically interacted with *E. coli* (20200626_0845) reasserted their conclusion. Additionally, 1 participant detected elevated ribosomal *E. coli* proteins, an indication of bacteriophage T4 infection (20210128_1356). Among the other submissions were the following (not all of which are correct): (a) the samples were generated from a laboratory mixture of *Bacillus*, *Escherichia*, and *Salmonella* (20200625_2338); (b) the samples were from an

anaerobic fermenter (20200626_1856); and (c) the samples represented a bacterial mixture cultivated in aerobiosis or insufficient de-aeration for *C. butyricum* (20200731_1630).

While providing additional information about the component organisms resulted in more accurate taxonomy detection among participants in the second phase as compared to the first phase, this did not translate into improved functional analysis or biological interpretation.

DISCUSSION

Through identification and relative quantification of component proteins, metaproteomics can provide insight into how a microbiome responds to its immediate environment. However, numerous challenges remain because of the complexity of the samples, resulting in the need for optimization of sample preparation, data acquisition, and data analysis.[4],[42] The current study was designed to assess the bioinformatic approaches available to address the difficult task of detecting taxonomy and deducing biological interpretation from a metaproteomics dataset.

In preparation for the study, 4 biological replicates of mixtures of bacterial cells were generated, each containing approximately equal quantities of 2 related bacteria (*E. coli* strain B^E and *S. enterica* strain UB-0015) and a substantially lower level of a third bacterium (*B. subtilis* strain 168). Since *E. coli* and *Salmonella* share many homologous proteins, it was anticipated that this would result in the identification of many peptides shared between the 2 bacteria in the downstream analysis of the mass spectral data. To introduce an additional dimension to the study, we used *E. coli* that had been infected with bacteriophage T4 as one of the components of the sample. We anticipated that phage infection of one of the bacteria would alter the quantity of one or more of the host proteins, thereby introducing permutations in the levels of some shared peptides.

Our analysis of the bioinformatic approaches used by participants indicated that there is a need for educating researchers who are new to the field of metaproteomics of best practices for data processing, especially for functional data analysis. In the first phase of the study, only 2 research groups reported the correct rank order of taxonomic abundance in the samples. Interestingly, 1 of the 2 groups accurately deduced the biological implications of the dataset. Other participants concluded that there had been bacteriophage infection of *E. coli*. It is important to note that determining biological relevance was an especially challenging question since no metadata, metagenomic sequencing information, or protein FASTA database were provided in the first phase. In real-life scenarios, metaproteomics searches are often performed against a matched metagenome or data from a public repository,[6] but in this study, we wanted to find out what level of information could be independently deduced from metagenomics data.

In the second phase, participants were much more successful in identifying the taxonomic composition because the necessary protein sequence database file was provided. Four groups correctly reported the taxa and their relative abundance. For a functional analysis though, apart from the group that reported the bacteriophage–*E.*

coli interaction in the first phase, only one more participant speculated that there was evidence for bacteriophage infection.

A valuable outcome of this iPRG 2020 study is the insight it provided into the repertoire of bioinformatics approaches in use for metaproteomics research, since a variety of software tools and data processing pipelines were used (Table 1). The study also highlighted the fact that there is a need for a wider dissemination of knowledge from expert metaproteomics laboratories about ways to process data and how to use the outputs to formulate biologically relevant conclusions.

In 2021, the Metaproteomics Initiative[43] was formally established. The mission of this consortium of microbiome researchers is to disseminate metaproteomics fundamentals, advancements, and applications through collaborative networking in microbiome research. This type of initiative is essential to provide educational resources for researchers who are interested in metaproteomics. Additionally, it brings together researchers of diverse backgrounds, perspectives, and skillsets to achieve higher goals via collaborative efforts.

Based on this study, we feel that the metaproteomics informatics field will benefit from improvements in taxonomy detection tools for mass spectrometry–based metaproteomics datasets. While there was a marked improvement in taxonomic detections in Phase 2, when a protein database had been made available, it is clear that further improvements in taxonomic ranking and accuracy are still needed. The lack of consistent results at the taxonomic level was also evident from the CAMPI Study [42] undertaken by the Metaproteomics Initiative. [43] We believe that there is a need for the generation of ground truth datasets both at taxonomic and functional levels to develop better algorithms and methods for taxonomic detection, protein grouping, and spectral assignment. With this in mind, the Metaproteomics Initiative has recently launched the CAMPI3 study[44] with a focus on taxonomic and functional annotations as well as resolving protein inference and spectral assignment issues in metaproteomics research.

Although proteomics can be considered to be a generally mature discipline, metaproteomics remains one of its more challenging extensions. Nevertheless, the performance of the participants in this study demonstrated that metaproteomics is sufficiently developed to be able to answer a wide range of interesting questions about taxonomic composition. The results from this study can serve as a starting point for larger, in-depth, community efforts in metaproteomics and related fields.

ACKNOWLEDGMENTS

Mass spectrometry analyses were conducted at the University of Texas Health Science Center at San Antonio (UTHSCSA) Institutional Mass Spectrometry Laboratory with the expert technical assistance of Sammy Pardo and Dana Molleur, supported in part by UTHSCSA and the University of Texas System Proteomics Core Network for the purchase of the Orbitrap Fusion Lumos mass spectrometer.

The identification of certain commercial equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology or the ABRF, nor does it imply that the products identified are necessarily the best available for the purpose.

The MS data files are available in the ProteomeXchange Consortium via the PRIDE[30] partner repository with the dataset identifier PXD034795 and 10.6019/PXD034795.

The authors would like to thank all the participants who took the time to analyze and return the data for this study. Some of the participants are the members of the Metaproteomics Initiative (https://metaproteomics.org/), the goals of which are to promote, improve, and standardize metaproteomics.

SUPPLEMENTARY MATERIALS

pdf	Supplementary Section_MP_Metaproteomics iPRG 2020_07132023-61689871685823.pdf	291 KB

References

1. Zhang X, Li L, Butcher J, Stintzi A, Figeys D. Advancing functional and translational microbiome research using meta-omics approaches. *Microbiome*. 2019;7(1):154. doi:10.1186/s40168-019-0767-6. ↩
2. Heintz-Buschart A, Wilmes P. Human gut microbiome: function matters. *Trends Microbiol.* 2018;26(7):563-574. doi:10.1016/j.tim.2017.11.002. ↩
3. Muth T, Benndorf D, Reichl U, Rapp E, Martens L. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol Biosyst.* 2013;9(4):578-585. doi:10.1039/c2mb25415h. ↩
4. Muth T, Kolmeder CA, Salojärvi J, et al. Navigating through metaproteomics data: a logbook of database searching. *Proteomics.* 2015;15(20):3439-3453. doi:10.1002/pmic.201400560. ↩
5. Hardouin P, Chiron R, Marchandin H, Armengaud J, Grenga L. Metaproteomics to decipher CF host-microbiota interactions: overview, challenges and future perspectives. *Genes (Basel).* 2021;12(6):892. doi:10.3390/genes12060892. ↩
6. Blakeley-Ruiz JA, Kleiner M. Considerations for constructing a protein sequence database for metaproteomics. *Comput Struct Biotechnol J.* 2022;20:937-952. doi:10.1016/j.csbj.2022.01.018. ↩
7. Galata V, Busi SB, Kunath BJ, et al. Functional meta-omics provide critical insights into long-and short-read assemblies. *Brief Bioinform.* 2021;22(6):bbab330. doi:10.1093/bib/bbab330. ↩
8. Rechenberger J, Samaras P, Jarzab A, et al. Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant Enterobacteriaceae. *Proteomes.* 2019;7(1):2. doi:10.3390/proteomes7010002. ↩

9. Jagtap P, Goslinga J, Kooren JA, et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics*. 2013;13(8):1352-1357. doi:10.1002/pmic.201200352. ↩
10. Kertesz-Farkas A, Keich U, Noble WS. Tandem mass spectrum identification via cascaded search. *J Proteome Res*. 2015;14(8):3027-3038. doi:10.1021/pr501173s. ↩
11. Potgieter MG, Nel AJM, Fortuin S, et al. MetaNovo: a probabilistic approach to peptide and polymorphism discovery in complex metaproteomic datasets. *PLoS Comput Biol*. 2023;19(6):e1011163. doi:10.1371/journal.pcbi.1011163. ↩
12. Kumar P, Johnson JE, Easterly C, et al. A sectioning and database enrichment approach for improved peptide spectrum matching in large, genome-guided protein sequence databases. *J Proteome Res*. 2020;19(7):2772-2785. doi:10.1021/acs.jproteome.0c00260. ↩
13. Park SKR, Jung T, Thuy-Boun PS, Wang AY, Yates JR III, Wolan DW. ComPIL 2.0: an updated comprehensive metaproteomics database. *J Proteome Res*. 2019;18(2):616-622. doi:10.1021/acs.jproteome.8b00722. ↩
14. Bassignani A, Plancade S, Berland M, et al. Benefits of iterative searches of large databases to interpret large human gut metaproteomic data sets. *J Proteome Res*. 2021;20(3):1522-1534. doi:10.1021/acs.jproteome.0c00669. ↩
15. Kleikamp HBC, Pronk M, Tugui C, et al. Database-independent de novo metaproteomics of complex microbial communities. *Cell Syst*. 2021;12(5):375-383.e5. doi:10.1016/j.cels.2021.04.003. ↩
16. Johnson RS, Searle BC, Nunn BL, et al. Assessing protein sequence database suitability using *de novo* sequencing. *Mol Cell Proteomics*. 2020;19(1):198-208. doi:10.1074/mcp.TIR119.001752. ↩
17. Zhang X, Figeys D. Perspective and guidelines for metaproteomics in microbiome studies. *J Proteome Res*. 2019;18(6):2370-2380. doi:10.1021/acs.jproteome.9b00054. ↩
18. Gurdeep Singh R, Tanca A, Palomba A, et al. Unipept 4.0: functional analysis of metaproteome data. *J Proteome Res*. 2019;18(2):606-615. doi:10.1021/acs.jproteome.8b00716. ↩
19. Saunders JK, Gaylord DA, Held NA, et al. METATRYP v 2.0: metaproteomic least common ancestor analysis for taxonomic inference using specialized sequence assemblies-standalone software and web servers for marine microorganisms and coronaviruses. *J Proteome Res*. 2020;19(11):4718-4729. doi:10.1021/acs.jproteome.0c00385. ↩

20. Schiebenhoefer H, Schallert K, Renard BY, et al. A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. *Nat Protoc*. 2020;15(10):3212-3239. doi:10.1038/s41596-020-0368-7. ↩
21. Sajulga R, Easterly C, Riffle M, Mesuere B, et al. Survey of metaproteomics software tools for functional microbiome analysis. *PLoS One*. 2020;15(11):e0241503. doi:10.1371/journal.pone.0241503. ↩
22. Huerta-Cepas J, Forslund K, Coelho LP, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol*. 2017;34(8):2115-2122. doi:10.1093/molbev/msx148. ↩
23. Huson DH, Weber N. Microbial community analysis using MEGAN. *Methods Enzymol*. 2013;531:465-485. doi:10.1016/B978-0-12-407863-5.00021-6. ↩
24. Riffle M, May DH, Timmins-Schiffman E, et al. MetaGOmics: a web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. *Proteomes*. 2017;6(1):2. doi:10.3390/proteomes6010002. ↩
25. Mehta S, Kumar P, Crane M, et al. Updates on metaQuantome software for quantitative metaproteomics. *J Proteome Res*. 2021;20(4):2130-2137. doi:10.1021/acs.jproteome.0c00960. ↩
26. Easterly CW, Sajulga R, Mehta S, et al. metaQuantome: an integrated, quantitative metaproteomics approach reveals connections between taxonomy and protein function in complex microbiomes. *Mol Cell Proteomics*. 2019;18(8 suppl 1):S82-S91. doi:10.1074/mcp.RA118.001240. ↩
27. Simopoulos CMA, Ning Z, Zhang X, et al. pepFunk: a tool for peptide-centric functional analysis of metaproteomic human gut microbiome studies. *Bioinformatics*. 2020;36(14):4171-4179. doi:10.1093/bioinformatics/btaa289. ↩
28. Cheng K, Ning Z, Zhang X, et al. MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome*. 2017;5(1):157. doi:10.1186/s40168-017-0375-2. ↩
29. Cheng K, Ning Z, Zhang X, et al. MetaLab 2.0 enables accurate post-translational modifications profiling in metaproteomics. *J Am Soc Mass Spectrom*. 2020;31(7):1473-1482. doi:10.1021/jasms.0c00083. ↩
30. Perez-Riverol Y, Bai J, Bandla C, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res*. 2022;50(D1):D543-D552. doi:10.1093/nar/gkab1038. ↩
31. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999;20(18):3551-3567.

doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2. ↩

32. C Silva AS, Bouwmeester R, Martens L, Degroeve S. Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics.* 2019;35(24):5243-5248. doi:10.1093/bioinformatics/btz383. ↩

33. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods*. 2017;14(5):513-520. doi:10.1038/nmeth.4256. ↩

34. Barsnes H, Vaudel M. SearchGUI: a highly adaptable common interface for proteomics search and de novo engines. *J Proteome Res*. 2018;17(7):2552-2555. doi:10.1021/acs.jproteome.8b00175. ↩

35. Vaudel M, Burkhart JM, Zahedi RP, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol*. 2015;33(1):22-24. doi:10.1038/nbt.3109. ↩

36. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics.* 2004;20(9):1466-1467. doi:10.1093/bioinformatics/bth092. ↩

37. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol*. 2016;428(4):726-731. doi:10.1016/j.jmb.2015.11.006. ↩

38. Millikin RJ, Solntsev SK, Shortreed MR, Smith LM. Ultrafast peptide label-free quantification with FlashLFQ. *J Proteome Res*. 2018;17(1):386-391. doi:10.1021/acs.jproteome.7b00608. ↩

39. Gokce E, Shuford CM, Franck WL, Dean RA, Muddiman DC. Evaluation of normalization methods on GeLC-MS/MS label-free spectral counting data to correct for variation during proteomic workflows. *J Am Soc Mass Spectrom.* 2011;22(12):2199-2208. doi:10.1007/s13361-011-0237-2. ↩

40. Kleiner M, Thorson E, Sharp CE, et al. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat Commun*. 2017;8(1):1558. doi:10.1038/s41467-017-01544-x. ↩

41. Pible O, Allain F, Jouffret V, Culotta K, Miotello G, Armengaud J. Estimating relative biomasses of organisms in microbiota using "phylopeptidomics". *Microbiome*. 2020;8(1):30. doi:10.1186/s40168-020-00797-x. ↩

42. Van Den Bossche T, Kunath BJ, Schallert K, et al. Critical assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows. *Nat Commun*. 2021;12(1):7305. doi:10.1038/s41467-021-27542-8. ↩

43. Van Den Bossche T, Arntzen MØ, et al. The Metaproteomics Initiative: a coordinated approach for propelling the functional characterization of microbiomes. *Microbiome.* 2021;9(1):243. doi:10.1186/s40168-021-01176-w. ↩

44. Metaproteomics Initiative. CAMPI 3 kickoff. Metaproteomics Initiative. June 7, 2023. Accessed July 27, 2023. https://metaproteomics.org/news/2023-06-07-campi3-announcement/ ↩