Critical Assessment of Metaproteome Investigation (CAMPI): A Multi-Lab Comparison of Established Workflows

*Tim Van Den Bossche[*1,2], Benoit J. Kunath*[3], Kay Schallert*[4], Stephanie S. Schäpe*[5], Paul E. Abraham[6], Jean Armengaud[7], Magnus Ø. Arntzen[8], Ariane Bassignani[9], Dirk Benndorf[4,10,11], Stephan Fuchs[12], Richard J. Giannone[6], Timothy J. Griffin[13], Live H. Hagen[8], Rashi Halder[3], Céline Henry[9], Robert L. Hettich[6], Robert Heyer[4], Pratik Jagtap[13], Nico Jehmlich[5], Marlene Jensen[14], Catherine Juste[9], Manuel Kleiner[14], Olivier Langella[15], Theresa Lehmann[4], Emma Leith[13], Patrick May[3], Bart Mesuere[1,16], Guylaine Miotello[7], Samantha L. Peters[6], Olivier Pible[7], Udo Reichl[4,11], Bernhard Y. Renard[12,17], Henning Schiebenhoefer[12,17], Alexander Scryba[18], Alessandro Tanca[19], Kathrin Trappe[12], Jean-Pierre Trezzi[3,20], Sergio Uzzau[19], Pieter Verschaffelt[1,16], Martin von Bergen[5], Paul Wilmes[3], Maximilian Wolf[4], and Lennart Martens[#1,2] and Thilo Muth[#21]*

* Shared first authors
\# Shared last authors, corresponding authors (lennart.martens@ugent.be, thilo.muth@bam.de)
Intermediate authors are listed alphabetically

(1) VIB - UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
(2) Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
(3) Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg
(4) Bioprocess Engineering, Otto-von-Guericke University Magdeburg, Magdeburg, Germany
(5) Helmholtz-Centre for Environmental Research - UFZ GmbH, Department of Molecular Systems Biology, Leipzig, Germany
(6) Chemical Sciences Division, Oak Ridge National Lab, Oak Ridge, TN, USA
(7) Université Paris Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), SPI, 30200 Bagnols-sur-Cèze, France
(8) Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences (NMBU), Ås, Norway
(9) Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350, Jouy-en-Josas, France.
(10) Microbiology,Department of Applied Biosciences and Process Technology, Anhalt University of Applied Sciences, Köthen, Germany
(11) Bioprocess Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
(12) Bioinformatics Unit (MF1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany
(13) Department of Biochemistry Molecular Biology and Biophysics, University of Minnesota, MN, USA
(14) Department of Plant & Microbial Biology, North Carolina State University, Raleigh, USA
(15) Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon, 91190, Gif-sur-Yvette, France
(16) Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium
(17) Data Analytics and Computational Statistics, Hasso-Plattner-Institute, Faculty of Digital Engineering, University of Potsdam, Potsdam, Germany
(18) Faculty of Technology, Bielefeld University, Bielefeld, Germany
(19) Department of Biomedical Sciences, University of Sassari, Sassari, Italy
(20) Integrated Biobank of Luxembourg, Luxembourg Institute of Health, 1, rue Louis Rech, L-3555, Dudelange, Luxembourg.
(21) Section eScience (S.3), Federal Institute for Materials Research and Testing, Berlin, Germany

Abstract

Metaproteomics has matured into a powerful tool to assess functional interactions in microbial communities. While many metaproteomic workflows are available, the impact of method choice on results remains unclear.

Here, we carried out the first community-driven, multi-lab comparison in metaproteomics: the critical assessment of metaproteome investigation study (CAMPI). Based on well-established workflows, we evaluated the effect of sample preparation, mass spectrometry, and bioinformatic analysis using two samples: a simplified, lab-assembled human intestinal model and a human fecal sample.

We observed that variability at the peptide level was predominantly due to wet-lab workflows, with a smaller contribution of bioinformatic pipelines. These peptide-level differences largely disappeared at protein group level. While differences were observed for predicted community composition, similar functional profiles were obtained across workflows.

CAMPI demonstrates the robustness of present-day metaproteomics research, serves as a template for multi-lab studies in metaproteomics, and provides publicly available data sets for benchmarking future developments.

Main

Microbial communities play a primary role in global biogeochemical cycling and form complex interactions that are crucial for the development and maintenance of health in humans, animals, and plants. To fully understand microbial communities and their interplay with their environment requires knowledge not only of the microorganisms involved and their biodiversity, but also of their metabolic functions at both the cellular and community level[1]. As proteins constitute the key operational units performing these functions, metaproteomics has emerged as the most relevant approach to characterize the functional expression of a given microbiome[2,3]. Metaproteomics corresponds to the large-scale characterization of the entire set of proteins accumulated by all community members at a given point in time, known as the metaproteome[4]. Since its first introduction in 2004[5], mass spectrometry (MS)-based metaproteomics has quickly emerged as a powerful tool to functionally characterize a broad variety of microbial communities *in situ.* This allows a direct link to the phenotypes on a molecular level and shows the adaptation of the microorganisms to their specific environment[6]. Metaproteomics thus complements other meta-omic approaches such as metagenomics and metatranscriptomics, as these only have the exploratory power to assess the diversity and functional potential of microorganisms, but cannot observe their actual phenotypes[7].

In metaproteomics, proteins are commonly measured using a bottom-up approach in which proteins are first extracted, isolated, and digested into peptides. These peptides are then separated and analyzed using liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). The resulting MS/MS spectra are typically matched against *in silico* generated spectra derived from a protein sequence database to identify the analyzed peptides and infer the original proteins. The inferred proteins are then used to describe the various active taxa in the community, their functions, and the relative gene expression levels[8].

Each of the aforementioned steps can potentially influence the outcomes of a metaproteomic analysis and every step brings specific benefits as well as challenges. As a result, multiple workflows have been established. While such diversity brings flexibility, it also complicates the comparison of results across different experiments. Wet-lab challenges include protein recovery due to the presence of different matrices[9], the

presence of different types of microorganisms with different optimal lysis conditions[10,11], and limited depth of analysis[3] and quantification[12] due to an increased sample complexity. Moreover, metaproteomics comes with many specific bioinformatic challenges[8,13]. First, the choice of an appropriate sequence database is critical for peptide identification[14,15]. Typically, large databases can strongly impact sensitivity and false discovery rate (FDR) estimation[16], while incomplete reference databases can lead to missing or false positive identifications[17,18]. Second, the protein inference problem[19] is more pronounced in metaproteomics due to many homologous proteins from closely related organisms[20]. As a result, several dedicated bioinformatic tools have been developed or extended for metaproteomic analysis[21–28]. Despite these challenges, the added value of metaproteomics has already been demonstrated in numerous examples from both the environmental and medical fields, providing unprecedented insights into the functional activity of microbial communities[7,20,29–41].

Nevertheless, a lingering concern is the potential risk of unintended, approach-based biases inherent in various metaproteomic workflows. This is important because reproducibility is key to translate metaproteome studies into applications (*e.g.* clinical or industrial). Consequently, a comprehensive evaluation of widely-used workflows is required to assess their respective outcomes. In the past, various reference data sets from defined microbial community samples (i.e., for which the composition is known *a priori*) have been used in individual benchmarking studies[42–44]. However, a ring trial with different laboratories involved has not yet been performed in the field of metaproteomics. To fill this gap, the 3rd International Metaproteomics Symposium (December 2018, Leipzig, Germany) hosted a multi-lab benchmarking study in the form of a community challenge. Participating laboratories received two microbial samples: a simplified mock community simulating the gut microbiome (SIHUMIx) and a complex, natural stool sample (fecal sample). Each group was allowed to use any preferred sample preparation, analysis, and data evaluation pipeline.

Here, we describe the results of this community-driven study, referred to as the Critical Assessment of MetaProteome Investigation (CAMPI). We compare and discuss the employed workflows covering all analysis steps from sample preparation to the bioinformatic identification and quantification. Moreover, we compare the metaproteome

results with sequencing read-based analyses (metagenomics and metatranscriptomics). We found that meta-omics databases performed better than public reference databases across both samples. More importantly, even though larger differences were observed in identified spectra and unique peptide sequences, the different protein grouping strategies and the functional annotations provided similar results across the provided data sets from all laboratories. When minor differences could be observed, these were largely due to differences in wet-lab methods and partially to bioinformatic pipelines. Finally, for the taxonomic comparison, we found that overall profiles were similar between read-based methods and proteomics methods, with few exceptions.

Apart from these immediate conclusions, the CAMPI study also delivers highly valuable benchmark data sets that can serve a foundation for future method development for metaproteomics.

Results

At the 3rd International Metaproteome Symposium in December 2018, individual lab outcomes of a collaborative, multi-lab effort to compare metaproteomic workflows were presented. In this study, metaproteomics data was acquired in seven labs, using a variety of well-established platforms. **Figure 1** provides a general overview of the study design showing (i) the provision of two types of samples (SIHUMIx and fecal) to the study participants, (ii) the various experimental workflows of biomolecule extraction and MS/MS acquisition, and (iii) the bioinformatic processing steps from protein database generation to database search identification and follow-up analyses (more details in the **Methods**, see **Supplementary Table 1** for an overview of all methods).

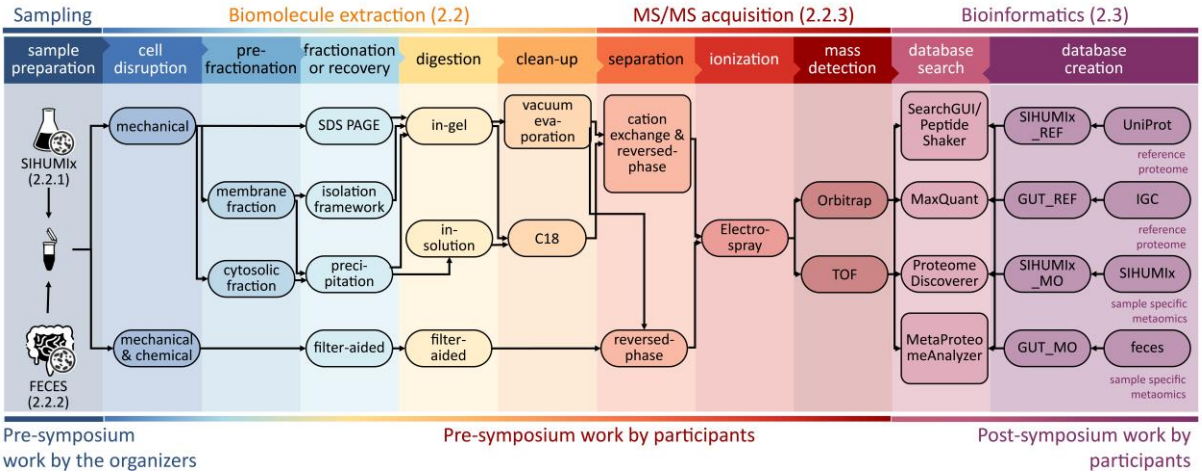

Figure 1. Schematic representation of the main sample preparation steps and follow-up analyses of the CAMPI study. The figure consists of three parts: (i) Pre-symposium work by the organizers (left panel). The two samples (SIHUMIx and FECES) were, prior to the symposium, aliquoted and distributed over the participating labs. (ii) Pre-symposium work by participants (middle panels). Every used method by the participants, going from cell disruption to mass detection, is displayed. (iii) Post-symposium work by participants (right panel). The bioinformatics analyses, i.e. database creation and database search for peptide and protein identification, were harmonized to make the results between all participating laboratories comparable.

At the Symposium, the decision was made to re-analyse the acquired data with different bioinformatics pipelines, to obtain the first multi-laboratory effort in metaproteomics to independently evaluate available methodological and computational approaches, in line with similar community-driven benchmarking studies[45–48]. In the first Results section, we analyzed 42 raw files (21 for the SIHUMIx sample and 21 for the fecal sample) from 24 different workflow combinations with X!Tandem using either public or in-house generated

protein databases (see **Figure 1** for a general overview, and **Figure 2** for the results; see online Methods for the database construction). A more in-depth comparison of sample preparations, bioinformatic pipelines, and taxonomic and functional annotations using a sub-selection of ten data sets is available after the first Results section.

Complex wet-lab workflows and sample-specific meta-omic search databases lead to more identifications

In order to study the effect of the different wet-lab and LC-MS/MS workflows on the identification outcome, we searched all submitted MS files using the widely used X!Tandem search engine[49]. To investigate the influence of the chosen database, we searched each file against a publicly available reference database (SIHUMIx_REF and GUT_REF) and against a multi-omic database (SIHUMIx_MO and GUT_MO). The comparison of all CAMPI workflows is displayed in **Figure 2** (raw data in **Supplementary Table 2**).

Figure 2. Comparison of identification rates across all CAMPI workflows. On the left side, the bar charts show the number of identified spectra using the reference (REF) database (orange), the number of identified spectra using the multi-omic (MO) database (dark blue) and total amount of measured spectra (red). On the right side, the light blue bars represent the identification rate calculated as the percentage of spectra that yielded a peptide identification at 1% FDR for both the REF database (orange) and the MO database (dark blue). The specific protocols can be found in **Supplementary Table 1**. For database searching, X!Tandem was used as a single search engine.

The results greatly differed between the samples and workflows in terms of absolute numbers of acquired spectra, identified spectra, and relative amount of identified spectra (identification rates). For the SIHUMIx data set, the number of acquired spectra varied between 37k to 260k, and identification rates varied between 29.99% and 68.64% for SIHUMIx_REF and between 32.52% and 73.34% for SIHUMIx_MO. For the fecal data set, between 9k and 223k spectra were acquired, with identification rates between 11.99% and 34.79% for GUT_REF, and between 15.70% and 40.49% for GUT_MO.

The differences in acquired spectra show a clear relation to the method used, as similar methods or replicates show highly similar numbers of acquired spectra. As expected, more complex methods with longer gradient lengths, fractionation, and additional separation methods such as MudPIT[50] or ion mobility (PASEF)[51] led to up to eight times more identified spectra, but at the cost of increased time and resources spent[52] (see **Supplementary Table 1** for a detailed description, and **Supplementary Table 2** for an overview of the samples). Notably, identification rates were not necessarily correlated with the total number of identifications. For example, between analyses S03 and S05, which used a 260 min and 460 min LC gradient length, respectively, a higher absolute number of identified spectra was found for the 460 min gradient, but also a lower identification rate. As expected, if an MS instrument is provided with the ability to acquire more spectra, it will do so. However, the gains in spectral acquisition do not readily translate into gains in identification. There is thus a potential for diminishing returns when going for more complex methods. There is also a somewhat consistent drop in number of acquired spectra of around 10% when comparing SIHUMIx samples with fecal samples for similar workflows (e.g., S09-S10 with F05-F06, and S13 Reps 1-3 with F09 Reps 1-3). However, occasionally this drop is much greater, as for S11_Fract1-4 and F07_Fract1-4. The overall limited drop might be attributed to the higher complexity of the fecal sample, and corresponding ion suppression effects. The differences in identification rate are likely to be derived from the choice of the search database. The identification rates for the publicly available databases were invariably lower, which is due to their larger and less specific search space, consistent with literature[14,16,18,42,53]. Here, these public reference databases (SIHUMIx_REF and GUT_REF) contained 1.6 and 16 times, respectively,

more unique *in silico* digested peptides than the corresponding multi-omic databases (SIHUMIx_MO and GUT_MO) (**Supplementary File 1**).

Overall, our results indicate that generating a sample-specific meta-omic database can be advantageous for complex metaproteomics samples, such as the human gut microbiome, and even more so for complex and poorly characterised samples such as soil microbiota. The smaller meta-omic databases require less computational resources (e.g., CPU and RAM) and tend to be more accurate due to their tailored composition. However, for their generation, meta-omic databases require additional experimental and computational resources, and are often not as well assembled and/or annotated as reference databases. Because the composition of SIHUMIx was known, the benefit of using a tailored meta-omic database was limited and the analysis was feasible with available reference proteomes. In contrast, the community for the fecal sample was unknown, which represents the typical scenario in metaproteomics.

For known reference samples (such as SIHUMIx), it is therefore reasonable to simply use the reference database, while the largely unknown fecal sample community is best analysed using a tailored meta-omic database. In the following sections, we thus opted to use only the SIHUMIx_REF and GUT_MO search databases for SIHUMIx and fecal data sets, respectively.

Different bioinformatic pipelines resulted in highly similar peptide identifications

To investigate the effect of the bioinformatic pipelines on peptide identification, we compared the two data sets with the most identified peptides (S11 and F07) (**Figure 3**). To ensure a robust and reliable comparison, we fixed the search parameters for the four different bioinformatic pipelines employed (see online Methods for details).

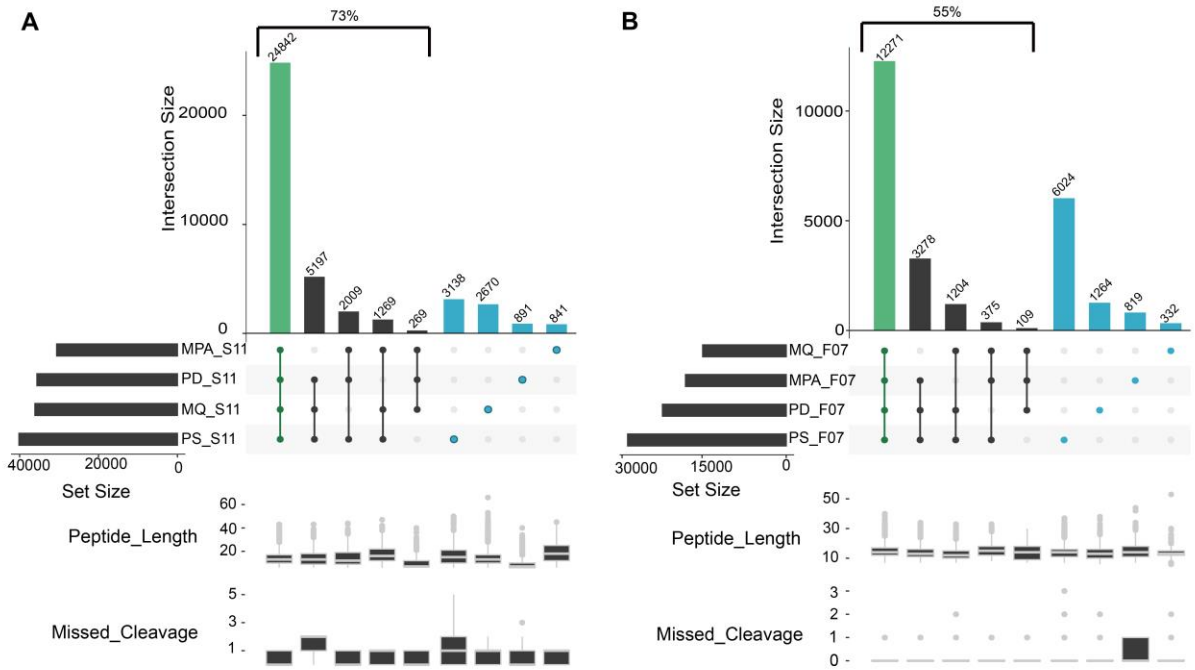

Figure 3. UpSet plot comparison of identified sets of peptides using different bioinformatic pipelines. The left panel displays the results for the SIHUMIx sample S11 (A), while the right panel corresponds to the results for the fecal sample F07 (B). The four different bioinformatic pipelines (MetaProteomeAnalyzer (MPA, using X!Tandem and OMSSA), Proteome Discoverer (PD, using SequestHT), MaxQuant (MQ, using Andromeda), SearchGUI/PeptideShaker (PS, using X!Tandem, OMSSA, MS-GF+, and Comet)) are indicated on the x-axis and sorted by increasing set size. Set size corresponds to the total number of peptides identified per tool, and intersection size corresponds to the number of shared peptides identified in different approaches. Green highlights the intersection, and blue shows unique peptides to each tool. The lower panel box plots show peptide lengths, and number of missed cleavages for each intersection.

For SIHUMIx, the majority of the identified peptides (54.2%) were found by all four bioinformatic pipelines (**Figure 3A**), while this ratio dropped to 40% for the more complex fecal F07 sample (**Figure 3B**). As expected, this percentage increased to 73% and 55%, respectively, when considering the peptides identified by at least three out of four tools. Interestingly, 16% of the peptides were uniquely identified by a single bioinformatic pipeline for the S11 data set (3138, 2670, 891, and 841 peptides for SearchGUI/PeptideShaker, MaxQuant, Proteome Discoverer, and MPA, respectively), while this was 27% for the F07 data set (6024, 1264, 819, and 332 peptides for the same pipelines, respectively). The number of search engines varies between pipelines, with one for MaxQuant (Andromeda) and ProteomeDiscoverer (SequestHT), two for MPA (X!Tandem, OMSSA), and four for SearchGUI (X!Tandem, OMSSA, MS-GF+, and

Comet). Furthermore, each algorithm uses its own score as a quality metric for finding the best matching peptide for a spectrum. This score varies between the search engines and can even result in different peptide identifications for the same spectrum[54].

Overall, the combination from multiple search engines as performed by SearchGUI/PeptideShaker (four algorithms) resulted in the highest number of identifications, which is in line with the previous studies in proteomics and proteogenomics[55,56]. This effect may be attributable to algorithms with more sophisticated scoring methods (e.g., MS-GF+[57] used in SearchGUI, but not in MPA), which generally lead to more identifications overall. However, we do expect that novel search engines based on machine learning algorithms can still boost the number of peptide identifications in the field of metaproteomics[58].

Additionally, we compared the pipelines in terms of peptide features using the peptide lengths and the number of missed cleavages (lower panels of **Figure 3A** and **3B**). While few outliers could be observed (*e.g.* peptide length over 50 AA for MaxQuant and missed cleavages over two for SearchGui/PeptideShaker and ProteomeDiscoverer), the features were overall equally distributed between pipelines. Most of the differences thus seemed to be simply linked to the search engines used.

Because the SearchGUI/PeptideShaker combination provided the most identifications, relatively few identifications were missed by excluding the other three pipelines. We therefore preferred to only use the results of the SearchGUI/PeptideShaker pipeline in the following sections, which investigate the effect of different wet-lab workflows on downstream peptide identifications. These analyses are performed on ten representative data sets that have been selected based on their type of fractionation and MS instrument. These include six SIHUMIx, and four fecal data sets (**Supplementary Table 2**).

Differences between laboratory workflows are mostly attributable to low abundance proteins

After we ruled out bioinformatic workflows as a source of significant difference between samples, we investigated differences arising from different laboratory workflows. We compared the overlap and uniqueness of identifications at the level of peptides, protein subgroups, and the 50% most abundant protein subgroups for the selected laboratory

workflows in **Figure 4**. The figure shows how many peptides and protein subgroups are uniquely identified by a single laboratory workflow and how many are identified by all laboratory workflows.

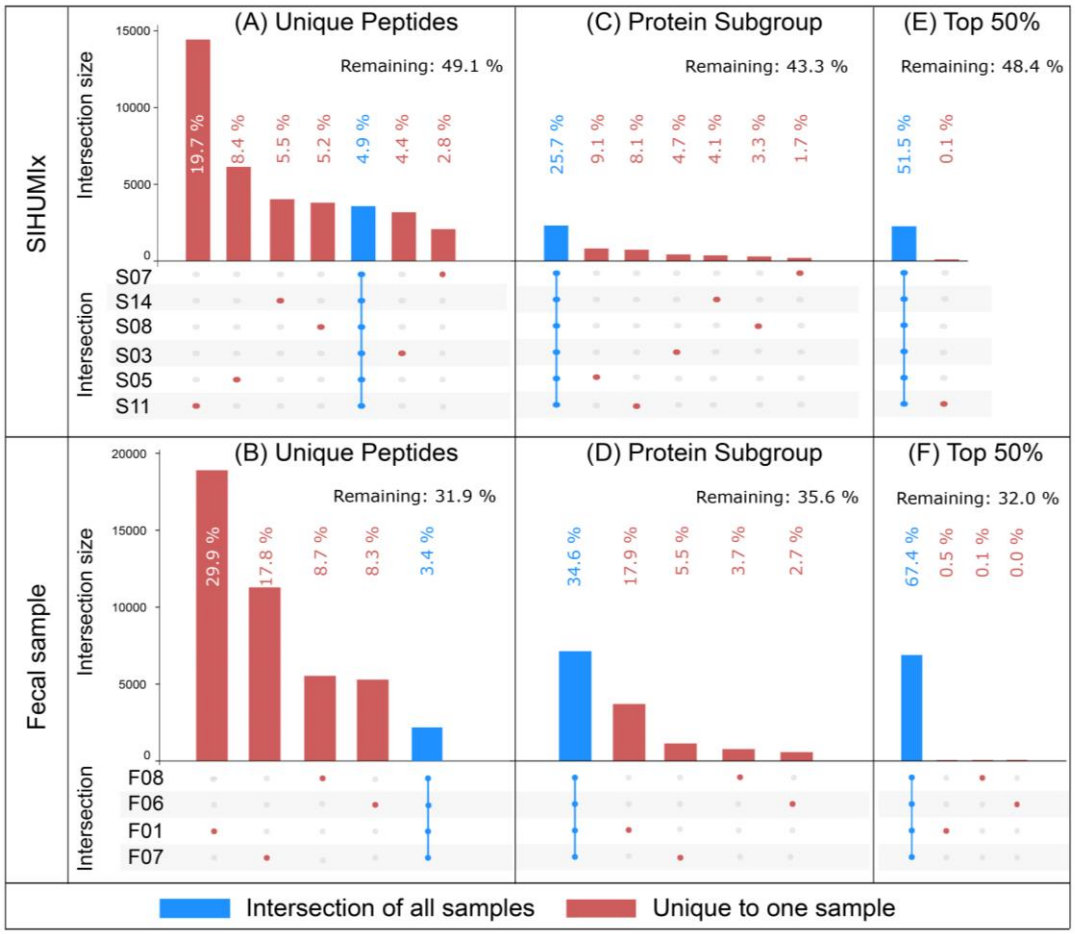

Figure 4. UpSet plot comparison of sets of identified peptides (A and B), protein subgroups (C and D), and 50% most abundant protein subgroups based on spectral counts (E and F). The figure is based on the identifications obtained using SearchGUI/PeptideShaker. The intersection size displays the number of features shared in an intersection. An intersection corresponds to features shared across multiple samples. This figure only displays features unique to a sample (red dot), and shared across all samples (blue bar overlapping all points).

At the peptide level (**Figure 4A and B**), more complex workflows, such as those with longer gradient length and fractionation, identified the most peptides in general (as shown earlier in **Figure 2**) as well as the most workflow-specific peptides, thus limiting the potential for overlap. The number of identified peptides shared between all workflows was quite limited: only 3,557 peptides (4.9% of all identified peptides) in the SIHUMIx data sets, and 2,186 peptides (3.4% of all identified peptides) in the fecal data set. At the

protein subgroup level (**Figure 4C and D**), the intersections of protein subgroups shared across all workflows were 25.7% and 34.6% for the SIHUMIx and fecal data sets, respectively. These percentages increased to 51.5% and 67.4% when we only considered the 50% most abundant protein subgroups (**Figure 4E and F**). Large differences between laboratory workflows observed at the peptide level were thus attenuated at the protein subgroup level, and further reduced for the 50% most abundant protein subgroups. This trend was also clearly visible when considering all intersections, including partial agreement among some samples (**Supplementary Figures 1 and 2**). Of note is that the data sets that only differed in a single laboratory method parameter such as LC gradient length (S03 and S05), or fractionation (F06 and F07) showed a much higher overlap. Also, the number of protein subgroups identified uniquely in a single sample mostly disappeared when only considering the 50% most abundant subgroups. We investigated this further by analyzing the agreement between samples at all Top-N-% values (**Supplementary Figure 3**). A clear trend emerged: the lower the agreement between samples on a given subgroup, the lower the abundance of this subgroup. Furthermore, subgroups that were identified with a single peptide - and therefore usually at the lowest abundance - track very closely with the subgroups identified in only a single sample. Finally, when considering the actual spectral abundance of subgroups, those subgroups that were found in all samples also explained at least 77% of the identified spectra. It is therefore clear that the low agreement between samples at the peptide level is mostly attributable to the identification of low abundant proteins. The complexity of the samples and the limited speed of mass spectrometers in DDA mode led to stochasticity in precursor selection at the low end of the dynamic range. Low abundant protein subgroups with only one peptide thus behave more like peptides, where stochastic selection causes large differences between samples. It is worth noting that this issue is completely avoided by only selecting the Top 50% of protein subgroups. Overall, it can be concluded that while different laboratory workflows provide very different peptide identifications, the protein subgroups are well preserved.

Because protein grouping plays such an important role in translating peptide identifications into biologically meaningful information, we decided to analyze two

commonly used grouping methods in more detail. Protein grouping is achieved using the algorithms PAPPSO[59] and MPA[26] (see **Supplementary Note 1.3**). These two methods use different rules for protein inference: PAPPSO uses Occam's razor, and MPA uses Anti-Occam's razor[60]. The first approach provides a minimum set of proteins that explains the presence of the detected peptides, while the second approach keeps all proteins matched by at least one peptide. Both PAPPSO and MPA can create two types of protein groups: comprehensive groups based on at least one shared peptide, and more specific subgroups based on a complete shared peptide set. Subgroups were deemed more suitable for this analysis, as comprehensive groups collated proteins that were too heterogeneous leading to diverse protein functions within the same group (see **Supplementary Table 3, Supplementary File 2**). This might not be the case for smaller data sets, as a smaller data set also decreases the chance for peptides that link highly dissimilar proteins together. For the SIHUMIx samples, the two protein grouping methods PAPPSO and MPA provided very similar numbers of both protein groups and subgroups, while substantial differences were found for the fecal samples (**Supplementary Table 4**). While cross sample correlation confirmed that the impact of bioinformatic pipelines on the analysis here was negligible, little else could be learned from this correlation analysis (**Supplementary Figure 4 and 5**). To shed some light on these differences between protein grouping methods, we analyzed the agreement between samples for different grouping approaches (**Supplementary Figure 6 and 7**). Notably, when applied to the fecal sample, both grouping algorithms resulted in an unusually high number of protein groups that are unique to F10. However, it remains unclear which of these approaches is better able to capture the actual composition of the sample, or even if the performance of the approaches varies for different types of samples. Because PAPPSO grouping removes likely wrong identifications from homologues, it could be more appropriate for single-organism proteomics or for taxonomically well-defined samples like SIHUMIx, while the grouping from MPA could be more appropriate for complex, unknown samples like the fecal sample (where shared peptides become much more likely) as it retains all information for the grouping (**Supplementary note 1.3**). To conclude, both protein grouping methods provide highly similar results for the SIHUMIx sample, but diverge on

the fecal sample, likely due to the increased complexity of the protein inference task in the latter.

Comparison of meta-omic methods reveals differences between peptide and protein-derived analysis of taxonomic community composition

To determine if differences between workflows have an effect on the overall biological conclusions, we quantitatively compared the identified taxa for each selected sample from both data sets, and this at the peptide, the protein subgroup, and the sequencing read level.

We found different trends between the SIHUMIx and fecal samples (**Figures 5 and 6**). For SIHUMIx, the taxonomic distributions were relatively similar between the metagenomic read, peptide, and protein group levels based on the principal component analysis. Hierarchical clustering highlighted clusters of samples, with the peptide and protein subgroup profiles for samples S07 and S14 clustering with the read-based profile (**Figure 5A**) (**Supplementary Figure 8A and B**). Interestingly, samples with more complex wet-lab methods (S03, S05 and S08) did not show clustering between the peptide and the protein subgroups level. While species were found to be similar between methods overall, there were some notable differences (**Figure 5B**). All methods agreed that *Bacteroides thetaiotaomicron* was the most abundant species, and found *Escherichia coli* at 10-13% abundance. However, differences were found for *Blautia producta*, which was barely found by the proteomics methods, while found at around 5% abundance by metagenomics. It is interesting to consider that this might be caused by the construction of the reference database: at the moment of construction, the UniprotKB reference proteome of *Blautia producta* was not available, and multiple *Blautia sp.* proteomes were therefore provided instead. When looking at the Unipept results in detail, 15% of the peptides were associated with the genus Blautia **(Supplementary Table 5)**, which indicates that the lower identification of *Blautia producta* at the peptide level is due to difficulties in resolving *Blautia* at the species level, rather than a lack of identified *Blautia* peptides during the metaproteomic search. Additionally, a false negative was observed for *Clostridium butyricum*, which was not found by the read-based method, while two false positives were observed for *Clostridiales bacterium* and *Bacteroides dorei*, which were

found by the protein-centric method even though these are not present in the SIHUMIx sample. However, these last two were both found at very low abundance.

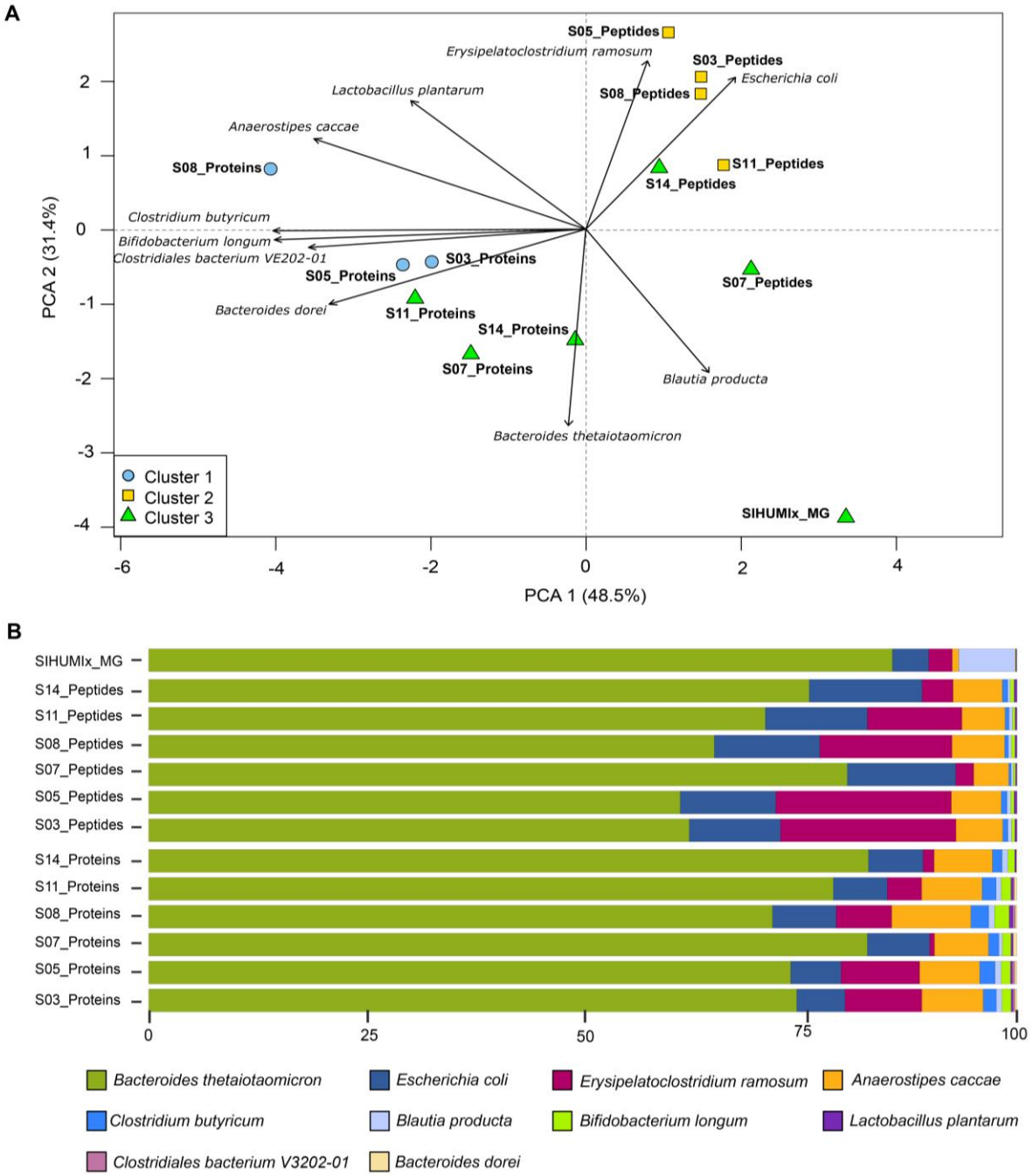

Figure 5. Comparisons of community composition for SIHUMIx. The upper panel shows PCA clustering of the results (A). Different approaches and tools used for taxonomic annotation (mOTU2, Unipept and Prophane) are indicated in the label. Clusters (k=3) were calculated using manhattan distance and are represented by blue, yellow, and green. Features not annotated at species level were considered unclassified and discarded for PCA calculation. Unclassified features accounted for 24.2% and 69.9% of data for peptide and protein subgroup levels. Variables driving differences between samples are represented by black arrows. The lower panel details taxonomic profiles of each sample as bar plots (B).

For the fecal data set, which was grouped at the family level, relatively distinct assessments of community composition were obtained from the read-based, peptide, and protein subgroup levels (**Figure 6A**). While the same families were identified, these had different proportions across methods (**Figure 6B**). Metatranscriptomic information (Feces_MT) was available for the fecal sample and RNA and DNA results were closely colocated, while proteins and peptides were spread out from the read-based methods, but also from each other (**Figure 6A**). The difference between metagenomics/metatranscriptomics and metaproteomics is not surprising because these different methods highlight community profiles from different angles. As already shown before, metagenomics provides a good assessment of community composition in terms of cell numbers for each species, while metaproteomics reflects proteinaceous biomass for each species[43].

Strikingly, for the fecal samples, the community composition as quantified at the peptide level proved to be more similar to the read-based than to the protein-based composition (**Figure 6A**) (**Supplementary Figure 9A and B**). This discrepancy is likely due to the fundamental issue of protein inference. Indeed, in metaproteomics, identification and quantification usually rely on discriminative peptides. As the data sets get more complex, higher levels of sequence homology for many proteins will be observed and will lead to a much greater level of peptide degeneracy across taxonomies[61]. Direct taxon inference from peptides thus likely results in more stringent taxonomy filtering, due to the necessity to rely only on taxon-specific peptides. In fact, the proportion of unclassified peptides between the SIHUMIx and the fecal samples went up from 24.2% to 73.4% due to the increased taxonomic complexity of the fecal data set. In contrast, the proportion of unclassified protein subgroups went down from 69.9% for SIHUMIx to 9.5% for the fecal samples. This latter difference, while large, is not that surprising because the fecal sample considered protein subgroups at the family level, while the SIHUMIx sample considered protein subgroups at the species level, and at also only considered the list of SIHUMIx species, thus greatly limiting peptide-level degeneracy. For the fecal sample, proteins within a subgroup are usually associated to the same family, which explains the higher proportion of protein subgroups that can be classified for the fecal samples. Additionally,

regarding quantification, protein grouping for the fecal samples was done using MPA, which includes all peptides (shared as well as unique), while peptide level quantification only took into account taxon-specific peptides.

Altogether, the degree of degeneracy at the peptide level combined with the grouping method employed for the proteins leads to a different amount of features used for each analysis and thus to different composition profiles between peptide-centric and protein-centric approaches.

Ultimately, due to the sequence homology issue, worse taxonomic resolution will be available for larger, more complex data sets as illustrated in the differences between the SIHUMIx and the fecal data sets. A promising approach to tackle these limitations can take advantage of shared rather than taxon-specific peptides (and thus avoiding the previously mentioned issues) to assess the biomass content of a given community[61] However, regardless of the chosen approach, it is clear that a higher level of peptide coverage will be quite helpful for higher resolution taxonomic annotation, and that metaproteomics will therefore benefit from focusing on analysis depth at the peptide level.

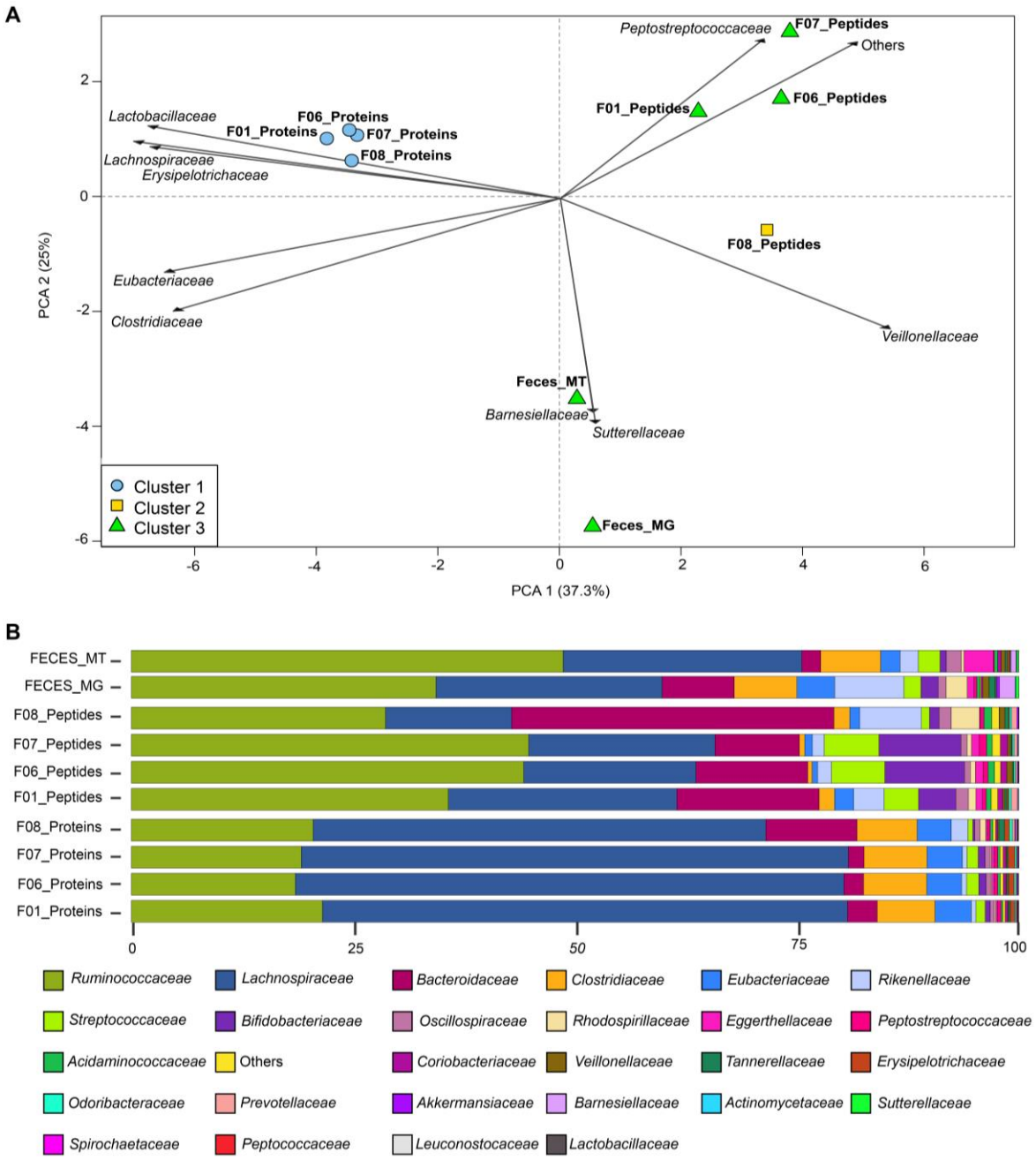

Figure 6. Comparisons of community composition for SIHUMIx. The upper panel shows PCA clustering of the results (A). Different approaches and tools used for taxonomic annotation (mOTU2, Unipept and Prophane) are indicated in the label. Clusters (k=3) were calculated using manhattan distance and are represented by blue, yellow, and green. Features not annotated at species level were considered unclassified and discarded for PCA calculation. Unclassified features accounted for 73.4% and 9.5% of data for peptide and protein subgroup levels. The top 10 variables driving differences between samples are represented by black arrows. The lower panel details taxonomic profiles of each sample as bar plots (B).

The functional profile is similar between different workflows

A major strength of metaproteomics is the ability to provide functional information that reflects the phenotype of the analyzed sample. In order to investigate the influence of post-processing steps on this functional information, we compared functional community profiles on both the SIHUMIx and the fecal samples (**Figure 7**). We observed that the functional similarity between data sets acquired with different workflows on each sample is extremely high, and this regardless of the approach chosen. For the peptide-centric approach, we compared the Gene Ontology (GO) terms (GO domain “biological process”) provided by Unipept for each of the identified peptides with MegaGO[62], resulting in MegaGO similarities of 0.96 or higher. Notably, 95% of the identified peptides were associated with at least one GO term. For the protein-centric approach, the protein families (PFAM) annotations provided by Prophane were compared, resulting in Pearson correlations of 0.98 or higher and Spearman correlations of 0.64 or higher. This continues the trend already observed in **Figure 4**: while peptide identifications may differ greatly between samples, the underlying biological meaning reflected by functional annotations are highly similar across different analysis workflows.

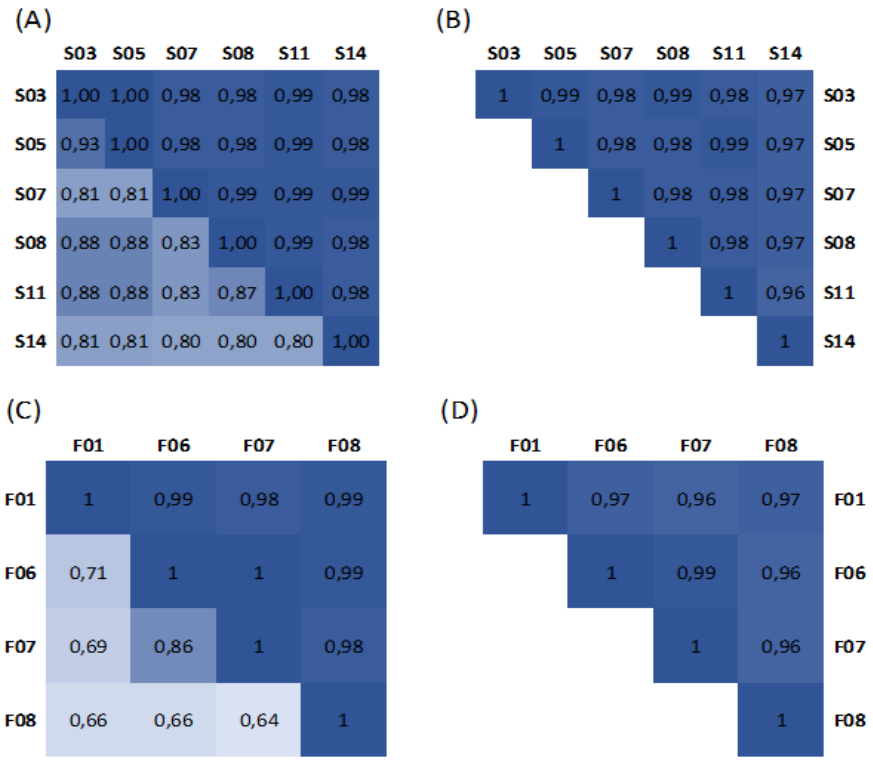

Figure 7. Functional similarity between SIHUMIx samples and fecal samples. The correlation matrices at the left show the Pearson correlation (upper triangle) and Spearman correlation (bottom triangle) for the (A) SIHUMIx data sets and (C) fecal data sets, calculated using the PFAM annotations returned by the protein-centric Prophane analysis. The correlation matrices at the right show the MegaGO similarity for the GO domain "biological process" for the (B) SIHUMIx data sets and (D) fecal data sets, calculated based on the GO terms returned by peptide-centric Unipept analyses.

Discussion

In this founding edition of CAMPI, we used both a simplified, lab-assembled sample as well as a human fecal sample to compare commonly used experimental methods and computational pipelines in metaproteomics at the peptide, protein subgroup, taxonomic and functional level, informed by and contrasted with metagenomics and metatranscriptomics. Our findings demonstrate some differences in the taxonomic profiles between peptide-centric metaproteomics, protein-centric metaproteomics, and read-based metagenomics and metatranscriptomics. This fits well with previous findings that assessment of microbial community structure via shotgun metagenomics and metaproteomics differs in the information obtained. While metagenomics has been shown to provide a good representation of per species cell numbers in a community, metaproteomics has been shown to provide a good representation of per species

biomass in a community[43]. When looking at different proteomics approaches, differences tend to show up primarily at the finest resolution, such as the sequences of the identified peptide sequences. When considering information from the protein subgroup level up, much of this variation disappears. Different protocols tend to primarily display different levels of analytic depth, which correlates with more extensive sample fractionation and faster instruments. Moreover, differences between search engines appear somewhat complementary, giving an advantage to integrative, multi-search engine approaches using more sophisticated scoring engines. Interestingly, there appears to be an important contribution to any observed differences from the sequence database used for identification. This is particularly evident in the protein inference step, where peptide-level degeneracy in the database becomes an important factor in the outcome of protein grouping, as already shown and discussed previously[63,64]. Overall, functional profiles of different proteomics workflows were quite similar, which is a reassuring characteristic due to the unique perspective provided by proteomics on the functional level.

Besides the direct conclusions of CAMPI as summarized here, another important outcome of this study is the availability of the acquired data sets. Indeed, these can serve as benchmark data sets for the field when developing novel algorithms and approaches for data processing and interpretation (see **Data Availability**).

Moreover, this first CAMPI study has highlighted that there is room for future editions of CAMPI studies. Indeed, based on the issues identified in this first study, we can already define interesting future research questions: how is quantification affected by the workflow used, and which quantification approach yields the most robust and accurate results; how are taxonomic resolution, functional profiling, and quantification affected by the dynamic range of the sample composition; and what is the potential of data independent acquisition (DIA) and targeted approaches in metaproteomics regarding reproducibility and analytical depth?

Obviously, relevant standardized samples will need to be defined for these studies, and should moreover be produced in sufficient amounts to allow their continued use by interested researchers after publication of these studies. These could take the form of a defined synthetic community with exactly known composition, including cell numbers and sizes, preferably stimulated under different biological conditions. With such a sample, we

will be able to validate a variety of quantification methods, but also investigate the effect of quantifying individual proteins in relation to their background. Moreover, it remains a question for now what the effect will be on the taxonomic resolution or functional profile. On another technical level, we could investigate the opportunities and challenges of the use of DIA on metaproteomics samples. Potentially, there will be new, AI-driven search engines that will enter the field of (meta)proteomics, which also brings new opportunities for the field.

Of course, all these follow-up CAMPI studies will contribute highly useful benchmark samples and data sets to the field as well, thus creating a strong, positive feedback loop with the metaproteomics community. Future CAMPI editions will be launched by the Metaproteomics Initiative (metaproteomics.org), a newly founded community of metaproteomics researchers which aims, among other things, to investigate, optimize, validate and standardise metaproteomics methods. This initiative can combine forces with existing initiatives such as the ABRF iPRG study group, who recently provided a metaproteomics data set to be analysed by the proteomics informatics community[65]. We believe that such ongoing efforts will continue to advance the field of metaproteomics, and make it more widely applicable. Metaproteomics will thus develop its full potential, and further increase its relevance across the life sciences.

Methods

The CAMPI study aims to evaluate the impact of different protein extraction protocols, MS/MS acquisition strategies, and bioinformatic pipelines used in metaproteomics (see **Figure 1** for a general overview, and **Supplementary Table 1** for an overview of all methods).

1 Sample description

1.1 Simplified human intestinal microbiota sample (SIHUMIx)

A simplified human intestinal microbiota (SIHUMIx) composed of eight species was constructed to embody a majority of known metabolic activities typically found in the human gut microbiome. The SIHUMIx sample contains the Firmicutes *Anaerostipes caccae DSMZ 14662, Clostridium butyricum DSMZ 10702, Erysipelatoclostridium ramosum DSMZ 1402 and Lactobacillus plantarum DSMZ 20174,* the Actinobacteria *Bifidobacterium longum NCC 2705,* the Bacteroidetes *Bacteroides thetaiotaomicron DSM 2079,* the Lachnospiraceae *Blautia producta DSMZ 2950,* and the Proteobacteria *Escherichia coli MG1655*, covering the most dominant phyla in human feces[66]. SIHUMIx was prepared as previously described, with an additional 24h of cultivation of one control bioreactor, to produce sufficient biomass to be sent out to each participating laboratory[66]. Participants received 3,5 x 10^9 cells/ml of frozen sample (-20 °C) in dry ice.

1.2 Human fecal microbiome sample

A natural human fecal microbiome sample was procured from a 33-year old omnivorous, non-smoking woman, with approval by the ethics committee of the University Magdeburg (number 99/10). The sample was immediately homogenized, treated with RNA-later, aliquoted, frozen, and stored at -20°C until aliquots were sent to each participating laboratory.

2 Biomolecule extraction and nucleotide sequencing

2.1 DNA/RNA extraction, sequencing, and processing

DNA was extracted from both SIHUMIx and the fecal samples. RNA could also be extracted from the fecal sample but not SIHUMIx as only the former was treated with RNA-later.

Extracted DNA and RNA were sequenced with Illumina technology, and the obtained sequencing reads subsequently co-assembled into contigs for further bioinformatic processing. Details on the extractions, libraries preparations, and sequencing can be found in **Supplementary Note 1.1**. Preprocessing of the sequenced reads was performed as part of the Integrated Meta-omic Pipeline (IMP)[67] and included the trimming and quality filtering of the reads, the filtering of rRNA from the metatranscriptomic data, and the removal of human reads after mapping against the human genome version 38. Preprocessed RNA and DNA reads were co-assembled using MEGAHIT v1.2.4[68] using minimum and maximum k-mer sizes of 25 and 99, respectively, and a k-step of 4. The resulting contigs were binned using MetaBAT 2.12.1[69] and MaxBin 2.2.6[70] with default parameters and minimum contig length of 2500 and 1500 bps, respectively. Bins were refined using DASTool 1.1.2[71] with default parameters and a score threshold of 0.5. Open reading frames (ORFs) were called from all contigs provided to DASTool using Prodigal 2.6.3[72] as part of the DASTool suite.

2.2 Protein extraction and processing

In total, eight different protein extraction protocols were applied and resulted in 24 different workflows when combined with MS/MS acquisition strategies (**Figure 1**). Key characteristics for each workflow can be found in the **Supplementary Table 1**. The most obvious workflow differences were found in protein recovery, cleaning, and fractionation strategies. In a wide comparative approach, the protein extract was processed by either filter-aided sample preparation (FASP)[73] (workflows 1-3, 5, 7-9,11-12,19-23 in **Supplementary Table 1**), in-gel (workflows 4, 6, 10, 13-18), or in-solution (workflows 21 and 24) digestion. In most workflows, proteins were directly extracted from the raw defrosted material (workflows 1-20, 22-23). In one lab, however, microbial cells were first enriched at the interface of a reverse iodixanol gradient (workflows 21, 24). In most

approaches, cell lysis was based on mechanical cell disruption by bead beating in a variety of chemical buffers (workflows 1-12, 19-23), or in water (workflows 13-18). Apart from bead beating, ultrasonication in a chaotrope-detergent-free buffer was employed to allow for further separation of cytosolic and envelope-enriched microbiome fractions (workflows 21 and 24) and, in another separate workflow, cryogenic grinding was employed for the simultaneous extraction of DNA, RNA, and protein using the Qiagen Allprep kit (workflows 22, 23). Recovery of proteins from the lysis mixture was carried out either by solvent extraction using a variety of solvents, with or without further washes (workflows 4-18, 22, 23), or by filter-aided methods (FASP) (workflows 1-3). All methods included trypsin as the sole proteolytic enzyme for digestion of DTT (or DTE)-reduced and iodoacetamide-alkylated proteins. Digestion was performed either on filters (workflows 1-3, 5, 7-9, 11-12, 19-24), in-gel with or without fractionation (workflows 6, 10, 13-18), or in-solution in the presence of a surfactant (workflows 21 and 24). Of note, the enzyme/substrate ratio varied from 1/50 to 1/10000, with digestion times from 2 to 16 hours. Finally, peptides were recovered from the gel or eluted from filters (FASP) using a salt solution (workflows 1-3, 5-21, 24). In some protocols, peptides were desalted using different commercial devices (workflows 4, 21, and 24).

3 LC-MS/MS acquisition

Each laboratory used its own LC-MS/MS protocol with the largest differences and similarities highlighted in the following and details provided in **Supplementary Table 1**. For LC, all laboratories separated peptides using reversed-phase chromatography with a linear gradient length ranging from 60 min to 460 min. Furthermore, one group performed an additional separation using a multidimensional protein identification technology (MudPIT) combining cation exchange and reversed-phase separation in a single column prepared in-house[74].

Six groups used an Orbitrap mass spectrometer (4x Q Exactive HF, 1x Q Exactive Plus, 1x Fusion Lumos, Thermo Fisher Scientific), while two groups employed a timsTOF mass spectrometer (Bruker Daltonik). All participants used data-dependent acquisition (DDA) with exclusion duration times ranging from 10s to 60s. All MS proteomics data and

X!Tandem results have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[75].

4 Bioinformatics

4.1 Generation of protein sequence databases

Two types of databases were used for each sample; a catalog (reference) database and a database that was generated from metagenomic and metatranscriptomic (when available) data sequenced from a matching sample (meta-omic database). The catalog database for SIHUMIx consisted of the combined reference proteomes of the strains extracted from UniProt in July 2019[76] except for *Blautia producta*, for which the whole genus *Blautia* was taken (SIHUMIx_REF). The IGC 9.9 database[77] (available at http://meta.genomics.cn/meta/dataTools) was used as the catalog database for the fecal sample (GUT_REF). Additionally, a meta-omic database from the assembled contigs was produced for both samples using the open reading frame generated with Prodigal (SIHUMIx_MO and GUT_MO).

The SIHUMIx database (SIHUMIx_REF) is composed of reference proteomes, containing 29,557 proteins (13.2 MB). In comparison, the metagenomic assembly for SIHUMIx (SIHUMIx_MO) produced 2,719 contigs, with an average contig length of 7.5 Kbp and the longest contigs being 468 Kbp, yielding 19,319 predicted ORFs (6.1 MB).

For the fecal sample, the IGC reference catalog (GUT_REF) contains 9,879,896 protein sequences (2.6 GB). The co-assembly of DNA and RNA for the fecal sample (GUT_MO) produced 247,518 contigs with an average length of 1.6 Kbp and the longest contigs being 600 Kbp. The database GUT_MO yielded protein sequences from 441,558 predicted ORFs (114.4 MB). All databases were concatenated with a cRAP database of contaminants (https://thegpm.org/cRAP; downloaded in July 2019) and the GUT databases were additionally concatenated with the human UniProtKB Reference Proteome (downloaded in September 2019).

The four databases were *in silico* digested into tryptic peptides with an in-house developed script, with two missed cleavages allowed, to compare their theoretical search spaces. Additionally, all peptides identified with each database in the explorative analysis, which was carried out using all data sets, were retrieved and compared.

For metaproteomic data analysis, the number of spectra, PSMs, and identification rates (calculated by dividing the number of identified spectra by the total number of acquired MS/MS spectra) were extracted for all data sets searched against the selected databases (SIHUMIx_REF and GUT_MO) and compared. Finally, a representative subset of data sets, based on the different methods, was selected for further analysis (S03, S05, S07, S08, S11, S14 for SIHUMIx and F01, F06, F07, and F08 for the fecal sample).

4.2 Data analysis using four different bioinformatic pipelines

All submitted MS/MS raw files were first analyzed with a single commonly used database search method to assess both the quality of the extraction and the MS/MS acquisition, as well as the effect of the search database composition (reference proteomes vs. multi-omics). For this, X!Tandem[49] was used as search engine with the following parameters: specific trypsin digest with a maximum of two missed cleavages; mass tolerances of 10.0 ppm for MS1 and 0.02 Da for MS2; fixed modification: Carbamidomethylation of C (+57.021464 Da); variable modification: Oxidation of M (+15.994915 Da); fixed modification during refinement procedure: Carbamidomethylation of C (+57.021464 Da). Peptides were filtered on length (between 6 and 50 amino acids), and charge state (+2, +3, and +4), and a maximum valid expectation value (e-value) of 0.1[78].

The following database search engines were used for the pipeline comparison: (i) MaxQuant[79] (including the search engine Andromeda) (ii) Galaxy-P workflows[80,81] consisting of SearchGUI[82,83] (using OMSSA[84], X!Tandem[49], MS-GF+[57], and Comet[85]) and PeptideShaker[86] to merge the results, (iii) MetaProteomeAnalyzer[26] (server version 3.4, using X!Tandem and OMSSA), and (iv) ProteomeDiscoverer 2.2 (using SequestHT, from Thermo Fisher). The identification settings for all search engines were the same as for the explorative analysis mentioned above. Refinement searches were allowed if implemented in the search engine (e.g., refinement search of X!Tandem), and the same for the inclusion of post-processing tools (e.g., Percolator within ProteomeDiscoverer).

4.3 Protein inference

To allow protein group comparison, groups were created using the combined peptide evidence of all compared samples. Two different protein grouping methods were tested:

MPA[26] and PAPPSO[59], and analyses were made on protein groups and subgroups (**Supplementary Note 1.3**).

Assigning peptides to their correct protein can be a difficult task, notably due to the protein inference issue[3], i.e., the same peptide can be found in different homologous proteins. This is particularly challenging in metaproteomics where the diversity and number of homologous proteins are much higher compared to single-species proteomics. To overcome this issue, most bioinformatic pipelines tend to automatically group homologous protein sequences into protein groups. However, each tool handles protein inference and protein groups in its own way, which prevents a straightforward output comparison at the protein group level. In order to allow robust comparison between approaches, the PSM output files of the four bioinformatic pipelines were combined. The peptides were then assigned to protein sequences in the FASTA file and the data was prepared for subsequent protein grouping. Two approaches of protein grouping were used and evaluated in this study: PAPPSO grouping[59], which excludes proteins based on the rule of maximum parsimony, and grouping from MPA[26], which does not exclude proteins. All data processing was done using a custom Java program except for PAPPSO grouping for which data was exported and imported using the appropriate XML format.

For both methods, protein groups were created using the loose rule “share at least one peptide” (groups) and the strict rule “share a common set of peptides” (subgroups), resulting in a total of four protein grouping analyses: (1) PAPPSO groups, (2) MPA groups, (3) PAPPSO subgroups, and (4) MPA subgroups. Finally, the resulting protein groups and subgroups were exported for further analysis (**Supplementary Note 1.3**).

4.4 Taxonomic and functional annotation

Annotations were performed at both the peptide, protein and the sequencing read level. Unipept was used for the peptide-centric approach[22,25,87]. For the taxonomic annotation of the SIHUMIx data sets, we used an advanced Unipept analysis that calculates the SIHUMIx-specific lowest common ancestor (LCA) (*i.e.* it calculates the LCA specific for its search database instead of the complete UniProtKB). Here, Unipept searched for the occurrence of each peptide in all species present in NCBI. For each peptide separately, we removed those species that cannot be present in the SIHUMIx sample (i.e., non-

SIHUMIx species and contaminating species in the cRAP database), after which we calculated the SIHUMIx-specific LCA. This advanced taxonomic analysis using Unipept is possible since the composition of the sample is known, and resulted in a more accurate taxonomic annotation of the peptides. For more information and examples of the advanced Unipept analysis **(Supplementary Note 1.4)**. For the taxonomic annotation of the fecal data sets with Unipept, the desktop[87] and CLI[21,88] versions were used. In both analyses for SIHUMIx and the fecal data sets, isoleucine (I) and leucine (L) were equated. For the functional analysis at the peptide level, we used the Unipept command line option to extract the GO terms for each identified peptide per data set (below 1% FDR). The functional similarity of these sets of GO terms was calculated with MegaGO[62].

Prophane was used for the protein-centric approach[89,90]. For both the functional and taxonomic annotations, a generic output format created by the in-house developed protein grouping script and the protein database for a given analysis were used. Within Prophane, the taxonomic annotation was performed with DIAMOND blastp against the latest NCBI non-redundant (nr) database (2019-09-30)[91], while two functional annotation tasks where performed against the eggNOG (database version 4.5.1)[92] and Pfam-A (db version 32) databases[93] using eggNOG-mapper[94,95] and hmmscan[96], respectively. Using eggNOG-mapper, the e-value threshold was set to 0.0005 while we applied a gathering threshold supported by PFAMs (cut_ga parameter) when searching using hmmscan. The result with the protein group identifiers from the previous analysis summary can be found in **Supplementary Files 3-5**.

Metagenomic and metatranscriptomic reads were both taxonomically annotated with the mOTUs profiler v 2.0[97] with default parameters at the species and family levels for SIHUMIx and the feces sample, respectively.

Quantification was based on read counts for metagenomic and metatranscriptomics data, and on spectral counts for peptides and protein subgroups. If two subgroups contained the same peptide, spectra would be counted twice, distorting the abundance of these particular subgroups inside a measurement, but preserving a consistent count for comparison with other samples. Comparisons were performed with normalised values as described in detail below.

4.5 Statistical analyses

Differences and overlap between search engines at the peptide level and between approaches at the peptide level using presence/absence data were visualized with UpSet plots with the UpSetR package[98]. For the peptides, sequences were extracted (without modifications and with leucine (L) and isoleucine (I) treated equally and replaced by J) from each result file and a table, indicating whether a peptide was found or not, was prepared (**Supplementary Note 1.4** and **Supplementary Files 6 and 7**). Similar tables and UpSet plots were generated to visualize differences and overlap between sample preparations for the peptides, the protein subgroups and the top 50% protein subgroups. The top 50% were first selected based on abundance data. The spectral counts were summed for each subgroup across all selected samples and only the top 50% was kept for UpSet plot comparison. Results from the taxonomic annotations for all approaches (peptides, proteins, metagenomic and metatranscriptomic reads) were compared and visualized using the PCA comparison feature of the R prcomp package. For the comparison, abundance values (number of reads and spectral counts) were used and normalized into percentage. The taxonomic annotations were harmonized across methods, unclassified values were filtered out and annotations with abundance lower than 0.05% after filtering were grouped into “other”.

All correlation plots were calculated using both Pearson and Spearman correlations with a p-value < 0.001. The correlations were calculated and plotted using the corrplot R packages.

Hierarchical clusterings were calculated with the R function hclust using the Manhattan distance and the Ward method.

Data availability

The metaproteomic data sets generated and analyzed in the current study are available via the PRIDE partner repository with the data set identifier PXD023217 (Username: reviewer_pxd023217@ebi.ac.uk Password: DulLrlme).

Assemblies and raw metagenomic and metatranscriptomic reads are available through the European Nucleotide Archive under the study accession number PRJEB42466.

Code availability

All scripts are made available on github.com/metaproteomics/CAMPI.

Acknowledgments

Part of the LC-MS/MS measurements were made in the Molecular Education, Technology, and Research Innovation Center (METRIC) at North Carolina State University, the ProGénoMIX platform at CEA-Marcoule supported by the IBISA network. Parts of the bioinformatics analysis was carried out using the high-performance computing facilities of the University of Luxembourg (https://hpc.uni.lu).

Funding

This work was supported by the Research Foundation - Flanders (FWO) [grant no. 1S90918N (SB) to TVDB; 12I5217N to BM; G042518N to LM]; by a FEBS Summer Fellowship [to TVDB]; by the European Union's Horizon 2020 Program (H2020-INFRAIA-2018-1) [823839 to LM]; by the FEMS [RTG to SSS]; by the Norwegian Centennial Chair program [to TJG, PDJ and MA]; the Novo Nordisk Foundation grant NNF20OC0061313 to MA; the USDA National Institute of Food and Agriculture Hatch project [1014212 to MK]; the U.S. National Science Foundation [OIA 1934844 and IOS 2003107 to MK]; the Foundation for Food and Agriculture Research [Grant ID: 593607 to MK]; the Agence Nationale de la Recherche [ANR-17-CE18-0023-01 to GM, OP, JA]; Deutsche Forschungsgemeinschaft (DFG) [RE3474/5-1 and RE3474/2-2 to SF, TM, BYR]. Research by T.J.G.,P.D.J, E.L were funded by National Cancer Institute-Informatics Technology for Cancer Research (NCI-ITCR) grant 1U24CA199347 and National Science Foundation (U.S.) grant 1458524 to T.J.G. The European Galaxy server that was used for some calculations is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and the German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR, 031L0103 MetaProtServ (de.NBI)). This work was supported by the Luxembourg National Research Fund (FNR) under grants PRIDE/11823097 and CORE-INTER/13684739 to BJK, PM and PW, and the European Research Council (ERC-CoG 863664) to PW.

References

1. Jansson, J. K. & Baker, E. S. A multi-omic future for microbiome studies. *Nat Microbiol* **1**, 16049 (2016).
2. Kleiner, M. Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. *mSystems* **4**, (2019).
3. Hettich, R. L., Pan, C., Chourey, K. & Giannone, R. J. 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 (2013).
4. Rodriguez-Valera, F. Environmental genomics, the big picture? *FEMS Microbiology Letters* **231**, 153–158 (2004).
5. Wilmes, P. & Bond, P. L. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environmental Microbiology* **6**, 911–920 (2004).
6. Michalak, L. *et al.* Microbiota-directed fibre activates both targeted and secondary metabolic shifts in the distal gut. *Nat. Commun.* **11**, 5773 (2020).
7. Kolmeder, C. A. *et al.* Colonic metaproteomic signatures of active bacteria and the host in obesity. *PROTEOMICS* **15**, 3544–3552 (2015).
8. Schiebenhoefer, H. & Van Den Bossche, T. *et al.* Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Review of Proteomics* **16**, 375–390 (2019).
9. Wang, D.-Z., Kong, L.-F., Li, Y.-Y. & Xie, Z.-X. Environmental Microbial Community

Proteomics: Status, Challenges and Perspectives. *IJMS* **17**, 1275 (2016).

10. Taylor, E. B. & Williams, M. A. Microbial Protein in Soil: Influence of Extraction Method and C Amendment on Extraction and Recovery. *Microb. Ecol.* **59**, 390–399 (2010).
11. Field, L. M., Fagerberg, W. R., Gatto, K. K. & Anne Böttger, S. A comparison of protein extraction methods optimizing high protein yields from marine algae and cyanobacteria. *J. Appl. Phycol.* **29**, 1271–1278 (2017).
12. Vaudel, M., Sickmann, A. & Martens, L. Peptide and protein quantification: A map of the minefield. *Proteomics* **10**, 650–670 (2010).
13. Heyer, R. *et al.* Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **261**, 24–36 (2017).
14. Tanca, A. *et al.* The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* **4**, 227 (2016).
15. Timmins-Schiffman, E. *et al.* Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J.* **11**, 309–314 (2017).
16. Muth, T. *et al.* Navigating through metaproteomics data: A logbook of database searching. *Proteomics* **15**, 3439–3453 (2015).
17. Sticker, A., Martens, L. & Clement, L. Mass spectrometrists should search for all peptides, but assess only the ones they care about. *Nat. Methods* **14**, 643–644 (2017).
18. Colaert, N., Degroeve, S., Helsens, K. & Martens, L. Analysis of the Resolution Limitations of Peptide Identification Algorithms. *J. Proteome Res.* **10**, 5555–5561 (2011).

19. Nesvizhskii, A. I. & Aebersold, R. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
20. Heyer, R., Kohrs, F., Reichl, U. & Benndorf, D. Metaproteomics of complex microbial communities in biogas plants. *Microbial Biotechnology* **8**, 749–763 (2015).
21. Verschaffelt, P. *et al.* Unipept CLI 2.0: adding support for visualisations and functional annotations. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa553.
22. Gurdeep Singh, R. *et al.* Unipept 4.0: Functional Analysis of Metaproteome Data. (2019) doi:10.1021/acs.jproteome.8b00716.
23. Park, S. K. R. *et al.* ComPIL 2.0: An Updated Comprehensive Metaproteomics Database. (2019) doi:10.1021/acs.jproteome.8b00722.
24. Sajulga, R. *et al.* Survey of metaproteomics software tools for functional microbiome analysis. *PLoS One* **15**, e0241503 (2020).
25. Van Den Bossche, T. *et al.* Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for seamless end-to-end metaproteomics data analysis. *J. Proteome Res.* (2020) doi:10.1021/acs.jproteome.0c00136.
26. Muth, T. *et al.* The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation. *Journal of Proteome Research* **14**, 1557–1565 (2015).
27. Heyer, R. *et al.* A Robust and Universal Metaproteomics Workflow for Research Studies and Routine Diagnostics Within 24 h Using Phenol Extraction, FASP Digest, and the MetaProteomeAnalyzer. (2019) doi:10.3389/fmicb.2019.01883.
28. Liao, B. *et al.* iMetaLab 1.0: a web platform for metaproteomics data analysis. *Bioinformatics* **34**, 3954–3956 (2018).

29. Zhang, X. *et al.* Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nature Communications* **9**, (2018).
30. Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology* **2**, (2017).
31. Erickson, A. R. *et al.* Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. *PLoS ONE* **7**, e49138 (2012).
32. Juste, C. *et al.* Bacterial protein signals are associated with Crohn's disease. *Gut* **63**, 1566–1577 (2014).
33. Starke, R., Jehmlich, N. & Bastida, F. Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics. *Journal of Proteomics* **198**, 50–58 (2019).
34. Schneider, T. *et al.* Proteome analysis of fungal and bacterial involvement in leaf litter decomposition. *PROTEOMICS* **10**, 1819–1830 (2010).
35. Teeling, H. *et al.* Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science* **336**, 608–611 (2012).
36. Morris, R. M. *et al.* Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *The ISME Journal* **4**, 673–685 (2010).
37. Petersen, J. M. *et al.* Chemosynthetic symbionts of marine invertebrate animals are capable of nitrogen fixation. *Nat Microbiol* **2**, 725 (2017).
38. Kleiner, M. *et al.* Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc.*

Natl. Acad. Sci. U. S. A. **109**, E1173–82 (2012).

39. Delogu, F. *et al.* Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. *Nat. Commun.* **11**, 4708 (2020).
40. Heyer, R. *et al.* Metaproteome analysis reveals that syntrophy, competition, and phage-host interaction shape microbial communities in biogas plants. *Microbiome* **7**, 69 (2019).
41. Rudney, J. D. *et al.* Protein relative abundance patterns associated with sucrose-induced dysbiosis are conserved across taxonomically diverse oral microcosm biofilm models of dental caries. *Microbiome* **3**, 69 (2015).
42. Tanca, A. *et al.* Evaluating the Impact of Different Sequence Databases on Metaproteome Analysis: Insights from a Lab-Assembled Microbial Mixture. (2013) doi:10.1371/journal.pone.0082981.
43. Kleiner, M. *et al.* Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **8**, 6 (2017).
44. Hinzke, T., Kouris, A., Hughes, R.-A., Strous, M. & Kleiner, M. More Is Not Always Better: Evaluation of 1D and 2D-LC-MS/MS Methods for Metaproteomics. (2019) doi:10.3389/fmicb.2019.00238.
45. Mangul, S. *et al.* Systematic benchmarking of omics computational tools. *Nat. Commun.* **10**, 157 (2019).
46. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 291 (2017).

47. Bell, A. W. *et al.* A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **6**, 423–430 (2009).

48. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).

49. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).

50. Cox, J. & Mann, M. Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. *Annu. Rev. Biochem.* **80**, 273–299 (2011).

51. Meier, F. *et al.* Online Parallel Accumulation–Serial Fragmentation (PASEF) with a Novel Trapped Ion Mobility Mass Spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).

52. Wenzel, L. *et al.* SDS-PAGE fractionation to increase metaproteomic insight into the taxonomic and functional composition of microbial communities for biogas plant samples. *Eng. Life Sci.* **18**, 498–509 (2018).

53. Rechenberger, J. *et al.* Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae. *Proteomes* **7**, 2 (2019).

54. Verheggen, K. *et al.* Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. *Mass Spectrom. Rev.* **39**, 292–306 (2020).

55. Park, G. W. *et al.* Integrated Proteomic Pipeline Using Multiple Search Engines for a Proteogenomic Study with a Controlled Protein False Discovery Rate. (2016) doi:10.1021/acs.jproteome.6b00376.

56. Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L. & Deutsch, E. W. Combining Results of Multiple Search Engines in Proteomics. *Mol. Cell. Proteomics* **12**, 2383–2393 (2013).

57. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).

58. Bouwmeester, R., Gabriels, R., Van Den Bossche, T., Martens, L. & Degroeve, S. The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics* e1900351 (2020).

59. Langella, O. *et al.* X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *J. Proteome Res.* **16**, 494–503 (2017).

60. Martens, L. & Hermjakob, H. Proteomics data validation: why all must provide data. *Mol. Biosyst.* **3**, 518–522 (2007).

61. Pible, O. *et al.* Estimating relative biomasses of organisms in microbiota using 'phylopeptidomics'. *Microbiome* **8**, 30 (2020).

62. Verschaffelt, P. & Van Den Bossche, T. *et al.* MegaGO: a fast yet powerful approach to assess functional similarity across meta-omics data sets. *Cold Spring Harbor Laboratory* 2020.11.16.384834 (2020) doi:10.1101/2020.11.16.384834.

63. Serang, O. & Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Stat. Interface* **5**, 3–20 (2012).

64. Huang, T., Wang, J., Yu, W. & He, Z. Protein inference: a review. *Brief. Bioinform.* **13**, 586–614 (2012).

65. Davis, D. L., Palmblad, M. & Weintraub, S. T. iPRG 2019 Metaproteomics Study. *J.*

Biomol. Tech. **30**, S53 (2019).

66. Schäpe, S. S. *et al.* The Simplified Human Intestinal Microbiota (SIHUMIx) Shows High Structural and Functional Resistance against Changing Transit Times in In Vitro Bioreactors. *Microorganisms* **7**, 641 (2019).
67. Narayanasamy, S. *et al.* IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biology* **17**, (2016).
68. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
69. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. (2019) doi:10.7717/peerj.7359.
70. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
71. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* **3**, 836–843 (2018).
72. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, (2010).
73. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
74. Wolters, D. A., Washburn, M. P. & Yates, J. R. An Automated Multidimensional

Protein Identification Technology for Shotgun Proteomics. *Analytical Chemistry* **73**, 5683–5690 (2001).

75. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
76. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
77. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
78. Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong. *J. Am. Soc. Mass Spectrom.* **22**, 1111–1120 (2011).
79. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367–1372 (2008).
80. Jagtap, P. D. *et al.* Metaproteomic analysis using the Galaxy framework. *PROTEOMICS* **15**, 3553–3565 (2015).
81. Blank, C. *et al.* Disseminating Metaproteomic Informatics Capabilities and Knowledge Using the Galaxy-P Framework. *Proteomes* **6**, (2018).
82. Vaudel, M., Barsnes, H., Berven, F. S., Sickmann, A. & Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *PROTEOMICS* **11**, 996–999 (2011).
83. Barsnes, H. & Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for

Proteomics Search and de Novo Engines. *Journal of Proteome Research* **17**, 2552–2555 (2018).

84. Geer, L. Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
85. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
86. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **33**, 22–24 (2015).
87. Verschaffelt, P., Van Den Bossche, T., Martens, L., Dawyndt, P. & Mesuere, B. Unipept Desktop: A Faster, More Powerful Metaproteomics Results Analysis Tool. *J. Proteome Res.* (2021) doi:10.1021/acs.jproteome.0c00855.
88. Mesuere, B. *et al.* The Unipept metaproteomics analysis pipeline. *PROTEOMICS* **15**, 1437–1442 (2015).
89. Schiebenhoefer, H. *et al.* A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane. *Nat. Protoc.* **362**, 776 (2020).
90. Schneider, T. *et al.* Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. *The ISME Journal* **6**, 1749–1762 (2012).
91. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* **35**, D61–D65 (2007).
92. Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous

groups of genes. *Nucleic Acids Research* **36**, D250–D254 (2007).

93. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
94. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
95. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
96. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Research* **46**, W200–W204 (2018).
97. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
98. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).