Freie Universität Berlin

Habilitationsschrift

Setting up the bioinformatics environment for employing meta-omics techniques in microbiome research and pathogen diagnostics

zur Erlangung der Lehrbefähigung
für das Fach Bioinformatik

vorgelegt dem Fachbereich Mathematik und Informatik
Freie Universität Berlin

von

Dr. Thilo Muth

Eingereicht: Juni 2021

Introduction

Motivation

In recent years, the general interest in microbiome research has strongly increased. This is mainly due to key findings suggesting a strong impact of microbial communities on the health of their human hosts. For example, the human gut harbors a massive number of mostly beneficial bacteria for which researchers have only gained prelimary insights about their key functions and high diversity so far. Yet, the cellular and molecular interplay of microorganisms with their host has already become apparent and many current research efforts focus on investigating associations between microbial dysbiosis and certain disease states. Along with a beneficial role of microbes, many bacteria and viruses can have harmful effects on their hosts and this pathogenic potential is of utmost interest. For public health, a robust and fast detection of pathogenic bacteria or viruses is critical: reducing the time for diagnosis and treatment of a severe infection can drastically lower the potential risk of further transmission and mortality for the affected patients. The detection of newly emerging and reoccurring pathogens requires diagnostic assays to cover a broad spectrum for various potentially disease-causing agents. However, handling global challenges such as the current SARS-CoV-2 pandemics does not only require a fast detection, but also powerful methods to investigate biomolecules of pathogens for an in-depth understanding at the functional level. For example, the SARS-CoV-2 spike protein had been identified as a key target for eliciting neutralizing antibodies and it is essential to investigate its role for the immune response.

In the past, high-throughput genome sequencing technologies have revealed both the immense variety of pathogens and the complexity of microbial communities. Reaching beyond the genetic potential that is addressed by metagenomics, the technique of metaproteomics as the mass spectrometry-based analysis of microbial communities enables investigating the metabolic and cellular pathways in which microbial enzymes are key players. In addition, proteomics itself has reached a mature stage where it cannot only be used in a research context, but also for diagnostics as it had been demonstrated with newly emerging SARS-CoV-2 applications. Recently, more powerful bioanalytical instrumentation and advances in machine learning provide exciting prospects for MS-based metaproteomics. However, the increasing quality and quantity of data from large-scale proteomics experiments demand the development of new robust algorithms and special-purpose software. Next to shotgun approaches, upcoming targeted proteomics approaches and data-independent acquisition can overcome intrinsic problems concerning sensitivity and specificity but require adapted developments of computational methods to unleash their full potential. Finally, a holistic view on biological processes through omics-based analyses can only be achieved by workflows that not only regard multiple omics levels separately but link and integrate multi-omics data to create an added value for the experimentalists. Overall, the technological advances in instrumentation and experimental setups have lead to new opportunities of analysis depth and speed, but the risk is high that data analysis and interpretation lag behind the fast progress in data acquisition. It therefore remains a general challenge to handle the large amounts of data from high-throughput experiments and to achieve a fast and reliable outcome.

During my studies, I could identify the following specific challenges:

- Algorithmic protocols and software for processing and analyzing metaproteomics data require robustness, user-friendliness, interoperability, and sustainability to be transferable into clinical and industrial applications.
- Research and diagnostic applications lack special-purpose methods such as reliable bioinformatic methods for detecting and characterizing viral pathogens.
- Workflows for integrating host and microbiome data from multiple omics levels (metagenomics, metatranscriptomics, and metaproteomics) are needed for a meaningful biological interpretation.
- Individual developments bare the potential risk of unintended, approach-based biases: comprehensive surveys and independent benchmarking studies on bioinformatics methods are therefore essential to evaluate the outcome of respective workflows.

Objectives and scope of the work

With the above mentioned challenges in mind, this habilitation thesis aims at revealing the various potentials offered by meta-omics technlogy and corresponding bioinformatics methods to process, analyze and interpret data derived from experiments on microbial and microbiome samples. During my work at the Robert Koch Institute I put a particular emphasis on developing computational methods for detecting and characterizing bacterial and viral pathogens as well as for integrating host and microbiome data at three different meta-omics levels. In addition, I contributed to both biology-focused experimental and data-oriented bioinformatics benchmarking studies. Finally, the focus of my work also broadened towards integrating methods and data for multiple omics levels, including metagenomics, metatranscriptomics, and metaproteomics.

The following sections give a synopsis of the major findings from my conducted research after finishing my doctoral dissertation in 2016 [1]. According to §2.I.1.(c) of the Habilitationsordnung the written habilitation thesis should be preceded by a detailed summary of published research results. For this habilitation work, ten scientific articles were chosen that I had contributed to with further co-authors since finishing my PhD. This habilitation thesis was written as a result of three first-author, three middle-author, and four last-author journal articles [2–11]. All original manuscripts have been deposited in the attachment of this work.

As stated in the title, the focus of my habilitation work was to establish a bioinformatics environment by developing, evaluating and improving processing and analysis methods for data from meta-omics technology with direct applications in microbiome research and pathogen diagnostics. In close collaboration with renowned national and international scientific co-workers, I conducted and steered projects for developing bioinformatic algorithms and software for computational analysis of pathogen and microbiome samples at the taxonomic and functional level. To adress the existing data-oriented challenges, representative scientific and diagnostic applications using bacterial and viral data sources have been selected in this work. The lack of special-purpose bioinformatics methods has been tackled by mainly concentrating on the development of robust and user-friendly software for (i) metaproteomics [2, 6, 8, 9], (ii) virus diagnostics [5, 7], and (iii) multi-omics integration of host and microbiome data [10]. A further pillar of my research within this habilitation framework represent survey and benchmarking investigations on evaluating bioinformatic methods for database-independent *de novo* sequencing in proteomics [3, 4] as well as an international multi-lab benchmarking study that aimed to evaluate existing algorithms for identification, quantification, and protein grouping in metaproteomics [11].

In the following sections, the motivation and main results of each article are highlighted together with a detailed statement describing my contribution with respect to the conceptual design and project implementation. In addition, I had further contributed to several peer-reviewed publications with middle and last authorships. These publications are not described in detail in this work, but are listed and briefly summarized in a separate paragraph at the end of this summary.

Finally, a step towards accomplishing this habilitation presents my continuous teaching of bioinformatics and meta-omics lectures and courses each semester in the Department of Mathematics and Computer Science at the FU Berlin since the summer term 2017.

Summary of research

MPA Portable Software for Analyzing Samples in Metaproteomics

Microbiome research has received an increased attention because of the key roles that microbial communities have in the environment and in the human, animal, and plant host. It is well studied that the human gut microbiome takes over crucial host-related functions, for example, it is a critical component of digestion and nutrient update and can also resolve an immune response following an infection [12]. The microbiome is that important because microorganisms are essential to the niche system (e.g., a human host) in which they are found. Common bioanalytical techniques such as 16S rRNA gene sequencing or whole metagenome sequencing are very useful to study the taxonomic composition and the functional potential of microbial communities. However, with their use one cannot assess the actual functional profile of microbes under specific conditions or at a given point in time. To gain insights from functionally active snapshots of microbial communities, metaproteomics, the mass spectrometry-based analysis of proteins from multi-species samples was established [13]. Various studies using metaproteomics have been conducted to characterize microbiomes at the protein level with respect to taxonomic composition and functional enzymes in relevant metabolic pathways [14, 15]. At the computational side, however, the field still faces specific challenges of data analysis and interpretation including (i) the more pronounced protein inference issue, (ii) the risk of selection bias, and (iii) the inflated search space leading to previously described issues of false discovery rate estimation. In general, existing bioinformatic methods for proteomics had not been adapted to overcome these issues and tailored software solutions for metaproteomics had been rare in general.

The first publication [2] describes the MetaProteomeAnalyzer (MPA) Portable software as full-featured application that overcomes several limitations of the original server-based application [16]. The software provides an end-to-end data processing workflow for metaproteomics: its features include a fully customizable experiment and project setup system with metadata annotations, loading and indexing of MS/MS spectra, protein sequence database search with multiple search algorithms, peptide and protein identification, false discovery rate estimation, protein grouping (using multiple user-defined grouping rules), taxonomic assignment (using NCBI taxonomy) and functional analysis based on KEGG pathways. For running the software, no specific computational expertise is required for installation and the prior dependence on a relational database system has been resolved. Further, the integrated database search algorithms have been updated to high-performance algorithms such as MS-GF+ and Comet. In order to tackle the issues of the decreased identification yield in metaproteomics, two different variants of a two-step search approach were added: a commonly used protein-based and a new taxon-based iterative search. Finally, besides the existing graphical user interface, a command line interface was added to run MPA on cluster environment supporting multi-threading for task parallelization. Along with the presentation of the new features of the software, we evaluated the developed methods using two experimental data sets derived from samples with a known species composition. The first benchmarking data set was created by mixing the five bacterial species Bacillus subtilis, Escherichia coli, Pseudomonas fluorescens, Micrococcus luteus, and Desulvofibrio vugaris. The second data set used for benchmarking was the published lab-assembled mixture of nine microbial organisms (9MM) [17]. The results from the benchmarking experiments indicate that combining multiple search algorithms significantly increases the number of correct unique and correct taxon-specific peptides and, at the same time, the number of incorrect hits could be reduced by using a taxon abundance threshold of 5%. In addition, the two-step searching showed to be beneficial when combined with the taxonomic filtering by increasing the number of correct taxonomic assignments while keeping the number of incorrect assignments at a similarly low level compared to standard database searching. Notably, it needs to be considered that the two-step searching the actual peptide false discovery rate (FDR) is likely higher than the FDR threshold set for the second search iteration.

The concept idea for the MPA Portable software arose from discussions with all authors, particularly, Fabian Kohrs, Robert Heyer, Bernhard Renard, and me. Fabian set up and conducted the web-lab experiments and provided the 5BCT benchmarking data set. I steered the software project and wrote the code for the new features such as the implementation of the two-step search approach as well as the command line interface with multi-threading support. Data evaluation of the benchmarking analysis was carried out by me with valuable suggestions from Bernhard. I wrote the manuscript with edits and suggestions from all authors.

Performance Evaluation of De Novo Sequencing Algorithms

In proteomics, searching tandem mass spectra against sequence databases presents the method of choice for peptide and protein identification. However, this approach runs into problems when sequence information is unavailable or when samples contain unexpected sequence variants, as in the case of samples from non-model organisms or cancer cell types, respectively. The method of *de novo* sequencing aims to overcome this problem by inferring amino acid sequences directly from a tandem mass spectrum without mapping to any reference sequence. In the past, *de novo* sequencing has been often been considered being slow and inaccurate compared to gold-standard database searching and therefore was not applicable in high-throughput proteomics studies. Recently, however, faster and more accurate *de novo* sequencing algorithms have been proposed that claim to benefit specifically from higher-resolution data derived from improved instrumentation in mass spectrometry.

In the second publication [3], we evaluated the performance of common *de novo* sequencing algorithms in an independent benchmarking study. While the outcome of new developments is often compared against competing algorithms, there is a lack of unbiased comparative studies and benchmarking is often performed without any solid ground truth: to our knowledge, *de novo* sequencing results had only been compared against peptide hits obtained from database search algorithms so far. Notably, searching against sequence databases is an error-prone process for which a suitable statistical error control such as FDR estimation needs to be applied. In order to provide a ground truth for assessing the *de novo* sequencing performance at the peptide level, we performed a simulation of tandem mass spectra by performing machine learning-based peak intensity predictions. In this study, we evaluated the *de novo* sequencing algorithms Novor [18], PepNovo+ [19] and PEAKS [20] using experimental high-resolution data sets from different instruments and fragmentation modes. Across all evaluations, the Novor algorithm showed the overall best performance concerning full length peptides and amino acid recall, followed by the commercial competitor PEAKS. On experimental spectra, none of the evaluated algorithms was highly accurate for delivering full-length peptide sequences. It should also be noted that the analysis of the simulated spectrum data resulted in an accuracy of over 80%, while gold standard database search algorithms failed to reach perfect accuracy. The latter results indicate that reference-based searching and *de novo* sequencing are closer in performance than it had been reported in previous studies. We could also demonstrate that using short peptide tags of few amino acids based on the *de novo* sequence predictions resulted in high accuracy values of over 90%. Finally, we also assessed the most common *de novo* sequencing errors and the impact of lacking fragment ion and noise peaks on the performance of the algorithms. The outcome was that peptides with long sequences, for example, caused by multiple missed cleavages, presented a yet unresolvable challenge for the algorithms. Based on the outcome of this benchmarking study, we could summarize that *de novo* sequencing has a high potential when its tools are improved or utilized in downstream analyses.

For this project, I had the initial idea of evaluating existing *de novo* sequencing tools in proteomics. This idea was refined in more depth in various fruitful discussions with Bernhard Renard. Together, we then planned and designed the stucture of the benchmarking study. Bernhard provided additional input concerning data simulations of MS/MS spectra and accuracy assessments of *de novo* sequence predictions. I wrote the benchmarking code, conducted the data processing as well as analysis and wrote the manuscript. Bernhard gave valuable advice and edited the manuscript. Felix Hartkopf critically read the manuscript and provided useful feedback.

Analyzing the current state of de novo sequencing in shotgun proteomics

In bottom-up proteomics, the gold standard is to search tandem mass spectra against in-silico digested protein sequences, commonly provided by reference databases such as UniProtKB or NCBI, to order to identify peptides and to infer proteins. However, this process is dependent on the quality and disposability of suitable reference proteomes. Therefore, when such proteomes are unavailable or incomplete (e.g. for samples from non-model organisms or microbial communities), the database-driven approach becomes limited. In addition, single amino acid variations or post-translationally modified proteins occurring in pathogenic or clinical cancer samples are problematic because reference database often do not cover such variabilities. While tailored protein sequence databases can be used, the increased search space limits the discrimination power of database algorithms for peptide identification. Thus, complementary approaches are required to overcome these specific challenges. Besides spectral library searching, de novo sequencing remains an appealing technique that infers partial or complete peptide sequences directly from tandem mass spectra.

The third publication [4] provides an overview on the current state of bioinformatic methods and software tools for de novo sequencing in proteomics. In this article, we first describe the algorithmic principles: many *de novo* sequencing algorithms use graph theory-based methods by constructing a spectrum graph for each tandem mass spectrum. By doing so, peaks are converted into nodes representing masses (i.e., m/z values) of partial peptides. When two nodes have the same or similar mass difference of one or multiple amino acids, a directed edge is applied. The best-scoring path over the spectrum graph is then used to find candidate peptide sequences from a given mass spectrum. Second, we analyzed the literature starting from the 1980s until today focusing on methods and software tools for *de novo* sequencing, including techniques based on dynamic programming, integer linear programming, divide-and-conquer, hidden Markov models, machine learning, and deep learning. Besides full-sequence (i.e., from N- to C-terminus) methods, we describe sequence tagging algorithms by which so-called sequence tags can be obtained: these are partial peptide sequences that consist of few amino acids surrounded by mass gaps. Third, we describe methods for sequence-to-protein mapping for combining *de novo* sequencing with database-driven methods with the ultimate goal of protein inference. Forth, we focus on use cases of *de novo* sequencing from previous studies such as antibody sequencing, application on non-model organisms and cross-species identification, venom-based studies, and carbohydrate analytics. Finally, we critically discuss shortcomings of existing methods and highlight future potential of improvement of *de novo* sequencing methods and software tools, eventually leading to a wider adoption of *de novo* sequencing in the proteomics community.

The idea of analyzing the current state of *de novo* sequencing evolved in discussions with Bernhard Renard and Marc Vaudel. I conducted the major part of the literature research with assistance from Felix Hartkopf. Felix also helped in designing the figures and evaluating the use cases and practical applications. I wrote the manuscript with edits and suggestions from all authors. I also thank Tobias Loka and Robert Rentzsch for critical reading of the manuscript and their valuable suggestions.

Peptide selection software for virus diagnostics based on targeted proteomics

Viruses are infectious agents that transmit biological information and replicate inside living cells of host organisms. On order to replicate, viruses reprogram the function of host cells from humans, animals, plants, and even bacteria to produce virus particles, also known as progeny virions. Despite their tiny size and commonly simple structure, viral pathogens such as HIV, Influenza, Ebola, Zika, and lately SARS-CoV-2 present serious health threats for millions of individuals worldwide. Human viral pathogens are highly diverse with currently hundreds of different pathogenic species and many more likely to be discovered in the future. For public health, a sensitive, specific, and fast detection of pathogenic viruses is critical: reducing the time for diagnosis and treatment of an infection can strongly decrease the potential risk of further viral transmission and mortality for the affected patients. As an example, for SARS-CoV-2, it is crucial to rapidly detect the virus in large cohorts of clinical samples to control the virus spread by specific containment measures. While detecting SARS-CoV-2 variants allows to identify attenuated or more infectious forms of the virus, specific viral mutations might undergo existing molecular detection methodologies. In this context, mass spectrometry-based targeted proteomics has recently emerged to a promising methodology for characterizing viral proteins in biological samples with high sensitivity, quantitative accuracy, and reproducibility.

In the fourth publication [5], we describe the development of Purple (Picking unique relevant peptides for viral experiments), a software for selecting target-specific peptides used in targeted proteomics assays for virus diagnostics. Developing such assays involves various steps of peptide candidate selection, peptide synthesis, and assay optimization. The peptide selection process requires comparing candidate peptides against a large search space of background proteomes. Equipped with a graphical user interface, Purple enables peptide candidate selection across various taxonomic levels and automated filtering for background protein sequence information with proteins that are not specific for the target virus proteome. In this process, candidate peptide sequences are validated against a user-selected sequence database of virus proteomes. In the publication, we provide practical use cases using sample data from different virus strains and species. Our software facilitates performing the crucial step of taxa-specific peptide selection and therefore can be used for pathogen screening and diagnostics.

The original work in preparation to the presented publication had been conducted by Johanna

Lechner in her master thesis (Institute of Bioinformatics at the Department of Computer Science and Mathematics, FU Berlin) at RKI under my supervision. The concept, methodology and formal analysis of a computational workflow for peptide selection was developed by me in discussions with Johanna, Felix Hartkopf Bernhard Renard and Jörg Döllinger. The resources were provided by Andreas Nitsche, Marica Grossegesse, Bernhard and Jörg. Johanna and Felix prepared the original draft of the manuscript and editing was done by all authors. Felix also helped in designing the figures and further data analysis visualizations.

Improving metaproteomics with automated sample comparison, metagenome annotation and peptide indexing

The most severe challenges in metaproteomics originate from limitations of experimental as well as computational protocols. At the experimental side, the sample preparation is time-consuming and is also affected by sensitivity issues to sample impurities. While a typical sample preparation workflow used in metaproteomics research currently takes up to one week, the overall diagnostics workflow would require a much shorter timeframe for the complete analysis to be employed in routine fashion. Another immanent problem in metaproteomics workflows concerns computational aspects with the lack of well-established data analysis methods and software tools for processing and interpretation of microbiome samples.

In the fifth publication [6], we describe the combination of a full-featured metaproteomics workflow that covers both important experimental as well as computational requirements. The aim of our work is to ensure overall time-efficiency, experimental simplicity, high-throughput performance, reproducibility, and robustness. All these features are strongly required in routine diagnostics protocols. Our newly developed workflow has the advantage of being very time-efficient in the sense that the overall processing from sample preparation to data analysis and visualization can be performed within 24 hours. The workflow features several experimental improvements with respect to protein extract and in-gel digestion. Concerning the bioinformatics part, important features are to compare results at the identification and quantification level from different experiments and to automatically annotate metagenome sequences with taxonomic and functional information. In addition, we developed a peptide index that is used when processing protein databases, overcoming the Occams razor implementation of peptide identification and protein inference algorithms tailored for single proteome use cases. For metaproteomics, however, inferring a reduced number of proteins can be highly disadvantageous because relevant information on potentially occurring species with homologous sequences is consequently lost. The peptide index prevents this problem, by performing a lookup step after the database search step that recovers all protein hits inferred from the identified peptides. This strategy works in combination with a subsequent protein grouping step to accurately represent homologous proteins across multiple species. This approach resulted in an increase of reported proteins by a factor of up to 16, while the number of reported metaproteins remained approximately the same or slightly decreased. Across all tested samples, our new workflow resulted in at least two times as many protein identifications and significantly more assigned taxa as well as annotated protein functions. These features were additionally implemented into the MPA server version 3.0, increasing the usability of the software to analyze and interpret metaproteome-based MS/MS data from microbial samples. It resulted in at least two times as many protein identifications and significantly more assigned taxa as well as annotated protein functions. This performance increase provides the basis for further steps to establish metaproteomics in the routine analysis of technical and environmental samples and

points to directions of applying it into diagnostic settings.
The concept and further development of the bioinformatics workflow was steered by me together with Robert Heyer. The software implementations were performed by Kay Schallert, Roman Zoun, Sebastian Dorl, Robert, and me. The original draft of the manuscript was written by Robert, Kay, Dirk Benndorf, Udo Reichl. I further assisted with editing and revising the manuscript.

Iterative database searching for strain-level identification of pathogenic samples

Acquiring strain-level knowledge is of utmost importance in any diagnostic setting, for example, in infection outbreak scenarios from emerging viruses. However, it is also required in a therapeutic context to infer virulence, i.e., the degree of damage caused by a microbe to its host, and to identify the resistance phenotype of pathogens. While various proteomics applications have emerged in the recent past, obtaining exact strain-level information from pathogenic samples remains challenging because the exact sample origin is often unknown. In this context, untargeted searches relying on large reference proteome databases are often biased towards frequently described organisms, have limited taxonomic depth at the species level, or contain closely related strains with very similar sequences. Moreover, extending databases with strain-level information increases computational runtime, while it reduces statistical power when the false discovery rate is estimated based on the target-decoy approach. One way to overcome these problems is to constrain the search space upfront, but this is prone to selection bias with potentially relevant strains remaining unidentified or being assigned to incorrect taxa in the end.
In the sixth publication [7], we describe a bioinformatics workflow called TaxIt for strain-level identification using tandem mass spectra from samples of unknown taxonomic background. The aim is to address the increasing search space of untargeted strain-level sequence databases using iterative searching. We adapted the general concept of multi-step procedures used in database-driven peptide identification to a comprehensive strain-level search space that is required for untargeted identification and taxonomic assignment. We apply two distinct identification stages for both species- and strain-level classification in iterative fashion: the first step performs an untargeted search aiming to select relevant species from an unknown sample. Based on the identified species candidates, the second step focuses on a limited number of adequate strain-level proteomes being automatically acquired from external database resources. This way, we overcome the issue of immediately requiring an overly large search space of a database containing all available strain-level proteomes at the same time. In addition, we address ambiguous taxonomic assignment originating from similar proteomes with a newly developed abundance weighting algorithm that increases the taxonomic assignment confidence. For benchmarking, we apply our workflow on selected bacterial and viral samples. In contrast to non-iterative strategies, we could show that TaxIt correctly identified all microbial strains in each of the cases. In summary, our workflow provides an accurate strain-level classification with increased identification confidence and reduced taxonomic ambiguity.
For this project, I participated in the workflow design and consulted for the individual software modules. Together with Mathias Kuhring and Bernhard Renard, I helped with drafting the original manuscript and assisted with editing and revising. Mathias performed the implementation and evaluation of the workflow. Bernhard participated in the computational and experimental design. Jörg Döllinger performed the experimental design and data acquisition.

Andreas Nitsche participated in the experimental design.

Combining established tools into a complete metaproteomics analysis workflow

Microbiome research has evolved to a highly relevant field with multiple applications in microbiology, ecology, and medicine. The increased attention of microbiome studies is mainly due to methodological and technological progress in omics domains. Complementary to metagenomics and metatranscriptomics, metaproteomics gives insights into the functioning of microbial communities and their interactions with the host at the protein level. For analyzing data from microbiome samples using metaproteomics methods and software tools have been developed. Two examples are UniPept [21] and MPA [16]: these tools fulfill different, yet complementary purposes with Unipept assigning peptides to taxaonomies and MPA providing a complete peptide identification and protein grouping.

In the seventh publication [8], we describe the implementation of a connection from MPA to Unipept that allows identified peptides to be uploaded directly to Unipept. The identified peptides can be also provided by PeptideShaker as another proteomics software that allows false discovery rate filtering of peptides equally to MPA. To evaluate the developed combined workflow, we reprocessed 45 spectrum files of data from a sample from a simplified model of the human gut. This sample aims to cover the most known metabolic activities typically found in the human gut. It consists of eight bacterial species covering the most dominant genera Firmicutes, Bacteroidetes, and Proteobacteria in the human intestine. For data processing and visualization, over 67,000 uniquely identified peptides were transferred via an export routine automatically to Unipept. In Unipept, the taxonomies can be visualized using different techniques either via a treeview, sunburst plot, treemap, or a hierarchical outline. In addition, it provides a functional analysis feature by assigning peptides to Enzyme Commission or Gene Ontology annotations. In conclusion, we provide the link of established metaproteome-based identification workflows with a user-friendly visualization module for downstream analysis of taxonomies and protein functions.

I supervised Tim van den Bossche during his research stay at the Robert Koch Institute and during this time I steered the project while developing the idea of integrating multiple metaproteomics tools together with Tim. Bart Mesuere and I helped Tim with the code implementation for the integration of the existing tools MPA and Unipept. Bernhard Renard, Lennart Martens, and I conceived and designed the experiments that were then performed by Tim. Tim wrote the manuscript together with edits and suggestions from me and the co-authors.

Data analysis protocol covering an end-to-end metaproteomics workflow

Mass spectrometry-based microbiome research faces various high technological requirements such as high throughput capability, large dynamic range, high sensitivity, and mass accuracy, but it also depends on sophisticated data analysis methods. On the one hand, these methods need to provide reliable peptide identification and protein inference similarly as for conventional single-organism proteomics tools. On the other hand, metaproteome-focused sample analysis requires much attention on adding meaningful protein annotations at the data analysis and interpretation stage. Previously, with MPA [16] and Prophane [22], two software tools were tailored towards metaproteomics data analysis. MPA supports peptide-spectrum matching with multiple database search algorithms and optimized protein grouping strategies. Prophane is a web-based application focusing on taxonomic and functional annotation using different annotation resources and results visualization. These tools have been applied in various scientific domains such as microbial community research, molecular ecology, and the study of host-pathogen interactions.

In the eighth publication [9], we describe a tailored workflow that integrates MPA and Prophane as two well-established software solutions developed for metaproteomics. The proposed protocol provides the researcher with guidelines for the step-by-step data processing and data analysis instructions for mass spectrometry-based microbiome samples. Besides describing peptide identification, protein inference and taxonomic assignment, it focuses on functional analysis aspects of integrating external databases such as UniProt, EggNOG, PFAM and CAZy. For evaluating the whole workflow, we provide an exemplary analysis on two different samples. In this evaluation, we assess the effect of different sample preparation methods and the influence of protein databases. Furthermore, we evaluate the taxonomic and functional composition of these samples using Sunburst plots and Sankey diagrams. Our protocol enables the user to perform the complete data analysis process in metaproteomics, including protein database creation, database search, protein grouping, taxonomic as well as functional annotation, and finally, specific features for results visualization and interpretation. While novice users are provided with a robust and user-friendly data analysis in a few hours, more advanced users benefit from the adaptability and flexibility and adaptability of the workflow.

The development of the protocol was supervised by Stephan Fuchs, Bernhard Renard, and me. Henning Schiebenhfer and Kay Schallert performed the implementation of the protocol and conducted the data analyses together with the help of Stephan, Bernhard, and me. Bernhard, Stephan designed the experiments that were then performed by Henning and Kay. Stephan and Henning wrote the manuscript together with edits and suggestions from the other co-authors and me.

Multi-omics pipeline for integrated host and microbiome analysis

Microbiome research has gained much attention lately driven by findings on impact of microbial communities on the health in humans, animals, and plants. Common analysis techniques are metagenomics, metatranscriptomics, and metaproteomics: taken separately, these methods can already powerfully complement and support each other. Also, many bioinformatics solutions have been developed for analyzing microbiomes based on the data of each of the methods alone. So far, however, there is a lack of bioinformatic software to process and analyze data sets from these individual techniques in an integrated manner. Currently, automated multi-omics bioinformatic workflows are almost non-existent and do cover two different omics levels at most. Further, existing workflows have been designed for human microbiome analyses and commonly neglect non-model organisms. The latter is crucial because non-model organisms are underrepresented in public repositories and therefore reference sequence data is rare, leading to problems with respect to data processing and analysis across different omics levels. Finally, a full integration across multiple omics levels is required to fully decipher the ongoing interactions between host and microbiota.

The ninth publication [10] describes the development and evaluation of gNOMO, a meta-omics pipeline that integrates the analysis levels metagenomics, metatranscriptomics, and metaproteomics. While these meta-omics techniques are already powerful individually, their combination allows investigating the interplay of microbial species with their host at both taxonomic and functional level. Besides the integration of meta-omics tools, we developed a specific workflow for generating tailored protein sequence databases directly from genomics and transcriptomics data allowing the analysis of host data without any proteome reference. The gNOMO pipeline makes use of the workflow management system Snakemake to ensure both automation capability and reproducibility. We demonstrate the use of the pipeline using experimental datasets from samples of *Blattella germanica*, the German cockroach. This non-model organism resides ordinarily in human habitats and harbors a complex gut microbiome and further symbionts. Focusing from a biological perspective, we show the capabilities of gNOMO with its complete meta-omics data integration, different sample abundance analysis, taxonomic and functional annotation as well as visualization features for results interpretation. While gNOMO can be customized for processing and analyzing multiple meta-omics data types for producing output visualizations, it was designed for using paired-end sequencing with high-resolution mass spectrometry data from non-model organisms. gNOMO closes the gap of lacking multi-omics pipelines for microbiome and host samples integrating and comparting data at the genome, transcriptome, and proteome level.

For this project, based on the initial ideas of Maria Muñoz-Benavent I conceptualized the work of developing a multi-omics pipeline at three different omics levels. Together with Maria, I conceived and designed the experiments that were then performed by her under my guidance. Maria and Felix Hartkopf implemented the workflow in Snakemake with assistance of Vitor Piro and Tim van den Bossche. Maria and I wrote and revised the manuscript together with edits and suggestions from Carlos García-Ferris, Amparo Latorre, and Bernhard Renard.

Multi-lab study of experimental and bioinformatic workflows in metaproteomics

It has been demonstrated in many studies that microbial communities are major drivers in biogeochemical cycles and have a strong impact on natural environments, industrial processes as well as the health and nutrition of both humans and important livestock animals. Common techniques for analyzing microbial communities are metagenomics, metatranscriptomics, and metaproteomics. While high-throughput genome sequencing technologies have revealed the immense variety of microbes and the complexity of microbiomes, metaproteomic approaches go beyond the mere genetic potential by investigating the functional protein level and thereby the enzymatic and metabolic pathways of both microbiome and host. While the field of metaproteomics has experienced a significant growth over the last decade, the use of different proposed workflows comes with the natural risk of unintended, method-specific biases. However, to bring metaproteome studies into real-world clinical or industrial applications, it important establish quality-controlled and reproducible methods both at the experimental and computational level. As the next step, it is therefore required to evaluate the most commonly used workflows in a larger benchmarking study involving different laboratories from the field of metaproteomics.

In the tenth publication [11], we describe the setup and outcome of the first international multi-laboratory study in metaproteomics, referred to as the Critical Assessment of MetaProteome Investigation (CAMPI). The study was conducted as a ring trial community-based effort for which each participating laboratory had received two different metaproteome samples: one sample with known composition from a simplified mock community simulating the gut microbiome and another from with unknown species origin in the form of a complex, natural stool sample. In the CAMPI study, we compare experimental and computational workflows including all analysis steps ranging from sample preparation to the bioinformatic analysis of peptide identification, protein and species inference and quantification.

We observed that meta-omics databases performed better than public reference databases across both samples. More importantly, we found that the functional profiles obtained from the diverse analyses were highly similar across workflows, while minor differences were observed for the inferred community composition. We also found that these differences originated primarily from the wet-lab protocols rather than from the bioinformatic pipelines. Our work demonstrates the robustness of metaproteomics serving as a template for further studies in the field by providing benchmarking data sets for developments to further increase the quality of metaproteomics analysis.

The original call for experimental laboratories to participate in a benchmarking effort was sent out by Dirk Benndorf and Nico Jehmlich as main organizers of the 3rd International Metaproteomics Symposium. The seven participating labs designed their experiments individually. At the symposium, the decision was made to reanalyze the acquired data with different bioinformatics pipelines. After that I steered the study, conceptualizing the design and refined further bioinformatics analysis steps. The data analyses were performed the first authors Tim van den Bossche, Kay Schallert, Benoit Kunath and Stephanie Schäpe under my supervision. Tim, Kay, Benoit, Stephanie and I wrote the manuscript together with edits and suggestions from all co-authors, in particular from Jean Armengaud, Catherine Juste, Manuel Kleiner, Lennart Martens, and Bernhard Renard.

Further published work

While the ten selected articles described above represent the main directions of my research in the last five years after completing my PhD, I co-authored several further publications that are briefly described in the following paragraph.

In one publication [23], we evaluated the impact of sequence database choice on the identification rate in gut microbiome studies. While this work appeared in 2016 after the submission of my doctoral thesis, I contributed mainly to the manuscript during my PhD studies. I also contributed as a first author during my PhD to a book chapter [24] that describes user-friendly tools for peptide identification using tandem mass spectrum sequencing. While working in the Bioinformatics Unit at RKI, I contributed as a first author for a review article [25] on the techniques, challenges and advances in compuational microbial community proteomics. Another publication [26] that I co-authored in 2018 describes modular bioinformatics workflows integrated in the Galaxy for proteomics platform (Galaxy-P). In this context, I participated in a contribution workshop with a group of other software developers and expert users from the metaproteomics research community. The publication describes the software tools that were selected, packaged and deployed via the Galaxy-P platform. In 2019, I co-authored another publication [27] reviewing the current state of bioinformatics at the interface of genomics and metaproteomics by surveying methods for metaproteogenomic data analysis. In the same year, I contributed to a book chapter [28] on the peptide-to-protein summarization step used for accurate protein quantification in label-based proteomics. In 2020, I co-authored a study [29] that evaluated several software tools in metaproteomics for analyzing microbiomes at the functional level by measuring their combined proteome-level response to environmental perturbations. In this survey, we evaluated the performance of tools that take the combined proteome-level response of microbiome to environmental perturbation into account. The aim was to enable scientists making informed decisions regarding software choice based on their research goals. Furthermore, I contributed as last author to a perspective article [30] in the context of the recent SARS-CoV-2 pandemics. In this article, we discussed the potential of proteomics for detecting viral infections. In this publication, we highlighted the challenges and the future potential of applying proteomics in routine virus diagnostics. In the time following this publication, several research articles have been published in which MS-based proteomics techniques were applied to SARS-CoV-2 samples.

References

[1] T. Muth, *Novel Computational Methods for the Analysis and Interpretation of MS/MS Data in Metaproteomics*. PhD thesis, Otto-von-Guericke-Universität Magdeburg, 2016.

[2] T. Muth, F. Kohrs, R. Heyer, D. Benndorf, E. Rapp, U. Reichl, L. Martens, and B. Y. Renard, "MPA Portable: A stand-alone software package for analyzing metaproteome samples on the go.," *Analytical chemistry*, vol. 90, pp. 685–689, Jan. 2018.

[3] T. Muth and B. Y. Renard, "Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification?," *Briefings in bioinformatics*, vol. 19, pp. 954–970, Sept. 2018.

[4] T. Muth, F. Hartkopf, M. Vaudel, and B. Y. Renard, "A potential golden age to come-current tools, recent use cases, and future avenues for de novo sequencing in proteomics.," *Proteomics*, vol. 18, p. e1700150, Sept. 2018.

[5] J. Lechner, F. Hartkopf, P. Hiort, A. Nitsche, M. Grossegesse, J. Doellinger, B. Y. Renard, and T. Muth, "Purple: A computational workflow for strategic selection of peptides for viral diagnostics using MS-based targeted proteomics.," *Viruses*, vol. 11, June 2019.

[6] R. Heyer, K. Schallert, A. Bdel, R. Zoun, S. Dorl, A. Behne, F. Kohrs, S. Pttker, C. Siewert, T. Muth, G. Saake, U. Reichl, and D. Benndorf, "A robust and universal metaproteomics workflow for research studies and routine diagnostics within 24 h using phenol extraction, FASP digest, and the MetaProteomeAnalyzer.," *Frontiers in microbiology*, vol. 10, p. 1883, 2019.

[7] M. Kuhring, J. Doellinger, A. Nitsche, T. Muth, and B. Y. Renard, "Taxit: An iterative computational pipeline for untargeted strain-level identification using ms/ms spectra from pathogenic single-organism samples.," *Journal of proteome research*, vol. 19, pp. 2501–2510, June 2020.

[8] T. Van Den Bossche, P. Verschaffelt, K. Schallert, H. Barsnes, P. Dawyndt, D. Benndorf, B. Y. Renard, B. Mesuere, L. Martens, and T. Muth, "Connecting metaproteomeanalyzer and peptideshaker to unipept for seamless end-to-end metaproteomics data analysis.," *Journal of proteome research*, vol. 19, pp. 3562–3566, Aug. 2020.

[9] H. Schiebenhoefer, K. Schallert, B. Y. Renard, K. Trappe, E. Schmid, D. Benndorf, K. Riedel, T. Muth, and S. Fuchs, "A complete and flexible workflow for metaproteomics data analysis based on metaproteomeanalyzer and prophane.," *Nature protocols*, vol. 15, pp. 3212–3239, Oct. 2020.

[10] M. Muñoz-Benavent, F. Hartkopf, T. Van Den Bossche, V. C. Piro, C. García-Ferris, A. Latorre, B. Y. Renard, and T. Muth, "gnomo: a multi-omics pipeline for integrated host and microbiome analysis of non-model organisms," *NAR Genomics and Bioinformatics*, vol. 2, no. 3, p. lqaa058, 2020.

[11] T. Van Den Bossche, B. J. Kunath, K. Schallert, S. S. Schpe, P. E. Abraham, J. Armengaud, M. . Arntzen, A. Bassignani, D. Benndorf, S. Fuchs, R. J. Giannone, T. J. Griffin, L. H. Hagen, R. Halder, C. Henry, R. L. Hettich, R. Heyer, P. Jagtap, N. Jehmlich, M. Jensen, C. Juste, M. Kleiner, O. Langella, T. Lehmann, E. Leith, P. May, B. Mesuere, G. Miotello, S. L. Peters, O. Pible, P. T. Queiros, U. Reichl, B. Y. Renard, H. Schiebenhoefer, A. Sczyrba, A. Tanca, K. Trappe, J.-P. Trezzi, S. Uzzau, P. Verschaffelt, M. von Bergen, P. Wilmes, M. Wolf, L. Martens, and T. Muth, "Critical assessment of metaproteome investigation (campi): a multi-laboratory comparison of established workflows.," *Nature communications*, vol. 12, p. 7305, Dec. 2021.

[12] J. L. Round and S. K. Mazmanian, "The gut microbiota shapes intestinal immune responses during health and disease.," *Nature reviews. Immunology*, vol. 9, pp. 313–323, May 2009.

[13] P. Wilmes and P. L. Bond, "Metaproteomics: studying functional gene expression in microbial ecosystems.," *Trends in microbiology*, vol. 14, pp. 92–97, Feb. 2006.

[14] C. A. Kolmeder, J. Ritari, F. J. Verdam, T. Muth, S. Keskitalo, M. Varjosalo, S. Fuentes, J. W. Greve, W. A. Buurman, U. Reichl, E. Rapp, L. Martens, A. Palva, A. Salonen, S. S. Rensen, and W. M. de Vos, "Colonic metaproteomic signatures of active bacteria and the host in obesity.," *Proteomics*, vol. 15, pp. 3544–3552, Oct. 2015.

[15] A. Tanca, M. Abbondio, A. Palomba, C. Fraumene, V. Manghina, F. Cucca, E. Fiorillo, and S. Uzzau, "Potential and active functions in the gut microbiota of a healthy human cohort.," *Microbiome*, vol. 5, p. 79, July 2017.

[16] T. Muth, A. Behne, R. Heyer, F. Kohrs, D. Benndorf, M. Hoffmann, M. Lehtev, U. Reichl, L. Martens, and E. Rapp, "The metaproteomeanalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation.," *Journal of proteome research*, vol. 14, pp. 1557–1565, Mar. 2015.

[17] A. Tanca, A. Palomba, M. Deligios, T. Cubeddu, C. Fraumene, G. Biosa, D. Pagnozzi, M. F. Addis, and S. Uzzau, "Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture.," *PloS one*, vol. 8, p. e82981, 2013.

[18] B. Ma, "Novor: real-time peptide de novo sequencing software.," *Journal of the American Society for Mass Spectrometry*, vol. 26, pp. 1885–1894, Nov. 2015.

[19] A. Frank and P. Pevzner, "Pepnovo: de novo peptide sequencing via probabilistic network modeling.," *Analytical chemistry*, vol. 77, pp. 964–973, Feb. 2005.

[20] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, "Peaks: powerful software for peptide de novo sequencing by tandem mass spectrometry.," *Rapid communications in mass spectrometry : RCM*, vol. 17, pp. 2337–2342, 2003.

[21] B. Mesuere, B. Devreese, G. Debyser, M. Aerts, P. Vandamme, and P. Dawyndt, "Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples.," *Journal of proteome research*, vol. 11, pp. 5773–5780, Dec. 2012.

[22] T. Schneider, E. Schmid, J. V. de Castro, M. Cardinale, L. Eberl, M. Grube, G. Berg, and K. Riedel, "Structure and function of the symbiosis partners of the lung lichen (lobaria pulmonaria l. hoffm.) analyzed by metaproteomics.," *Proteomics*, vol. 11, pp. 2752–2756, July 2011.

[23] A. Tanca, A. Palomba, C. Fraumene, D. Pagnozzi, V. Manghina, M. Deligios, T. Muth, E. Rapp, L. Martens, M. F. Addis, and S. Uzzau, "The impact of sequence database choice on metaproteomic results in gut microbiota studies.," *Microbiome*, vol. 4, p. 51, Sept. 2016.

[24] T. Muth, E. Rapp, F. S. Berven, H. Barsnes, and M. Vaudel, "Tandem mass spectrum sequencing: An alternative to database search engines in shotgun proteomics.," *Advances in experimental medicine and biology*, vol. 919, pp. 217–226, 2016.

[25] T. Muth, B. Y. Renard, and L. Martens, "Metaproteomic data analysis at a glance: advances in computational microbial community proteomics.," *Expert review of proteomics*, vol. 13, pp. 757–769, Aug. 2016.

[26] C. Blank, C. Easterly, B. Gruening, J. Johnson, C. A. Kolmeder, P. Kumar, D. May, S. Mehta, B. Mesuere, Z. Brown, J. E. Elias, W. J. Hervey, T. McGowan, T. Muth, B. Nunn, J. Rudney, A. Tanca, T. J. Griffin, and P. D. Jagtap, "Disseminating metaproteomic informatics capabilities and knowledge using the galaxy-p framework.," *Proteomes*, vol. 6, Jan. 2018.

[27] H. Schiebenhoefer, T. Van Den Bossche, S. Fuchs, B. Y. Renard, T. Muth, and L. Martens, "Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis.," *Expert review of proteomics*, vol. 16, pp. 375–390, May 2019.

[28] M. Fischer, T. Muth, and B. Y. Renard, "Peptide-to-protein summarization: An important step for accurate quantification in label-based proteomics.," *Methods in molecular biology (Clifton, N.J.)*, vol. 1977, pp. 159–180, 2019.

[29] R. Sajulga, C. Easterly, M. Riffle, B. Mesuere, T. Muth, S. Mehta, P. Kumar, J. Johnson, B. A. Gruening, H. Schiebenhoefer, C. A. Kolmeder, S. Fuchs, B. L. Nunn, J. Rudney, T. J. Griffin, and P. D. Jagtap, "Survey of metaproteomics software tools for functional microbiome analysis.," *PloS one*, vol. 15, p. e0241503, 2020.

[30] M. Grossegesse, F. Hartkopf, A. Nitsche, L. Schaade, J. Doellinger, and T. Muth, "Perspective on proteomics for virus detection in clinical samples.," *Journal of proteome research*, vol. 19, pp. 4380–4388, Nov. 2020.

MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome Samples on the Go

doi: 10.1021/acs.analchem.7b03544

Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification?

doi: 10.1093/bib/bbx033

A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics

doi: 10.1002/pmic.201700150

viruses

MDPI

Article

Purple: A Computational Workflow for Strategic Selection of Peptides for Viral Diagnostics Using MS-Based Targeted Proteomics

Johanna Lechner [1,†], Felix Hartkopf [1,†], Pauline Hiort [1], Andreas Nitsche [2], Marica Grossegesse [2], Joerg Doellinger [3], Bernhard Y. Renard [1,*] and Thilo Muth [1]

[1] Bioinformatics Unit (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany; LechnerJ@rki.de (J.L.); HartkopfF@rki.de (F.H.); p.hiort@web.de (P.H.); MuthT@rki.de (T.M.)

[2] Centre for Biological Threats and Special Pathogens, Highly Pathogenic Viruses (ZBS1), Robert Koch Institute, 13353 Berlin, Germany; NitscheA@rki.de (A.N.); GrossegesseM@rki.de (M.G.)

[3] Centre for Biological Threats and Special Pathogens, Proteomics and Spectroscopy (ZBS 6), Robert Koch Institute, 13353 Berlin, Germany; DoellingerJ@rki.de (J.D.)

[*] Correspondence: RenardB@rki.de; Tel.: +49-(0)30-18754-2561

[†] These authors contributed equally to this work.

Received: 19 March 2019; Accepted: 4 June 2019; Published: 8 June 2019

Abstract: Emerging virus diseases present a global threat to public health. To detect viral pathogens in time-critical scenarios, accurate and fast diagnostic assays are required. Such assays can now be established using mass spectrometry-based targeted proteomics, by which viral proteins can be rapidly detected from complex samples down to the strain-level with high sensitivity and reproducibility. Developing such targeted assays involves tedious steps of peptide candidate selection, peptide synthesis, and assay optimization. Peptide selection requires extensive preprocessing by comparing candidate peptides against a large search space of background proteins. Here we present Purple (Picking unique relevant peptides for viral experiments), a software tool for selecting target-specific peptide candidates directly from given proteome sequence data. It comes with an intuitive graphical user interface, various parameter options and a threshold-based filtering strategy for homologous sequences. Purple enables peptide candidate selection across various taxonomic levels and filtering against backgrounds of varying complexity. Its functionality is demonstrated using data from different virus species and strains. Our software enables to build taxon-specific targeted assays and paves the way to time-efficient and robust viral diagnostics using targeted proteomics.

Keywords: virus proteomics; mass spectrometry; virus diagnostics; data analysis; targeted proteomics; peptide selection; parallel reaction monitoring

1. Introduction

Virus infections present serious health threats to millions of individuals worldwide. For public health, the accurate detection of pathogenic viruses is time-critical because reducing the time for diagnosis and treatment lowers the risk of disease transmission and patient mortality. Fast and robust diagnostic assays are therefore required to rapidly detect re-emerging and newly emerging viruses (e.g., Influenza, Ebola, Zika, or Hepatitis C virus). These diagnostic methods need to cover a broad spectrum of potentially disease-causing viral agents.

Classical diagnostic strategies for detecting viral infection can be divided into two different categories: on the one hand, virus detection can be established by targeted methods, such as agent-specific polymerase chain reaction (PCR) or immunological techniques. On the other hand, detection approaches exist that provide an open view, such as electron microscopy or next-generation sequencing (NGS). Besides their unbiased view, the latter methods have the advantage of identifying multiple pathogens in a single experimental run. Due to its specificity (hybridization and sequencing) and sensitivity (qPCR), the detection of nucleic acids is the gold standard in diagnostics. Conversely, the detection of viral proteins is used less frequently in diagnostic settings and is usually based on interaction with affine binding molecules such as antibodies or aptamers. However, producing these binding molecules is generally time-consuming and laborious, as is the validation of their specificity.

While in clinical microbiology the analysis of subproteomes (<12 kDa) using matrix assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry (MS) has become a standard method for the identification of bacteria and fungi, no comparable proteomic approach exists in virology for technical reasons [1]. In recent years, MS-based targeted proteomics has evolved into a technique for detecting proteins in complex samples with high sensitivity, quantitative accuracy, and reproducibility [2,3]. Targeted proteomics is commonly used to test hypotheses on a subset of proteins of interest, in contrast to discovery shotgun proteomics. The latter provides global proteome profiling of thousands of proteins in a sample, however, at the expense of sensitivity and reproducibility. Unlike discovery methods, targeted methods of selected/multiple reaction monitoring (SRM/MRM) [4] and parallel reaction monitoring (PRM) [5] nowadays allow for detecting and analyzing preselected proteins and peptides in sensitive, specific, and time-efficient manner. Furthermore, the development of targeted proteomics assays has become easier in the past few years, owing to the advances of analytical methods, instrumental capabilities, and computational workflows [6].

Targeted MS-based proteomics assay development typically involves (i) peptide candidate selection, (ii) peptide synthesis, and (iii) assay optimization. This procedure now enables the transfer of a process highly similar to the design of multiplex PCRs to the proteome level for detecting pathogens. While MS-based targeted assays have not been used for detecting viruses in any diagnostic setting yet, promising findings could already be achieved for identifying and quantifying pathogenic bacterial species. For example, targeted proteomics methods were successfully used in previous studies on *Streptococcus pyogenes* [7] and *Mycobacterium tuberculosis* [8].

Although targeted proteomics has gained much popularity with many use cases in experimental research by now, relatively few research-oriented algorithms and software tools have been developed that support the user-defined selection of peptides for designing targeted SRM or PRM assays. In this context, Skyline [9] is a powerful and widely used software for designing targeted proteomics assays. Besides its wide applicability to different targeted methods and its intuitive use, it also has some internal limitations: first, Skyline is dependent on the operating system Windows, and can therefore not be used under a Linux cluster server environment, and second, it does perform only exact string matching during the peptide selection process without considering any homologies between related organisms. PeptidePicker [10] is a web-based workflow to select peptides by providing, amongst further options, the protein accession number and was designed for human and mouse proteomes. PeptideManager [11] is a tool developed to select peptide candidates as protein surrogates from a defined proteome. It was optimized for the use case of xenografts, i.e., human tumors orthotopically implanted into a different species. While this software allows for constructing a peptide database from any species-specific proteome, sequence homologies, and multiple taxonomic levels are disregarded. Picky [12]—a web-based method designer for targeted assays—only provides support for human and mouse sequences, while it relies on synthetic peptide data from the human-focused ProteomeTools project [13,14]. In the context of targeted metaproteomics, the Unique Peptide Finder of the UniPept web application [15] was developed to select unique peptides for user-defined taxa. Furthermore, various computational tools have been developed to predict proteotypic peptides for targeted proteomics experiments [16–18]. These methods often make use of machine learning training setups that incorporate the probability of observing a peptide in a standard proteomics analysis, referred to as peptide detectability [19] or observability [20], and commonly involve physicochemical properties of the proteins to select high-responding peptides [21]. To our best knowledge, however, no software tool is currently available to select taxon-specific peptides for targeted proteomics assays that also accounts for sequence homologies between different species or strain proteomes. Effectively considering homologies is crucial for accurate taxon-specific diagnostics, because proteins measured in virus samples frequently have a high sequence similarity either in closely related strains or due to highly conserved functional domains.

Here we present Purple (Picking unique relevant peptides for viral experiments), a platform-independent software that returns a set of taxon-specific peptides, after the user has specified the viral target (i.e., a particular virus species or genus), as candidates for targeted proteomics experiments. Equipped with a user-friendly graphical user interface and a threshold-based filtering strategy for homologous sequences, it simplifies the design of MS-based targeted proteomics assays for the end user. Purple enables peptide candidate selection and considers background sequence information, i.e., proteins that are not related to a specific virus target, at various taxonomic levels. Thus, all peptide candidates are validated against a user-defined database of virus proteomes. While the design of MS-based targeted assays requires further steps, our software greatly facilitates the cumbersome, yet important task of peptide selection and thereby paves the way to time-efficient and robust pathogen screening and viral diagnostics. Purple is open source software available at https://gitlab.com/rki_bioinformatics/Purple.

2. Materials and Methods

2.1. Purple Workflow

Purple is implemented in Python (version 3.6) and makes use of additional Python libraries such as tqdm (https://github.com/tqdm/tqdm) for process bar calculation and Biopython [22] to calculate the molecular weight of peptides. Purple is available as portable standalone version that already includes all required libraries or Purple can be installed using pip or conda, which are managing dependencies. The workflow of Purple is depicted in an overview diagram (Figure 1). Purple requires the input of protein sequence databases and a configuration file. The databases are automatically rearranged into a target and a background database. The “exact matching” step is used to remove exact sequence matches with the background from the target peptide set. The remaining target peptides are used to detect and remove homologous peptides. A result file containing the final unique peptides is created together with various intermediate result files. These are outputs of all Purple processing tasks, namely (i) digested peptides, (ii) exact matching peptides, (iii) non-homologous matching peptides and (iv) background shared peptides.

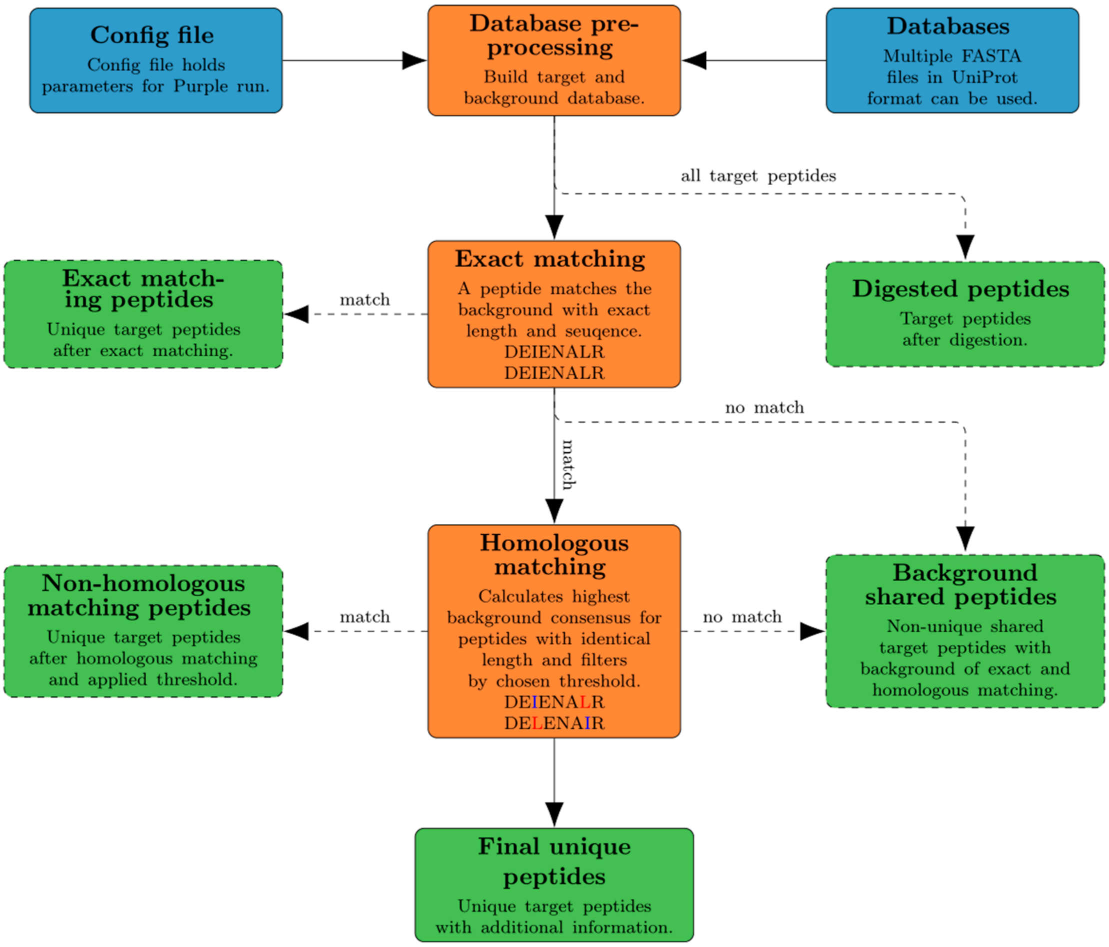

Config file
Config file holds parameters for Purple run.
Database pre-processing
Build target and background database.
Databases
Multiple FASTA files in UniProt format can be used.
all target peptides
Exact matching
A peptide matches the background with exact length and seqence.
DEIENALR
DEIENALR
match
Exact matching peptides
Unique target peptides after exact matching.
Digested peptides
Target peptides after digestion.
no match
match
Homologous matching
Calculates highest background consensus for peptides with identical length and filters by chosen threshold.
DEIENALR
DELENAIR
match
Non-homologous matching peptides
Unique target peptides after homologous matching and applied threshold.
no match
Background shared peptides
Non-unique shared target peptides with background of exact and homologous matching.
Final unique peptides
Unique target peptides with additional information.

Figure 1. Overview of the Purple workflow. A configuration file and a directory path to the location of FASTA databases serve as input (blue). In the database preprocessing step, the databases are separated into target and background (orange). Any target peptides exactly matching to the background database are removed. In the homologous matching step, any target peptides that have similar sequences are filtered out (orange). All intermediate and final results are exported automatically to a user-defined output folder (green).

2.1.1. Preprocessing (Target Selection)

The selection of a target virus proteome is handled by input and preprocessing routines in Purple. For target selection, protein sequence databases in FASTA format serve as main input and are required to be provided in UniProt format. To select the database input, a directory needs to be specified by the user and multiple FASTA files can be considered for the processing. Two options of database specification are available in Purple: the first option is to explicitly define target species names as a list, which leads to the merging of all provided input databases. Each protein entry that contains one of the defined target species names in the protein header is considered as a target protein. The protein sequences not matching the defined target species are used as background database. The second option is to select a specific FASTA file in the database directory as target database. All remaining databases in the directory are then automatically merged to a single background database. As the background database may still consist of proteins originating from one of the target species, each protein in the background database is checked once more: if a protein header matches any species in the specified target database file, the protein entry is removed from further processing accordingly.

Both options result in two types of databases, namely a target and a background database. In the following, each protein sequence in these databases is *in silico*-digested using the enzymatic rule of trypsin with optional proline digestion. The in silico digest step results in multiple peptides for each protein entry, and peptide sequences beyond the user-defined length boundaries are filtered out. In addition, preprocessing includes the option of removing protein fragments and also allows replacing each isoleucine by leucine: this option was implemented because these amino acids share identical molecular masses and are therefore commonly not distinguishable in mass spectrometry. When the preprocessing is completed, both a target and a background database are provided for further analysis, which in this stage consist of peptides instead of proteins.

2.1.2. Exact Matching

Exact matching presents the first actual processing step in Purple: here, each of the previously *in silico*-digested target peptides is compared against the provided background database (see previous paragraph). In this procedure, target and background peptides of identical length are compared and only those target peptides that are not contained in the background are considered further; thus, peptide sequences with one or more exact sequence matches in the background database are filtered out at this stage, because they are not unique to the user-defined taxa of the target space. This procedure is performed iteratively until all *in silico*-digested peptides have been evaluated. The remaining peptides that have not been filtered out are stored as unique peptide candidates for further processing and are exported as intermediate result of the exact matching step.

2.1.3. Homologous Matching

Homologous matching is performed subsequently to the exact matching step. The goal is to evaluate each of the unique peptide candidates concerning its potential sequence consensus to homologous proteins in the background. The rationale behind this approach is that the more similar a target peptide is to the background, the less appropriate it is as candidate for a taxon-specific targeted assay. To assess the similarity of each peptide to the background proteomes, a sequence background consensus metric is introduced (see next paragraph). The target peptides that are discarded either during the exact or the homologous matching step are exported as so-called "shared" peptides. Shared peptides have either an exact sequence match with the background or have background consensus value above a user-defined threshold. To keep track of all processed data, target peptides with a background consensus below the threshold are exported as well.

2.1.4. Background Consensus Metric and Threshold Generation

Owing to mutational effects on conserved viral proteins, peptides can often be shared within a virus genus or family with minor sequence variations between them. This is problematic for targeted assays because such peptide candidates are not specific for species- or strain-level identification. To remove such taxon-unspecific peptides from the final sequence set, the background consensus metric $f(A, B)$ is used in Purple as the essential part of the homologous matching. Basically, the background consensus presents the Hamming distance of a target peptide A and background peptide B of the same length in relation to the length of the peptide n (Equation (2)). An amino acid is shared if the same amino acid ($d(x, y)$) is at the same position in A and B (Equation (1)).

$$d(x,y) = \begin{cases} 1, & if\, x = y \\ 0, & if\, x \neq y \end{cases} \tag{1}$$

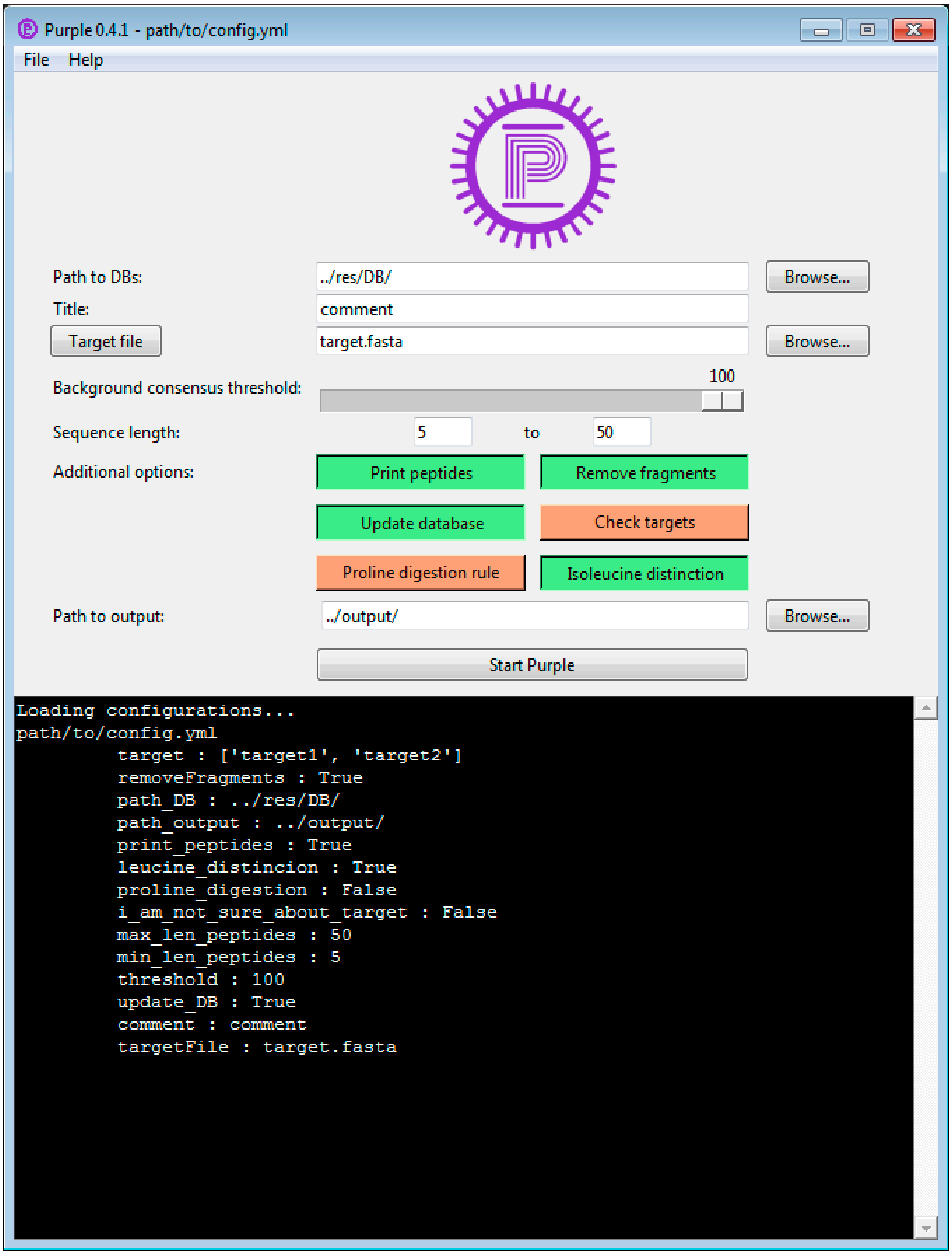
In other words, the background consensus is the sum of shared amino acids at a specific position i divided by the number of amino acids in both (target and background) peptides. Even though the Hamming distance is a simple metric, it provides a proof-of-concept and validation of Purple, as adding more sophisticated methods should only slightly improve the homologous matching while increasing the computational effort and complexity.

$$\begin{gathered} For\ A = \{a_1, a_2, \ldots, a_n\}\ and\ B = \ A = \{b_1, b_2, \ldots, b_n\}\ and\ n = |A| = |B| : \\ f(A,B) = \frac{\sum_{i=1}^{n} d(a_i, b_i)}{n}, for\ a_i \in A\ and\ b_i \in B \end{gathered} \tag{2}$$

This metric is applied to each of the target peptides that are compared to all background peptides of the same length. For each target peptide, the maximum consensus is stored when being below a user-defined background consensus threshold. A target peptide with a high background consensus is likely to originate from a homologous protein or common protein domain. Therefore, the consensus metric evaluates the conservation of peptides in the target and background database. A low background consensus marks target peptides that are unique in sequence in the target species. All peptides with a high background consensus below the previously chosen threshold are exported into the final results file and the remaining shared peptides are exported as part of the intermediate output. The results are supplemented with the peptide weight, the number of occurrences in the target database, as well as species and proteins names. This enables the user to conduct further analysis with the previously retrieved unique peptides. The Purple documentation is available for a complete description of all output files and more details about the data interpretation.

2.2. Graphical User Interface

A graphical user interface (GUI) was developed for using Purple (Figure 2). This interface allows researchers with less expertise in handling bioinformatics methods on the command line to use Purple in a efficient and user-friendly manner. The Purple GUI makes software configuration and execution straightforward and complex tasks can be rapidly accomplished. Any parameter can be adjusted in the GUI, and the background consensus threshold can be set by the user. Furthermore, the processing status can be inspected in a logging panel and a file menu provides options for saving and loading configuration files. Note that configuration files are optional in Purple and a default configuration is provided; thus, only system-specific parameters must be set in the GUI. Using configuration files makes each task reproducible and the GUI-integrated configuration file choice allows for switching between multiple settings easily. Figure 3 shows the final output in the tab separated values (TSV) format that can be further processed and visualized using common spreadsheet software.



Path to DBs: ../res/DB/

Title: comment

Target file: target.fasta

Background consensus threshold: 100

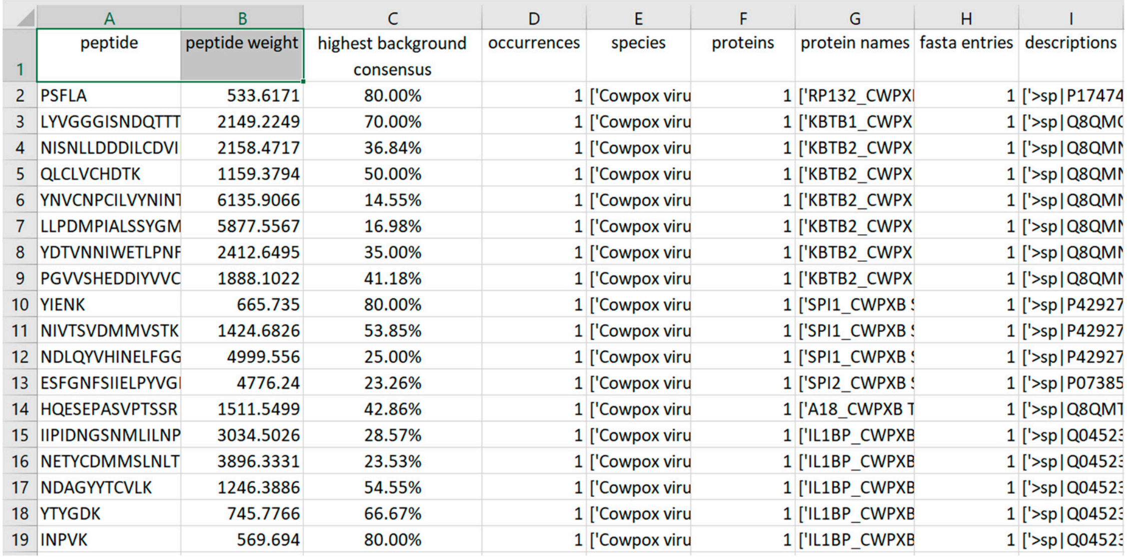
Sequence length: 5 to 50

Additional options:

Path to output: ../output/

```
Loading configurations...
path/to/config.yml
         target : ['target1', 'target2']
         removeFragments : True
         path_DB : ../res/DB/
         path_output : ../output/
         print_peptides : True
         leucine_distincion : True
         proline_digestion : False
         i_am_not_sure_about_target : False
         max_len_peptides : 50
         min_len_peptides : 5
         threshold : 100
         update_DB : True
         comment : comment
         targetFile : target.fasta
```

Figure 2. The graphical user interface of Purple. In the top file menu, configurations files can be loaded and saved. The top menu also includes a link to the documentation and manual. The listed GitLab page provides direct user support from the developers via an issue tracking system. The upper panel shows default parameters and allows modifying the configuration settings and processing start. The lower panel displays the current processing status with logging information on the current run, configuration, and progress of the analysis.



	A	B	C	D	E	F	G	H	I
1	peptide	peptide weight	highest background consensus	occurrences	species	proteins	protein names	fasta entries	descriptions
2	PSFLA	533.6171	80.00%	1	['Cowpox viru	1	['RP132_CWPXI	1	['>sp|P17474
3	LYVGGGISNDQTTT	2149.2249	70.00%	1	['Cowpox viru	1	['KBTB1_CWPX	1	['>sp|Q8QMC
4	NISNLLDDDILCDVI	2158.4717	36.84%	1	['Cowpox viru	1	['KBTB2_CWPX	1	['>sp|Q8QMN
5	QLCLVCHDTK	1159.3794	50.00%	1	['Cowpox viru	1	['KBTB2_CWPX	1	['>sp|Q8QMN
6	YNVCNPCILVYNINT	6135.9066	14.55%	1	['Cowpox viru	1	['KBTB2_CWPX	1	['>sp|Q8QMN
7	LLPDMPIALSSYGM	5877.5567	16.98%	1	['Cowpox viru	1	['KBTB2_CWPX	1	['>sp|Q8QMN
8	YDTVNNIWETLPNF	2412.6495	35.00%	1	['Cowpox viru	1	['KBTB2_CWPX	1	['>sp|Q8QMN
9	PGVVSHEDDIYVVC	1888.1022	41.18%	1	['Cowpox viru	1	['KBTB2_CWPX	1	['>sp|Q8QMN
10	YIENK	665.735	80.00%	1	['Cowpox viru	1	['SPI1_CWPXB S	1	['>sp|P42927
11	NIVTSVDMMVSTK	1424.6826	53.85%	1	['Cowpox viru	1	['SPI1_CWPXB S	1	['>sp|P42927
12	NDLQYVHINELFGG	4999.556	25.00%	1	['Cowpox viru	1	['SPI1_CWPXB S	1	['>sp|P42927
13	ESFGNFSIIELPYVGI	4776.24	23.26%	1	['Cowpox viru	1	['SPI2_CWPXB S	1	['>sp|P07385
14	HQESEPASVPTSSR	1511.5499	42.86%	1	['Cowpox viru	1	['A18_CWPXB T	1	['>sp|Q8QMT
15	IIPIDNGSNMLILNP	3034.5026	28.57%	1	['Cowpox viru	1	['IL1BP_CWPXB	1	['>sp|Q04523
16	NETYCDMMSLNLT	3896.3331	23.53%	1	['Cowpox viru	1	['IL1BP_CWPXB	1	['>sp|Q04523
17	NDAGYYTCVLK	1246.3886	54.55%	1	['Cowpox viru	1	['IL1BP_CWPXB	1	['>sp|Q04523
18	YTYGDK	745.7766	66.67%	1	['Cowpox viru	1	['IL1BP_CWPXB	1	['>sp|Q04523
19	INPVK	569.694	80.00%	1	['Cowpox viru	1	['IL1BP_CWPXB	1	['>sp|Q04523

Figure 3. Graphical representation of the Purple output. The tabular TSV output of Purple can be imported into various spreadsheet software tools. This exemplary table shows the peptide sequence, the calculated theoretical mass weight (Da), the highest background consensus, and the number of peptide occurrences in the target proteome. The species, protein name and full description of the associated protein are stored in a list for further analysis. In addition, the number of proteins and FASTA entries are listed separately, because they can diverge, e.g., when a protein has multiple sequence variants.

2.3. Data

2.3.1. Target Virus Databases

To evaluate the performance of Purple, selected target virus species from sequence databases were used. This section provides an overview on the virus species used with respect to database composition and further background information on the virus type. The virus species were selected based on their relevance for current or upcoming diagnostic settings.

Arenaviruses

Arenaviruses are enveloped RNA viruses with an average diameter of 120 nanometers that have a bisegmented negative-strand RNA genome. The Latin term "arena" refers to the grainy ribosomal particles acquired from the virus-host cells that can be viewed in cross-section with electron microscopy imaging. Arenaviridae is a virus family whose members are generally associated with causing chronic infections in rodents and zoonotically acquired severe diseases, such as lymphocytic choriomeningitis or hemorrhagic fever, in humans. In this work, nine disease-causing Old and New World arenavirus species are taken as targets for evaluating the performance of Purple (Table 1). Besides Lymphocytic choriomeningitis virus, strain members of which cause aseptic meningitis, encephalitis, or meningoencephalitis, all listed arenaviruses are causative agents for viral hemorrhagic fever (VHF).

Table 1. Alphabetically ordered list of arenavirus species used for the performance benchmarking. The reader is referred to [23] for further details on these arenaviruses.

Virus Species	Abbreviation	NW/OW [2]	NW - Clade [3]	No. Proteins	No. Peptides [1]
Chapare mammarenavirus	CHAV	NW	B	4	252
Guanarito mammarenavirus	GTOV	NW	B	4	244
Junin mammarenavirus	JUNV	NW	B	4	246
Lassa virus	LASV	OW	-	4	242
Lujo mammarenavirus	LUJV	OW [4]	-	4	250
Lymphocytic choriomeningitis virus	LCMV	OW	-	4	245
Machupo virus	MACV	NW	B	4	237
Sabia mammarenavirus	SABV	NW	B	4	248
Whitewater Arroyo mammarenavirus	WWAV	NW	A/rec	4	240

[1] Number of in silico-digested peptide sequences, [2] New World (NW)/ Old World (OW), [3] New World clade [4] Based on genome sequence clustering, Lujo mammaarenavirus shows its own cluster [23].

Cowpox virus

Cowpox virus (CPXV) is a large double-stranded DNA virus with a proteome of over 200 proteins [24] that belongs to the genus Orthopoxvirus (OPV) of the Poxviridae family. CPXV has been described as the source of the first vaccine used by Edward Jenner, who was the first to scientifically describe the vaccination process against the smallpox-causing variola virus. Recent findings based on a conducted analysis on the smallpox vaccine gave evidence of the suspected role of horsepox (instead of cowpox) in the origin of the vaccine [25,26]. Since the pathogenicity and zoonotic potential of CPXV are investigated at the Robert Koch Institute, detailed data acquired from MS measurements were available (see Section 2.3.3). For performance evaluations, CPXV is further beneficial because this virus species has several close relatives. In addition to the cowpox strains Brighton Red and Grishak-90, four very close relatives with high sequence similarity are given: a genome comparison performed with BLAST [27] showed that variola virus, monkeypox virus, horsepox virus, and vaccinia virus share sequence identities of up to 98% (Supplementary Table S1).

Vaccinia virus (VACV Copenhagen and VACV Western Reserve)

Vaccinia virus is a member of the Orthopoxvirus (OPV) genus [28] and has been used for vaccination against smallpox since the 19th century. Due to the high sequence similarity of members of the OPV genus, it is possible to provide cross-protection vaccination by one member of the OPV genus. Hence, the classification can be an issue, because it can be challenging to find peptides to reliably classify a species or a strain. In this work, we investigate whether it is possible to distinguish between the two strains VACV Copenhagen and VACV Western Reserve by finding strain-specific peptides using Purple. Similar to CPXV, experimental data was publically available (see Section 2.3.3).

2.3.2. Background Virus Databases

The target databases mentioned above are species-specific and therefore cannot represent all available virus proteomes. From the target databases, Purple only yields to species-specific unique peptides. To extend this space to all virus proteomes and subsequently be able to find unique peptides in that relation, we added a database that consists of all reviewed virus proteins available on UniProt/Swiss-Prot [29]. In contrast to the target databases, this database is used exclusively as a background database. At the time of writing, UniProt/Swiss-Prot contains 16,846 reviewed viral proteins, which results in 301,387 in silico-digested tryptic peptides. In this work, we evaluate Purple with and without the use of the larger background database.

2.3.3. Background Human Databases

To account for samples mixed with human proteins we added a human database to the background. This database originates from UniProt/Swiss-Prot [29] and enables Purple to discard human peptides. Subsequently, this reduces false positives in experiments using virus-infected human samples. The database consists of 20,428 proteins and was used exclusively for the CPXV analysis in this work.

2.3.4. Experimental Data

The MS/MS datasets used for the benchmarking of Purple originate from a previous study published by Doellinger et al. in 2015 [24] (PRIDE project accession: PXD003013). In this work, a subset of the data available was used including three CPXV Brighton Red, three VACV Copenhagen, and three VACV Western Reserve MS/MS raw files. These raw files were acquired by an LTQ Orbitrap in data-dependent manner. Further experimental details are listed and described in the above-mentioned publication. Subsequently, three CPXV Brighton Red raw files were converted into MGF files using the MSConvert function of ProteoWizard [30] with the peak picking parameter of MS-level two and with zero sampling removal activated. Table 2 shows the number of MS/MS spectra for each virus strain (CPXV Brighton Red, VACV Copenhagen and VACV Western Reserve). For peptide and protein identification, these spectra were searched against proteome databases using the MS-GF+ [31] (version v20181015) database search engine. The database search was performed with eight threads, an activated decoy search, a chosen precursor with mass tolerance of five ppm, optimized for Orbitrap instruments, and trypsin was selected as digestion enzyme. The sequence databases used for protein identification are described in detail in Section 2.3.1. The database searches produced mzid output files that were converted into TSV files using the build-in MS-GF+ conversion tool. Afterwards, the results were filtered by applying a 1% false discovery rate (FDR) threshold at the PSM-level.

Table 2. This table shows the number of spectra from each sample replicate for CPXV Brighton Red, VACV Copenhagen, and VACV Western Reserve virus species/strains.

Species/Strain	No. Spectra in Replicate 1	No. Spectra in Replicate 2	No. Spectra in Replicate 3	No. Total Spectra
CPXV Brighton Red	19,396	19,352	18,920	57,668
VACV Copenhagen	19,740	19,265	19,170	58,175
VACV Western Reserve	19,421	19,453	19,076	57,950

3. Results

We here present three different use cases to illustrate the possibilities of targeted proteomics using Purple in viral diagnostic settings. The first analysis focuses on the species-level resolution for arenaviruses, the second evaluates the taxonomic classification using cowpox data from shotgun proteomics measurements, and the third tests the capabilities of strain-level differentiation using experimental data from two closely related vaccinia virus strains.

3.1. Analysis of Species-Level Resolution using Nine Arenavirus Species

In the first analysis, we aimed to evaluate the species-level resolution of our diagnostic approach using sequence data from the Arenaviridae family. For this purpose, we investigated the resolution of Purple by evaluating different viral species as target organisms against a proteome background of similar species and viruses in general. We used nine arenavirus species (MACV, JUNV, SABV, CHAV, GTOV, LASV, LCMV, WWAV, and LUJV; see Table 1) with proteomes containing four proteins, namely (1) RNA-directed RNA polymerase L, (2) nucleoprotein N, (3) pre-glycoprotein polyprotein GP complex and (4) RING finger protein Z. As background proteomes, we added all reviewed virus proteins available on UniProt/Swiss-Prot to remove frequently occurring peptides (e.g., from conserved sequences of functional domains). The removal of target peptides from similar virus proteomes intends to eliminate false positive detections (i.e., to increase the specificity). Since the protein sequences differ strongly between the arenavirus species, we expected to retrieve sufficient unique peptides for each species that serve as candidates for designing a targeted assay. For a benchmarking, we examined the relative amount of taxon-specific target peptides for each of the arenavirus species using both exact and homologous matching mode (Tables 3 and 4). The homologous matching was performed to evaluate the impact of sequence homologies for the arenaviruses and between these and all other virus species.

Table 3. This table shows the number of taxon-specific peptides from nine arenavirus species after (i) *in silico* digest, (ii) exact matching, and (iii) homologous matching (80% background consensus threshold). Each target species was compared against the background of eight remaining arenavirus species proteomes. The second column provides the number of nonspecific peptides, i.e., the ones being shared with the background.

Species	No. Digested Peptides	No. Background Shared	No. Exact Matching	No. Homologous Matching
MACV	237	119	178	**118**
SABV	248	127	191	**121**
LUJV	250	24	241	**226**
CHAV	252	121	197	**131**
GTOV	244	75	205	**169**
JUNV	246	123	187	**123**
LASV	242	35	227	**207**
LCMV	245	31	232	**214**
WWAV	240	31	226	**209**

Table 4. This table shows the number of taxon-specific peptides from nine arenavirus species after (i) *in silico* digest, (ii) exact matching, and (iii) homologous matching (80% background consensus threshold). Each target species was compared against the background of eight remaining arenavirus species proteomes and additionally against all reviewed virus proteomes (from UniProt/Swiss-Prot). The second column provides the number of nonspecific peptides, i.e., the ones being shared with the background.

Species	No. Digested Peptides	No. Background Shared	No. Exact Matching	No. Homologous Matching
MACV	237	143	162	**94**
SABV	248	144	183	**104**
LUJV	250	52	229	**198**
CHAV	252	137	190	**115**
GTOV	244	118	189	**126**
JUNV	246	139	171	**107**
LASV	242	126	171	**116**
LCMV	245	110	187	**135**
WWAV	240	130	181	**110**

First, we investigated the ratios of taxon-specific unique peptides and in silico-digested peptides with a background database consisting of the four arenavirus proteins, as mentioned above. The exact matching yielded to taxon-specific peptide ratios between 75.1% (MACV) and 96.4% (LUJV) (Figure 4). This can be explained by the high sequence diversity between the nine arenavirus species: when generating multiple sequence alignments (MSA) of these species for their four proteins, overall, a low consensus of the sequences was found (Supplementary Data S1–S4). When applying a background consensus threshold of 80%, significantly fewer taxon-specific peptides were obtained with relative numbers between 48.8% and 90.4% for SABV and LUJV, respectively (Figure 4). Overall, the mean decrease in the ratio of all species is 16.6% and the strongest ratio decrease can be found for MACV (25.3%), SABV (28.2%), CHAV (26.2%), and JUNV (26.0%). These four species are all New World arenaviruses and part of the clade B (see Table 2). The close relationship of these four virus species (as shown in the phylogenetic tree in Figure 5) causes high numbers of shared peptides which explains the decline in taxon-specific peptides. The Old World arenavirus LUJV shows the highest taxon-specific peptide ratio after homologous matching (90.4%) and even after homologous analysis against all virus proteomes (79.2%). This illustrates that LUJV has the lowest sequence similarity with the other arenaviruses. The low similarity can be explained by the isolated geographical distribution of LUJV in Southern Africa [32]. In 2008, an outbreak of LUJV led to a high case fatality rate of 80% (4/5 cases), and a follow-up analysis of its genome confirmed that LUJV is a novel virus species being only distantly related to known arenaviruses and groups genetically closer to Old World viruses not associated with VHF [33].

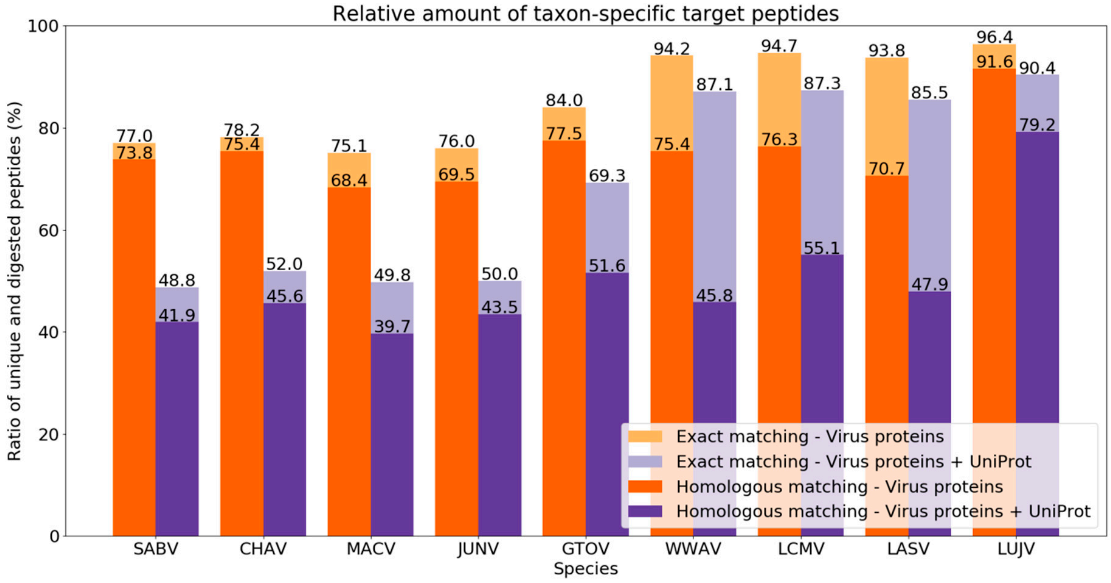


Figure 4. Relative amount of taxon-specific target peptides from nine arenavirus species proteomes. The ratio of unique to in silico-digested peptides is shown for exact (lighter colors) and homologous (darker colors) matching mode with a background consensus threshold of 80%. Orange bars show the results for the database consisting of four virus proteins for each arenavirus species. Purple bars indicate results that were generated when adding protein sequences from all reviewed virus proteomes (from UniProt/Swiss-Prot) as additional background.

Next, we assessed the protein sequence coverage on the basis of Purple-selected unique peptides for all four arenavirus proteins (RNA-directed RNA polymerase L; Nucleoprotein N; Pre-glycoprotein polyprotein GP complex GLYC; RING finger protein Z). We evaluated two different backgrounds here: (i) a small background with the arenavirus proteomes (containing the four proteins) of the remaining eight non-target species and (ii) a large background containing all arenavirus proteomes combined with all reviewed virus proteomes from UniProt/Swiss-Prot (see Section 2.3.2).

The analysis of the protein sequence coverage shows that L, GLYC and Z are relatively well covered by the taxon-specific peptides across all nine species for the small background (Figure 6). Nucleoprotein NCAP has the highest variability in protein coverage with an interquartile range (IQR) of 35.22% on the small background, suggesting that NCAP is the best-conserved protein among the considered arenavirus species. When taking a closer look at the results of the larger background analysis with all reviewed virus proteins, it can be found that the coverage decreases for all four proteins. The NCAP protein shows the lowest median in protein coverage (20.18%). This shows that NCAP has the lowest sequence consensus of taxon-specific peptides with other virus proteomes, indicating that it is the best-conserved of the four proteins. Indeed, the other three proteins (L, GLYC, and Z) have above 40% sequence coverage, thus more taxon-specific peptides can be obtained from these proteins. This analysis shows that, depending on the use case, it may make sense to investigate individual proteins instead of whole proteomes. For example, proteins with low sequence coverage based on taxon-specific peptides may be excluded.

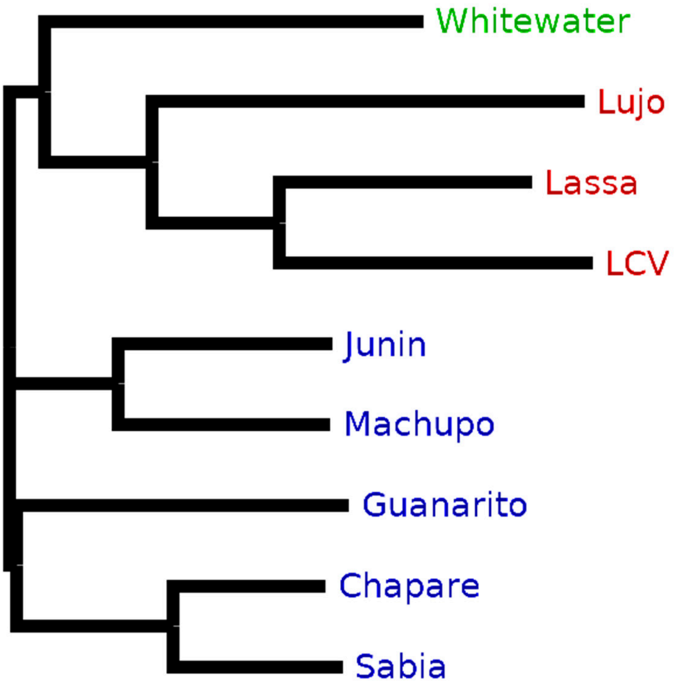


Figure 5. Phylogenetic tree of the pre-glycoprotein polyprotein GP complex (GLYC) of nine arenaviruses. The Whitewater strain is the only New World clade A/rec arenavirus (green). Lujo (LUJV), Lassa (LASV), and Lymphocytic choriomeningitis (LCV) are geographical Old World arenaviruses (red). Junin (JUNV), Machupo (MACV), Guanarito (GTOV), Chapare (CHAV), and Sabia (SABV) are members of the New World arenaviruses clade B (blue). The neighbor-joining tree without distance corrections was created using CLUSTAL Omega [34] for the multiple sequence alignment and the tree visualization software FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

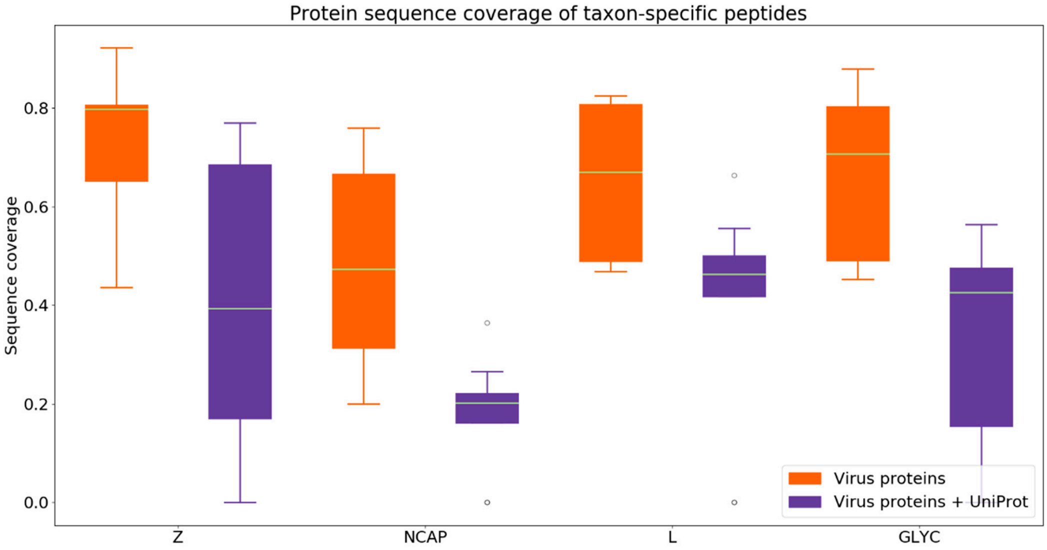


Figure 6. Protein sequence coverage of taxon-specific peptides selected by Purple on proteins for nine arenavirus species proteomes. The four proteins of the arenavirus proteomes are RNA-directed RNA polymerase L (L), nucleoprotein N (NCAP), pre-glycoprotein polyprotein GP complex (GLYC), and RING finger protein Z (Z). The coverage of selected peptides is displayed for homologous matching when applying a background consensus threshold of 80%.

3.2. Evaluating Species-Level Classification Based on Detected Peptides from Viral Shotgun Proteomics Measurements

To evaluate the peptide selection method in Purple on experimental data, we used representative MS/MS datasets derived from human cowpox virus (CPXV) samples. The main goal was to test whether peptides identified in a typical shotgun proteomics experiment can be used for differentiating viruses at the species level. We also aimed for estimating the expected accuracy gain for taxonomic classification when using a targeted proteomics assay on the basis of peptides suggested by Purple.

In a pre-analysis, we performed a Purple run using CPXV as target proteome to select species-specific peptides. For the peptide selection process, 18 reviewed (from UniProt/Swiss-Prot) and 208 unreviewed (from UniProt/TrEMBL) CPXV-specific protein sequences were used as target database, which is part of the PRIDE project (see Section 2.3.4). We used this combined database consisting of reviewed and unreviewed protein sequences because the available reviewed protein sequences for the Brighton Red strain yielded to a very limited number of peptide identifications during the database search (Supplementary Table S2). All available virus proteomes (a total of 16,846 sequences) and all reviewed human proteins were taken as background. These proteomes were obtained from UniProt/Swiss-Prot (see Section 2.3 for database details).

The Purple run resulted in 1509 in silico-digested peptides after exact matching and 885 peptides after homologous matching (using a background consensus threshold of 80%). The distribution of the homologous background consensus shows a normal distribution below 50% (Supplementary Figure S2). 3986 peptides were discarded, because they were shared with other (i.e., non-CPXV) viral proteomes or the human proteome. The remaining 885 CPXV-specific peptides have a mean background consensus of 53.9%, which means that on average around half of the amino acids of each peptide are equal to residues of peptides in the background.

Next, we searched experimental MS/MS spectra from CPXV samples using the search algorithm MS-GF+ [31] against a CPXV and human sequence database for peptide identification (see Section 2.3). In this analysis, CPXV datasets from MS measurements of three technical replicates, each with ~19,000 MS/MS spectra, were evaluated. The database search resulted in 4028, 4125, and 3967 identified peptides per sample replicate with sequence duplicates removed. More than twice the amount of CPXV peptides were identified as human peptides in this sample before applying a FDR filtering. After applying an FDR threshold of 1%, 1067, 1028, and 1004 CPXV peptides were identified (Table 5). Subsequently, the identified peptides (below 1% FDR threshold) were compared against the set of taxon-specific CPXV peptides suggested by Purple using both exact and homologous matching mode. Between 83 and 94 peptides selected by Purple were detected in the MS/MS experiments (without applying any FDR threshold). When filtered by 1% FDR, the peptides decreased to numbers between 78 and 84. Consequently, this analysis demonstrates that it would be possible to reliably identify CPXV for these three sample replicates.

Table 5. This table shows the number of peptides from the cowpox virus (CPXV) Brighton Red strain after (i) database search with duplicates removed (CPXV); (ii) database search with duplicates removed (human); (iii) intersection of peptides obtained from Purple and peptide identifications from database search; (iv) database search, duplicates removed and filtered by 1% FDR threshold; and (iiv) intersection of peptides suggested by Purple and peptide identifications from FDR-filtered database search. The CPXV Brighton Red strain was compared against the background of all reviewed virus proteomes and the reviewed human proteome. In addition, the second column specifies the sample replicate data that was used for the database search.

Strain	Replicate	No. Database Search (CPXV)	No. Database Search (HUMAN)	No. Intersection	No. Database Search Filtered	No. Intersection Filtered
Brighton Red	1	4028	10319	94	1067	**84**
Brighton Red	2	4125	10286	83	1028	**78**
Brighton Red	3	3967	10068	92	1004	**84**

When considering the results of all three replicates, it can be observed that 61 CPXV-specific peptides were detected without any applied FDR threshold (Figure 7A). Filtered by 1% FDR, 56 peptides across all replicates can be used to specifically identify the species within the sample as a member of CPXV (Figure 7B).

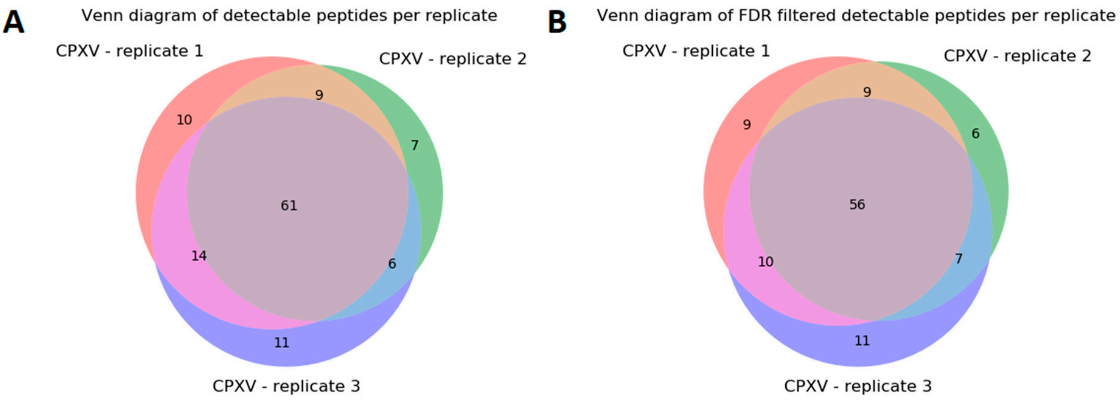


Figure 7. Intersection of detectable peptides of CPXV sample replicates. These Venn diagrams show the intersection of the detectable peptides in replicates 1–3. The subfigures depict the number of peptides without applying any false discovery rate (FDR) threshold (**A**) and filtered by 1% FDR (**B**).

When examining the peptides shared by the target and background proteomes, it can be found that the Cowpox virus shares ~3000 peptide sequences per strain with the Vaccinia virus strains and Variola virus strains (Figure 8). Other Orthopoxviruses were found as well, although the number of peptides is low, due to fewer proteins of these strains in the background database. The CPXV Brighton Red strain-specific peptides are small in number because most matches originate from the Cowpox virus species proteome without giving any details about a particular strain. Around 500 peptides were shared with the human proteome and were consequently discarded.

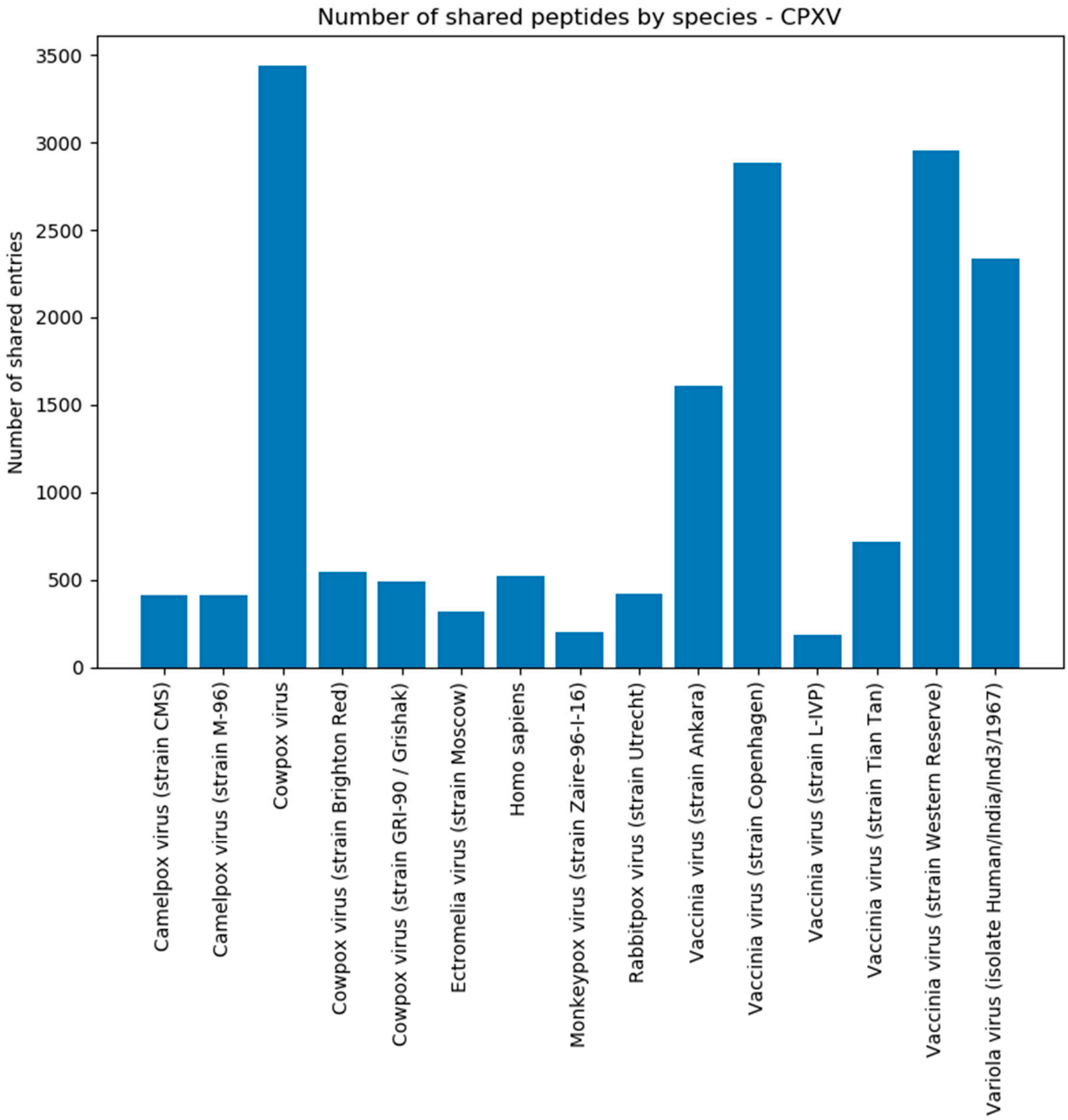


Figure 8. Number of shared CPXV peptides by species/strain assigned. This plot shows the number of shared peptides that Purple detected in the background for a species/species after the CPXV Brighton Red target analysis. All species that contribute less than 0.5% to the total amount of shared peptides were removed.

3.3. Comparison of Strain vs. Strain and Strain vs. All Virus Level Resolution

Next, we conducted a performance evaluation using two different, yet highly similar Vaccinia virus strains, namely VACV Copenhagen and VACV Western Reserve. The objective was to test whether Purple can retrieve strain-specific peptides that are then used in the targeted proteomics assay for accurate taxonomic classification. In this analysis, the target database contained sequences from one of the two VACV virus strains (either Copenhagen or Western reserve). Consequently, the background database contained the remaining VACV strain and all reviewed virus proteins available on UniProt. This procedure was repeated with the remaining VACV strains as target. The goal was to find strain-specific peptides to accurately detect the virus strain. We used a background consensus threshold of 80% to filter out homologous peptides. Afterwards, experimental data (see Section 2.3.3) was used to validate the results and to show if the selected strain-specific peptides are found in the acquired tandem mass spectrometry (MS/MS) data. For peptide identification, we used the software MS-GF+ [31] with an 1% FDR threshold (see Section 2.3.3).

In the case of VACV Copenhagen, Purple discarded 3848 peptides because a perfect sequence match was present in the background with a peptide of another strain or virus (Table 6). Equally, 3971 VACV Western Reserve peptides are marked as shared with the background and discarded. After exact matching, 498 and 341 strain-specific peptides could be obtained for VACV Copenhagen and VACV Western Reserve, respectively. The homologous matching removed additional 157 (VACV Copenhagen) and 172 (VACV Western Reserve) peptides from the set of unique peptides. The remaining 352 (VACV Copenhagen) and 169 (VACV Western Reserve) peptides can be used to uniquely identify the strain in a mixture of all reviewed virus proteins available on UniProt/Swiss-Prot.

Table 6. This table shows the number of taxon-specific peptides from the VACV Copenhagen and VACV Western Reserve strain after (i) *in silico* digest, (ii) exact matching, and (iii) homologous matching (80% background consensus threshold). Each target strain was compared against the background of the other strain and all reviewed virus proteomes. The second column provides the number of nonspecific peptides, i.e., the ones being shared with the background.

Species	No. Digested Peptides	No. Background Shared	No. Exact Matching	No. Homologous Matching
Copenhagen	4200	3848	498	**352**
Western Reserve	4140	3971	341	**169**

In addition, we categorized the shared peptides by virus species to check for close relationships in the background. For VACV Copenhagen, it can be observed that most peptide matches are found in the Vaccinia species (Figure 9), owing to a high protein sequence similarity of involved Vaccinia strains. Other contributing species are Camelpox virus, Cowpox virus, Monkeypox virus, Rabbitpox virus, and Ectromelia virus. All these viruses are, as expected, members of the orthopoxvirus genus. Similar findings could be observed for the results of the VACV Western Reserve strain (Supplementary Figure S1). Note here that Figure 9 shows the number of peptides and if a species is underrepresented in the databases, it will affect the outcome concerning the number of peptides that contribute to the shared peptides.

To evaluate the detectability of taxon-specific peptides for the given DDA experiments, we performed database searches for peptide identification using three different technical replicates of VACV Copenhagen. Without any FDR cut-off, we could identify between 60 and 66 strain-specific peptides selected by Purple (Table 7). However, when filtered by an FDR of 1% the number of peptides decreased drastically and only one or two taxon-specific peptides were confirmed in the shotgun proteomics data. It was possible to identify Replicate 1 and 2 as VACV Copenhagen by using the peptide sequence ILFWPYIEDELR. The number of peptides can be increased by switching to a targeted proteomics approach and by considering PTMs or by an improved homologous matching. The three technical replicates of the VACV Western Reserve strain resulted in fewer peptides in the intersection with the database search results (between 32 and 42), but when filtered by 1% FDR, the number of peptides was increased up to 11-fold (with nine to 11 peptides) in comparison to the VACV Copenhagen replicates. Six peptides were detected, and their sequences were identical among all three replicates.

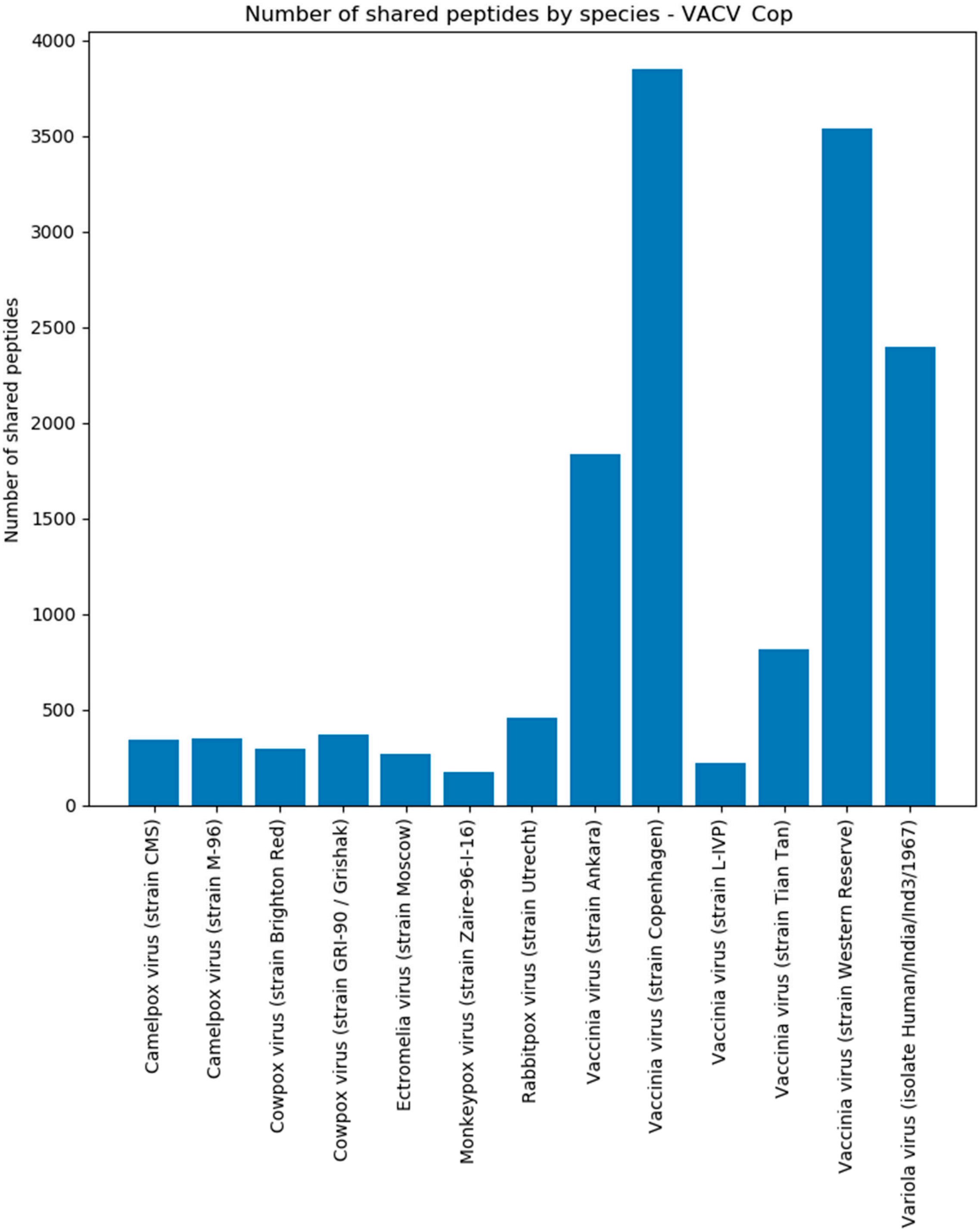


Figure 9. Number of shared peptides by species. This plot shows the number of shared peptides that Purple detected in the background for a species after the VACV Copenhagen analysis. All species that contribute less than 0.5% to the total amount of shared peptides were removed here.

Table 7. This table shows the number of peptides from VACV Copenhagen and VACV Western Reserve strain after (i) database search with duplicates removed; (ii) intersection of peptides obtained by Purple and database search; (iii) database search, duplicates removed and filtering by FDR; and (iv) intersection of peptides obtained by Purple and filtered database search. Each target strain was compared against the background of the other strain and all reviewed virus proteomes. The second column specifies the replicate data that was used for the database search.

Strain	Replicate	No. Database Search	No. Intersection	No. Database Search Filtered	No. Intersection Filtered
Copenhagen	1	3585	66	825	**2**
Copenhagen	2	3507	62	800	**1**
Copenhagen	3	3525	60	828	**1**
Western Reserve	1	3636	35	841	**9**
Western Reserve	2	3736	42	800	**11**
Western Reserve	3	3507	32	809	**9**

In conclusion, we were able to identify every strain in each sample with an applied FDR of 1%. For VACV Western Reserve, the number of peptides was higher than for the VACV Copenhagen strain. The number of detectable peptides could be increased by improving scoring and filtering or by switching from shotgun to targeted proteomics methods or by considering PTMs.

Figure 10 reveals a normal distributed homologous consensus in the interval from 10% to 50%. This is caused by random matches with background peptides and these peptides should be unique for the strain. We could not observe a distinct distribution above 50%. This could be improved by moving from identity to a similarity-based matching, as this would differentiate peptides with the same amount of matching consensus residuals.

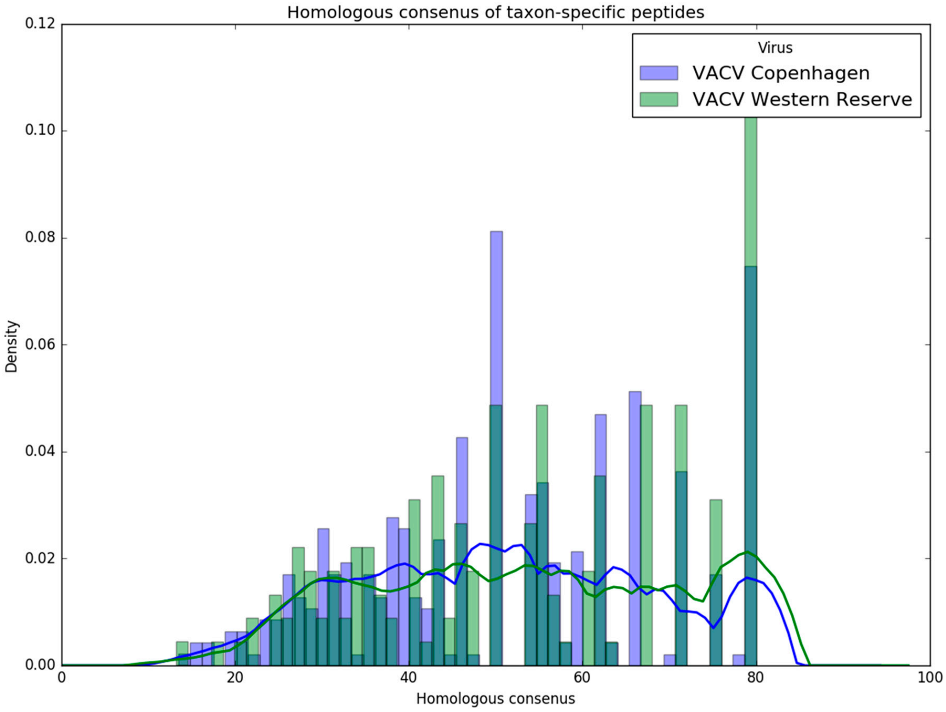


Figure 10. Histogram and density plot of homologous consensus. This histogram shows the distribution of the homologous consensus for the VACV Copenhagen (blue) and Western Reserve (green) strains. Additionally, the kernel density was calculated utilizing the Epanechnikov kernel and a Silverman bandwidth estimation.

4. Discussion

The main goal of our developed Purple software is to provide taxon-specific peptides for a targeted proteomics assay. These targeted assays can be used in a diagnostic setting to identify a virus species/strain or even a whole virus family in a sample in sensitive and time-efficient manner. In this work, we validated the software in three different benchmarking experiments.

Purple enabled us to retrieve taxon-specific peptides to distinguish between arenavirus species proteomes that are very similar in their sequences (see Section 3.1). Accordingly, we observed a comparable decrease in the ratio of unique to *in silico*-digested peptides for New and Old World arenaviruses based on differences between their proteomes (Figure 4). This effect could also be recognized also on the clade level for the New World viruses.

The data analysis of CPXV (see Section 3.2) resulted in 56 taxon-specific peptides (Figure 7). These peptides were present in each MS/MS sample replicate and can be used to uniquely identify CPXV in a mixed biological sample, although its proteome is very similar to other Orthopoxvirus species and strains (Figure 8). By changing to a Brighton Red strain-specific target database, a reliable determination of the strain would be possible as well. This underlines that Purple relies on a correct and complete database to yield to the best possible results. Missing or incorrectly assigned protein sequences could result in incorrect selected unique peptides or discarded ones. Furthermore, although many spectra in the shotgun proteomics experiment were assigned to human peptides, this does not present a limitation for the targeted proteomics approach, because unique virus peptides selected by Purple can be detected using a targeted (e.g., PRM-based) assay in specific and sensitive manner; for example, in a recently published study [35], a PRM-based assay was used to identify dengue virus species directly from clinical serum samples. Nevertheless, to validate the resulting set of peptides, it would be recommended to test them on other CPXV samples and to check if the peptides are detectable in these samples likewise. In addition, the selected background database might be incomplete, e.g., when proteome references were missed to be included for the Purple analysis. In this case, it is useful to validate Purple-selected peptides using secondary tools such as Unipept [36] for resolving the taxonomic origin of any tryptic peptide based on the complete UniProt database. Furthermore, false negatives may result from issues during sample preparation or poor instrument performance. Therefore, these parameters need to be controlled in diagnostic PRM assays, e.g., by using internal standards and running further quality control samples.

It can be crucial in virus infection scenarios to accurately distinguish between specific strains. To cover these cases, we examined the strain-level resolution of our tool using data of VACV Copenhagen and VACV Western Reserve strains (see Section 3.3). Purple was able to find a reliable amount of strain-specific peptides (Table 7). The intersection between the Purple-selected peptides and the peptide identification from the database search showed that it is possible to detect these peptides. In general, strain-level identification was possible even for an applied FDR threshold of 1%, however, it became apparent that the shotgun proteomics approach becomes limited due to the spurious numbers of identified peptides. The number of peptides could be increased by adjusting the FDR filtering or by using a targeted proteomics approach with higher sensitivity.

In comparison to other tools, Purple offers several advantages, such as cross-platform compatibility on multiple operating systems. Purple allows a homology-based analysis of multiple proteome databases at once and produces an aggregated and summarized export on various levels. In addition, Purple is not limited to specific organisms, but can be used with general UniProt databases, also including eukaryotic and bacterial databases. High sequence similarity between strains and horizontal gene transfer may complicate taxon-specific classification for bacterial samples. However, Purple could help to overcome complications and can be helpful for creating targeted assays for bacterial detection as well. The graphical user interface and compatibility with all UniProt databases enables researchers without bioinformatics background to find taxon-specific peptides in an easy and straightforward manner.

A potential improvement to the software would be to move from a sequence identity-based metric based on the Hamming distance to similarity-based matching for the homologous matching mode. In this case, amino acid substitutions are not weighted equally, for example by using a PAM or BLOSUM matrix [37]. This similarity-based metric might allow a more accurate homologous matching in Purple. For example, an approach based on a structural alignment as introduced by Ogata et al. [38] might be useful. Further potential improvements with useful features in Purple include adding plots for better data exploration and a tabular view for inspecting the results (that are currently exportable as text files to spreadsheet software).

In summary, the most promising application of Purple is to select taxon-specific peptides for creating tailored SRM or PRM assays with high sensitivity and specificity. This application will allow for new time- and cost-efficient diagnostic methods in healthcare and further biological applications. It could even be used to identify multiple organisms in a single sample in the context of targeted metaproteomics [39].

Purple is available for download on our GitLab website (https://gitlab.com/rki_bioinformatics), by using the Python package manager pip (https://pypi.org/project/purple-bio/) or via the Bioconda channel (https://anaconda.org/bioconda/purple-bio) [40]. The software is available as graphical user interface version, Python package and command line version for Windows, Linux, and MacOS. In addition, user support, tutorials, and the documentation manual can be found on the GitLab webpages.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/11/6/536/s1, Table S1: Genome sequence similarities of cowpox virus; Table S2: Number of peptides from CPXV Brighton Red strain processing; Figure S1: Number of shared peptides by species for VACV Western Reserve; Figure S2: Histogram and density plot of homologous consensus—CPXV; Data S1: MSA of the pre-glycoprotein polyprotein GP complex (GPC gene); Data S2: MSA of nucleocapsid protein (N gene); Data S3: MSA of RNA-directed RNA polymerase L (L gene); Data S4: MSA of RING finger protein Z (Z gene).

Author Contributions: Conceptualization: J.D., B.Y.R., and T.M.; methodology: J.D., J.L., P.H., and T.M.; software: F.H., J.L., and P.H.; validation: F.H. and J.L.; formal analysis: F.H., J.L., and T.M..; investigation: F.H. and J.L..; resources: A.N., B.Y.R., J.D., and M.G.; data curation: F.H.; writing—original draft preparation: F.H. and J.L.; writing—review and editing: A.N., B.Y.R., J.D., M.G., P.H., and T.M.; visualization: F.H.; supervision: B.Y.R. and T.M.; project administration: T.M.

Funding: This research was funded by the German Research Foundation (DFG), grand number RE3474/2-2, and intramural funds (#8321566) of the Robert Koch Institute.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Singhal, N.; Kumar, M.; Kanaujia, P.K.; Virdi, J.S. MALDI-TOF mass spectrometry: An emerging technology for microbial identification and diagnosis. *Front. Microbiol.* **2015**, *6*, 791. [CrossRef] [PubMed]
2. Ebhardt, H.A.; Root, A.; Sander, C.; Aebersold, R. Applications of targeted proteomics in systems biology and translational medicine. *Proteomics* **2015**, *15*, 3193–3208. [CrossRef] [PubMed]
3. Deutsch, E.W. The PeptideAtlas project. *Methods Mol. Biol.* **2010**, *604*, 285–296. [PubMed]
4. Picotti, P.; Aebersold, R. Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. *Nat. Methods* **2012**, *9*, 555–566. [CrossRef] [PubMed]
5. Peterson, A.C.; Russell, J.D.; Bailey, D.J.; Westphall, M.S.; Coon, J.J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteom.* **2012**, *11*, 1475–1488. [CrossRef] [PubMed]
6. Borràs, E.; Sabidó, E. What is targeted proteomics? A concise revision of targeted acquisition and targeted data analysis in mass spectrometry. *Proteomics* **2017**, *17*, 17–18.
7. Karlsson, C.; Malmström, L.; Aebersold, R.; Malmström, J. Proteome-wide selected reaction monitoring assays for the human pathogen *Streptococcus pyogenes*. *Nat. Commun.* **2012**, *3*, 1301. [CrossRef]
8. Peters, J.S.; Calder, B.; Gonnelli, G.; Degroeve, S.; Rajaonarifara, E.; Mulder, N.; Blackburn, J.M. Identification of quantitative proteomic differences between *Mycobacterium tuberculosis* lineages with altered virulence. *Front. Microbiol.* **2016**, *7*, 813. [CrossRef]

9. MacLean, B.; Tomazela, D.M.; Shulman, N.; Chambers, M.; Finney, G.L.; Frewen, B.; MacCoss, M.J. Skyline: An. open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, *26*, 966–968. [CrossRef]
10. Mohammed, Y.; Domański, D.; Jackson, A.M.; Smith, D.S.; Deelder, A.M.; Palmblad, M.; Borchers, C.H. PeptidePicker: A scientific workflow with web interface for selecting appropriate peptides for targeted proteomics experiments. *J. Proteom.* **2014**, *106*, 151–161. [CrossRef]
11. Demeure, K.; Duriez, E.; Domon, B.; Niclou, S.P. PeptideManager: A peptide selection tool for targeted proteomic studies involving mixed samples from different species. *Front. Genet.* **2014**, *5*, 305. [CrossRef] [PubMed]
12. Zauber, H.; Kirchner, M.; Selbach, M. Picky: A simple online PRM and SRM method designer for targeted proteomics. *Nat. Methods* **2018**, *15*, 156–157. [CrossRef] [PubMed]
13. Zolg, D.P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Yu, P. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **2017**, *14*, 259–262. [CrossRef] [PubMed]
14. Zolg, D.P.; Wilhelm, M.; Schmidt, T.; Medard, G.; Zerweck, J.; Knaute, T.; Kuster, B. ProteomeTools: Systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides. *Mol. Cell. Proteom.* **2018**, *17*, 1850–1863. [CrossRef] [PubMed]
15. Mesuere, B.; Van der Jeugt, F.; Devreese, B.; Vandamme, P.; Dawyndt, P. The unique peptidome: Taxon-specific tryptic peptides as biomarkers for targeted metaproteomics. *Proteomics* **2016**, *16*, 2313–2318. [CrossRef] [PubMed]
16. Mallick, P.; Schirle, M.; Chen, S.S.; Flory, M.R.; Lee, H.; Martin, D.; Kuster, B. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **2007**, *25*, 125–131. [CrossRef] [PubMed]
17. Eyers, C.E.; Lawless, C.; Wedge, D.C.; Lau, K.W.; Gaskell, S.J.; Hubbard, S.J. CONSeQuence: Prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol. Cell. Proteom.* **2011**, *10*, M110-003384. [CrossRef]
18. Qeli, E.; Omasits, U.; Goetze, S.; Stekhoven, D.J.; Frey, J.E.; Basler, K.; Ahrens, C.H. Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data. *J. Proteom.* **2014**, *108*, 269–283. [CrossRef]
19. Tang, H.; Arnold, R.J.; Alves, P.; Xun, Z.; Clemmer, D.E.; Novotny, M.V.; Radivojac, P. A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics* **2006**, *22*, e481–e488. [CrossRef]
20. Sanders, W.S.; Bridges, S.M.; McCarthy, F.M.; Nanduri, B.; Burgess, S.C. Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinform.* **2007**, *8*, S23. [CrossRef]
21. Fusaro, V.A.; Mani, D.R.; Mesirov, J.P.; Carr, S.A. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.* **2009**, *27*, 190–198. [CrossRef] [PubMed]
22. Cock, P.J.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; De Hoon, M.J. Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef] [PubMed]
23. Olayiwola, J.O.; Bakarey, A.S. Epidemiological trends of Lassa fever outbreaks and insights for future control in Nigeria. *Int. J. Trop. Dis. Heal.* **2017**, *24*, 1–14. [CrossRef]
24. Doellinger, J.; Schaade, L.; Nitsche, A. Comparison of the cowpox virus and Vaccinia virus mature virion proteome: Analysis of the Species- and strain-specific proteome. *PLoS ONE* **2015**, *10*, e0141527. [CrossRef] [PubMed]
25. Schrick, L.; Tausch, S.H.; Dabrowski, P.W.; Damaso, C.R.; Esparza, J.; Nitsche, A. An early American smallpox vaccine based on horsepox. *N. Engl. J. Med.* **2017**, *377*, 1491–1492. [CrossRef] [PubMed]
26. Esparza, J.; Schrick, L.; Damaso, C.R.; Nitsche, A. Equination (inoculation of horsepox): An. early alternative to vaccination (inoculation of cowpox) and the potential role of horsepox virus in the origin of the smallpox vaccine. *Vaccine* **2017**, *35*, 7222–7230. [CrossRef] [PubMed]
27. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
28. Jacobs, B.L.; Langland, J.O.; Kibler, K.V.; Denzler, K.L.; White, S.D.; Holechek, S.A.; Baskin, C.R. Vaccinia virus vaccines: Past, present and future. *Antivir. Res.* **2009**, *84*, 1–13. [CrossRef]

29. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Martin, M.J. UniProt: The Universal protein knowledgebase. *Nucleic Acids Res.* **2004**, *32*, D115–D119. [CrossRef]
30. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534–2536. [CrossRef]
31. Kim, S.; Pevzner, P.A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277. [CrossRef] [PubMed]
32. Fehling, S.; Lennartz, F.; Strecker, T. Multifunctional nature of the arenavirus RING finger protein Z. *Viruses* **2012**, *4*, 2973–3011. [CrossRef] [PubMed]
33. Briese, T.; Paweska, J.T.; McMullan, L.K.; Hutchison, S.K.; Street, C.; Palacios, G.; Nichol, S.T. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog.* **2009**, *5*, e1000455. [CrossRef] [PubMed]
34. Chojnacki, S.; Cowley, A.; Lee, J.; Foix, A.; Lopez, R. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Res.* **2017**, *45*, W550–W553. [CrossRef] [PubMed]
35. Wee, S.; Alli-Shaik, A.; Kek, R.; Swa, H.L.; Tien, W.P.; Lim, V.W.; Gunaratne, J. Multiplex targeted mass spectrometry assay for one-shot flavivirus diagnosis. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 6754–6759. [CrossRef]
36. Mesuere, B.; Devreese, B.; Debyser, G.; Aerts, M.; Vandamme, P.; Dawyndt, P. Unipept: Tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.* **2012**, *11*, 5773–5780. [CrossRef]
37. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915–10919. [CrossRef] [PubMed]
38. Ogata, K.; Ohya, M.; Umeyama, H. Amino acid similarity matrix for homology modeling derived from structural alignment and optimized by the Monte Carlo method. *J. Mol. Gr. Model.* **1998**, *16*, 178–189. [CrossRef]
39. Saito, M.A.; Dorsk, A.; Post, A.F.; McIlvin, M.R.; Rappé, M.S.; DiTullio, G.R.; Moran, D.M. Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* **2015**, *15*, 3521–3531. [CrossRef]
40. Dale, R.; Grüning, B.; Sjödin, A.; Rowe, J.; Chapman, B.A.; Tomkins-Tinch, C.H.; Köster, J. Bioconda: A sustainable and comprehensive software distribution for the life sciences. bioRxiv 2017, bioRxiv:207092. *bioRxiv* **2017**. bioRxiv:207092.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

frontiers in Microbiology

ORIGINAL RESEARCH
published: 16 August 2019
doi: 10.3389/fmicb.2019.01883

A Robust and Universal Metaproteomics Workflow for Research Studies and Routine Diagnostics Within 24 h Using Phenol Extraction, FASP Digest, and the MetaProteomeAnalyzer

*Robert Heyer[1†], Kay Schallert[1†], Anja Büdel[1], Roman Zoun[2], Sebastian Dorl[3], Alexander Behne[4], Fabian Kohrs[1], Sebastian Püttker[1], Corina Siewert[5], Thilo Muth[6], Gunter Saake[2], Udo Reichl[1,5] and Dirk Benndorf[1,5]**

[1] Bioprocess Engineering, Otto von Guericke University Magdeburg, Magdeburg, Germany, [2] Database Research Group, Otto von Guericke University Magdeburg, Magdeburg, Germany, [3] Bioinformatics Research Group, University of Applied Sciences Upper Austria, Hagenberg, Austria, [4] glyXera GmbH, Magdeburg, Germany, [5] Bioprocess Engineering, Max Planck Institute for Dynamics of Complex Technical Systems Magdeburg, Magdeburg, Germany, [6] Bioinformatics Unit (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany

OPEN ACCESS

Edited by:
*Esteban A. Hernandez-Vargas,
Frankfurt Institute for Advanced Studies, Germany*

Reviewed by:
*Brook Leanne Nunn,
University of Washington, United States
Martin Taubert,
Friedrich Schiller University Jena, Germany
Carolin Adriane Kolmeder,
Max Planck Institute for Developmental Biology, Germany*

***Correspondence:**
*Dirk Benndorf
benndorf@mpi-magdeburg.mpg.de*

† Equal authorship

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *20 February 2019*
Accepted: *30 July 2019*
Published: *16 August 2019*

Citation:
Heyer R, Schallert K, Büdel A, Zoun R, Dorl S, Behne A, Kohrs F, Püttker S, Siewert C, Muth T, Saake G, Reichl U and Benndorf D (2019) A Robust and Universal Metaproteomics Workflow for Research Studies and Routine Diagnostics Within 24 h Using Phenol Extraction, FASP Digest, and the MetaProteomeAnalyzer. Front. Microbiol. 10:1883. doi: 10.3389/fmicb.2019.01883

The investigation of microbial proteins by mass spectrometry (metaproteomics) is a key technology for simultaneously assessing the taxonomic composition and the functionality of microbial communities in medical, environmental, and biotechnological applications. We present an improved metaproteomics workflow using an updated sample preparation and a new version of the MetaProteomeAnalyzer software for data analysis. High resolution by multidimensional separation (GeLC, MudPIT) was sacrificed to aim at fast analysis of a broad range of different samples in less than 24 h. The improved workflow generated at least two times as many protein identifications than our previous workflow, and a drastic increase of taxonomic and functional annotations. Improvements of all aspects of the workflow, particularly the speed, are first steps toward potential routine clinical diagnostics (i.e., fecal samples) and analysis of technical and environmental samples. The MetaProteomeAnalyzer is provided to the scientific community as a central remote server solution at www.mpa.ovgu.de.

Keywords: bioinformatics, software, sample preparation, environmental proteomics, microbial communities, mass spectrometry, gut microbiome

INTRODUCTION

The metabolism of microbial communities is determined by the proteome, the total set of proteins of the microbial cells, including enzymes for growth and maintenance. The expression of proteins depends on the environmental conditions, community composition, and the metabolic activity of the individual microorganisms (Wasinger et al., 1995). Metaproteomics, the identification

Abbreviations: BGP, biogas plant; de.NBI, German Network for Bioinformatics Infrastructure; DTT, dithiothreitol; FASP, filter aided sample prep; Hgut, human gut; IAA, iodoacetamide; LC-MS/MS, liquid chromatography tandem mass spectrometer; MPA, MetaProteomeAnalyzer; MPAv1, MetaProteomeAnalyzer version 1.0.5; MPAv2, MetaProteomeAnalyzer version 2.12; MS, mass spectrometry/mass spectrometer; PCoA, principal coordinates analysis; RT, room temperature; SOP, standard operation procedure; TFA, trifluoroacetic acid; WWTP, wastewater treatment plant.

of microbial proteins using MS (Wilmes and Bond, 2006), is crucial to understand microbial communities. Due to the rapid development of MS, the number of conducted metaproteomics studies has increased over the last years. Microbiomes from the human gut (Kolmeder et al., 2012; Xiong et al., 2015; Zhang et al., 2018a), rumen (Deusch et al., 2017), soil (Bastida and Jehmlich, 2016; Keiblinger et al., 2016), or BGPs (Heyer et al., 2016; Hagen et al., 2017) were measured. Metaproteomics aims at deeper insights into microbiomes by analyzing taxonomic and functional composition of complex microbial communities in diverse environments and technical applications. Based on metaproteome data the state of microbial communities can be linked with certain environmental conditions or process parameters. However, metaproteomics also has the potential to serve as a tool for diagnostics in clinical settings or routine process monitoring (Heyer et al., 2017). For example, proteins of the microbial community in the human gut or in a BGP may represent valuable markers for diseases or process disturbances in BGP, respectively. Such routine application of metaproteomics is not common yet, due to two major challenges (i) sample preparation due to high complexity and contamination of samples, and (ii) data analysis due to the required computational effort for large datasets, missing corresponding annotated protein sequence databases, and protein inference causing ambiguity of protein annotation.

The first challenge is the time-consuming sample preparation workflow and its sensitivity to sample impurities (Heyer et al., 2015). Common metaproteomics workflows comprise of protein extraction and purification, tryptic digestion of proteins into peptides, and measurement by LC-MS/MS. The amount of extracted proteins is measured by different assays, and the complexity of protein extracts is often reduced by fractionation using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (Heyer et al., 2015; Wenzel et al., 2018) or two dimensional chromatography (Erickson et al., 2012; Kleiner et al., 2017). In consequence, the total workflow for sample preparation can take up to 1 week, but routine diagnostics should not exceed 24 h for complete analysis. Therefore, we choose to sacrifice fractionation, since monitoring of the main microbial processes and highly abundant marker proteins do not require such a high coverage of the metaproteome. Different protocols exist for protein extraction and protein purification (Keiblinger et al., 2012; Zhang et al., 2018b), depending on the sample type. Samples from microbial communities from fresh water or the ocean are almost free of impurities, and proteins can be extracted easily (Colatriano and Walsh, 2015). In contrast, soil and BGP samples contain high amounts of humic substances (Heyer et al., 2015; Keiblinger et al., 2016), which require specialized extraction methods such as phenol extraction (Heyer et al., 2013) or trichloroacetic acid precipitation (Chourey et al., 2010). Adaptation of the workflow for each sample type is time consuming and not feasible for routine application, therefore, we choose phenol extraction in this study, since it provides robust protein recovery from different sample types (Benndorf et al., 2007, 2009; Keiblinger et al., 2012; Heyer et al., 2013; Püttker et al., 2015).

The second challenge concerns the data analysis. Proteins are commonly identified by comparing experimental peptide spectra against theoretical spectra derived from protein sequence databases (Mann and Wilm, 1994). Subsequently, identified proteins are assigned by taxonomy and function. However, three issues specific to metaproteomics hamper and delay bioinformatics evaluation (Muth et al., 2013). First, the amount of acquired data is huge due to the high complexity of microbial communities, which results in enormous demands regarding computing resources. Modern LC-MS/MS instruments produce tens of thousands high-resolution spectra per hour. This enables in-depth analysis of the metaproteome but increases the computational load significantly. Second, protein identification can be difficult due to the lack of suitable protein or metagenome databases. Third, the interpretation of taxonomic and functional results is difficult due to the problem of protein inference (Nesvizhskii and Aebersold, 2005) from conserved sequences in homologous proteins.

To tackle these issues, the MPA was developed as an intuitive open-source software platform for metaproteomics data analysis and interpretation (Muth et al., 2015a). Among other features, it supports the handling of protein inference by grouping proteins into protein groups (called metaproteins hereafter). The generation of metaproteins is a strategy that was developed specifically for the metaproteomics field. The latest implementation of the MPA (version 3.0.0) also allows for easy comparison of results from different experiments and provides supplementary annotation functions for protein entries from metagenome sequences (regarding taxonomies or protein functions).

In this paper, a complete metaproteomics workflow is described where all processing steps from sample preparation to visualization are performed within 24 h, referred to as "new workflow" hereafter. The objectives of our new protocols were speed, simplicity, high throughput, reproducibility, and robustness to establish metaproteomics as routine application in applied research and diagnostics. This new workflow was applicable to various types of samples and drastically decreased overall processing time from at least 3 days to only 1 day. The aim of the presented workflow was not to provide discovery oriented, in-depth analysis of microbial communities. Instead, it constituted an important milestone toward routine monitoring of biotechnological processes and analysis of clinical samples, since such routine analyses should not exceed a 24 h time period or require complicated adaptations of the laboratory procedures. In order to achieve this goal, phenol extraction was optimized compared to previous studies (Heyer et al., 2013), in-gel digestion was replaced by FASP digestion (new sample preparation)(Wisniewski et al., 2009), and the MPA software (Muth et al., 2015a) was continuously updated (current MPA version 3.0.0).

MATERIALS AND METHODS

For an overview, refer to the complete workflow steps A1-A6, B1-B8, and X1-X3 in **Figure 1**. For a more detailed

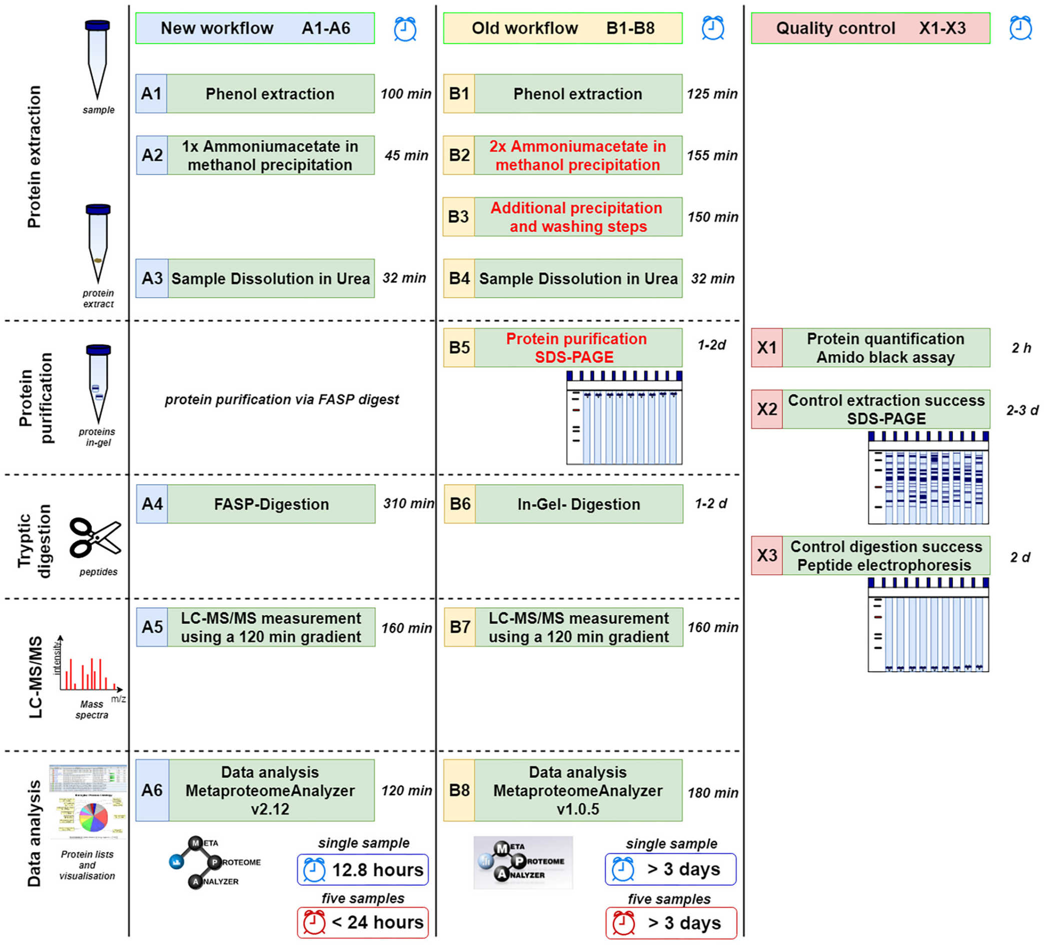

New workflow A1-A6
Old workflow B1-B8
Quality control X1-X3
Protein extraction
sample
protein extract
A1 Phenol extraction 100 min
A2 1x Ammoniumacetate in methanol precipitation 45 min
A3 Sample Dissolution in Urea 32 min
B1 Phenol extraction 125 min
B2 2x Ammoniumacetate in methanol precipitation 155 min
B3 Additional precipitation and washing steps 150 min
B4 Sample Dissolution in Urea 32 min
Protein purification
proteins in-gel
protein purification via FASP digest
B5 Protein purification SDS-PAGE 1-2d
X1 Protein quantification Amido black assay 2 h
X2 Control extraction success SDS-PAGE 2-3 d
Tryptic digestion
peptides
A4 FASP-Digestion 310 min
B6 In-Gel- Digestion 1-2 d
X3 Control digestion success Peptide electrophoresis 2 d
LC-MS/MS
intensity
m/z
Mass spectra
A5 LC-MS/MS measurement using a 120 min gradient 160 min
B7 LC-MS/MS measurement using a 120 min gradient 160 min
Data analysis
Protein lists and visualisation
A6 Data analysis MetaproteomeAnalyzer v2.12 120 min
B8 Data analysis MetaproteomeAnalyzer v1.0.5 180 min
META PROTEOME ANALYZER
single sample
12.8 hours
five samples
< 24 hours
single sample
> 3 days
five samples
> 3 days

FIGURE 1 | Comparison of new **(A1–A6)** and old workflow **(B1–B8)** for metaproteomics sample preparation and analysis. In addition, methods for quality control are shown **(X1–X3)**. The time shown represents the shortest possible time in which a single sample can be processed. Under reasonable circumstances five samples can be done in less than 24 h (or 15 samples within 48 h) using the new workflow limited by the number of available mass spectrometer. Similarly, at least 3 days are required for multiple samples using the old workflow.

descriptions and chemicals please consider the SOPs listed as **Supplementary Data Sheet S1**.

Improvements of the Laboratory Workflow

In order to reduce the time required for the phenol extraction protocol, dispensable washing steps with organic solvents were removed and incubation times were reduced. Protein purification by SDS-PAGE (Kohrs et al., 2014) and subsequent in-gel-digestion into peptides were the most time-consuming steps of old workflows, and were replaced with the FASP protocol (Wisniewski et al., 2009). The FASP protocol replaced these steps, allowing for direct digestion and simultaneous purification of the protein extract on the FASP filter (Wisniewski et al., 2009). In contrast to previous applications of the FASP protocol to environmental samples (Tanca et al., 2014; Brum et al., 2016), several steps of the FASP digestion were optimized. In particular, trypsin incubation time was reduced from the previous 12 h (overnight) to only 2 h (**Supplementary Table S1**). Furthermore, re-buffering of peptide extracts by time-consuming lyophilisation was omitted. Instead, extracts after FASP digestion were injected directly into the LC-MS/MS system after acidification.

Improvements of the MetaProteomeAnalyzer Software

An updated version of the MPA software was developed (see **Figures 2**, **3**). It not only improved the existing features but also added new functionalities (Muth et al., 2015a). The MPA offered a complete workflow from peak lists exported by the MS-software to protein database searching, and result analysis, visualization and export. A major feature of the MPA was the grouping of proteins into metaproteins based on shared peptides or sequence similarity. The provided manual (**Supplementary Table S2**) gives an in-depth description of the new version of the MPA software. Video tutorials, the download and other material are available on the MPA website[1]. All analyses for this manuscript were

[1] www.mpa.ovgu.de

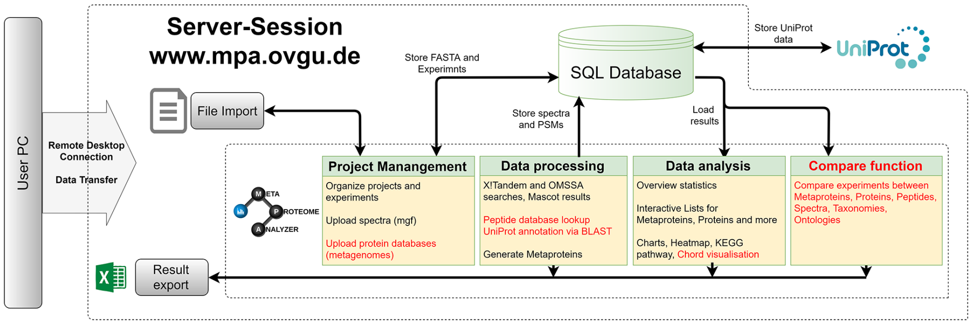
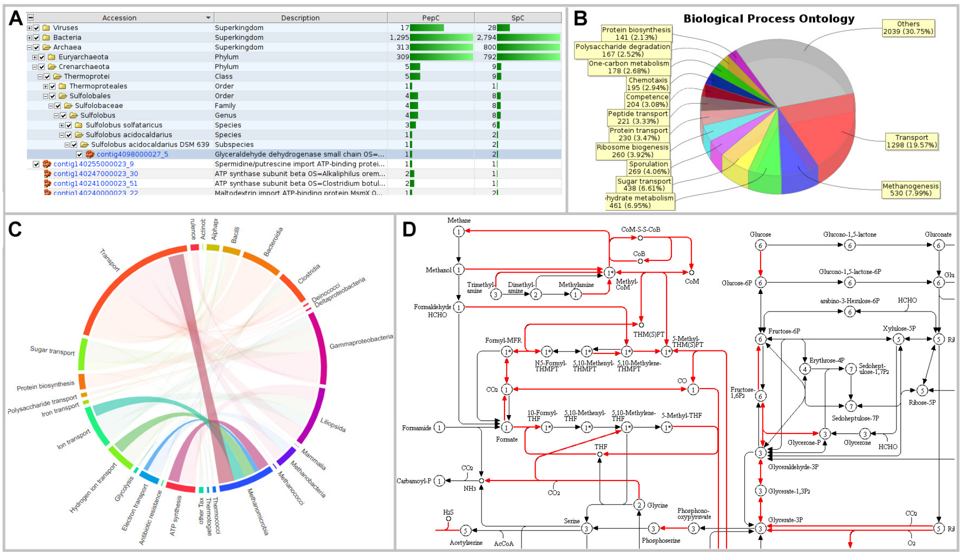

Server-Session
www.mpa.ovgu.de
User PC
Remote Desktop Connection
Data Transfer
File Import
Store FASTA and Experimnts
SQL Database
Store UniProt data
UniProt
Store spectra and PSMs
Load results
META PROTEOME ANALYZER
Project Manangement
Organize projects and experiments
Upload spectra (mgf)
Upload protein databases (metagenomes)
Data processing
X!Tandem and OMSSA searches, Mascot results
Peptide database lookup
UniProt annotation via BLAST
Generate Metaproteins
Data analysis
Overview statistics
Interactive Lists for Metaproteins, Proteins and more
Charts, Heatmap, KEGG pathway, Chord visualisation
Compare function
Compare experiments between Metaproteins, Proteins, Peptides, Spectra, Taxonomies, Ontologies
Result export

FIGURE 2 | MetaProteomeAnalyzer. Workflow of the MetaProteomeAnalyzer software including improvements and additions to the first MetaProteomeAnalyzer version (Muth et al., 2015a). Improvements were highlighted in red.



Accession	Description	PepC	SpC
Viruses	Superkingdom	17	28
Bacteria	Superkingdom	1,295	2,794
Archaea	Superkingdom	313	800
Euryarchaeota	Phylum	309	792
Crenarchaeota	Phylum	5	9
Thermoprotei	Class	5	9
Thermoproteales	Order	1	1
Sulfolobales	Order	4	8
Sulfolobaceae	Family	4	8
Sulfolobus	Genus	4	8
Sulfolobus solfataricus	Species	3	6
Sulfolobus acidocaldarius	Species	1	2
Sulfolobus acidocaldarius DSM 639	Subspecies	1	2
contig4098000027_5	Glyceraldehyde dehydrogenase small chain OS=...	1	2
contig140255000023_9	Spermidine/putrescine import ATP-binding protei...	1	1
contig140247000023_30	ATP synthase subunit beta OS=Alkaliphilus orem...	2	1
contig140241000023_51	ATP synthase subunit beta OS=Clostridium botul...	2	1
contig140240000023_22	Maltodextrin import ATP-binding protein MsmX O...	1	2

FIGURE 3 | Visualizations of MetaProteomeAnalyzer using data from NewWF_BGP_3_B. **(A)** The taxonomy results view of the protein tables hierarchically orders proteins by taxonomy allowing for easy selection and filtering of specific taxonomies. **(B)** Pie Chart with spectral counts of the biological process ontology of the Phylum *Euryarchaeota* selected through the taxonomy view. **(C)** Interactive chord diagram visualizing the relationship between taxonomy (rank = family) and functional ontology (UniProt keywords for Biological Process) (Zoun et al., 2017). Biological processes for *Methanosarcinaceae*, as an example, are highlighted. **(D)** KEGG pathway map for central carbon metabolisms (KEGG map 01200) highlighting enzymes identified with the MPA.

carried out with MPAv2.2.12. Meanwhile the version number was updated to number 3.0.0, which contains only minor changes.

Memory and speed constraints were reduced by improving the existing implementation of the algorithms and the database queries. Metagenome databases can now be uploaded by the user, providing a more user-friendly and efficient access. Further improvements include an update of internal parser routines, and the retrieval of protein meta-information using UniProtJAPI (Patient et al., 2008) for obtaining complete protein databases during upload. Additionally, the database search engines X!Tandem (Craig and Beavis, 2004) and OMSSA (Geer et al., 2004) were supplemented with a peptide database lookup. Furthermore, an integrated protein BLAST allowed the user to link unannotated protein sequences with UniProt metadata. The new MPA version now includes a sample comparison function that allows for a quantitative comparison of metaproteins, taxonomies, and protein ontologies across a multitude of samples. The newly implemented cord diagram tool visualizes taxonomy-function-relationships (Zoun et al., 2017).

Sampling

For this study, a total of nine samples were taken: three BGP samples (BGP_1–3), three human gut samples (Hgut_1–3), a soil sample, a compost sample and one WWTP sample. Samples were stored directly at −20°C. For subsequent phenol extraction, samples were defrosted and weighed. For processing of the WWTP sample, sludge flakes were centrifuged (10 min, 4°C, 10,000 g) before weighing and discarding of the supernatant.

Phenol Extraction (A1, B1)

For phenol extraction (**Supplementary Data Sheet S1**), 2 g sample, 5 g silica beads (0.5 mm), 2 mL 2 M sucrose solution, and

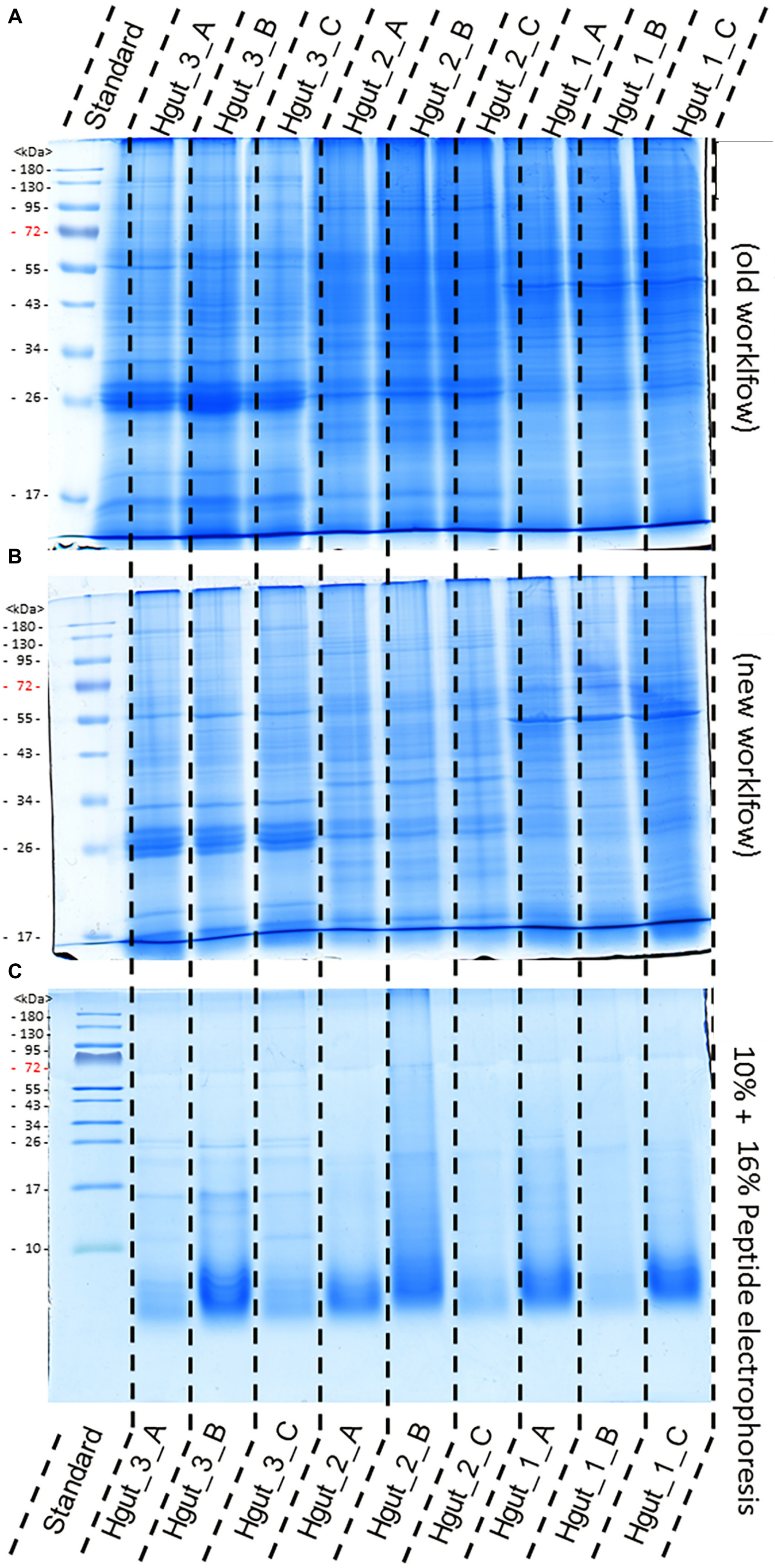


FIGURE 4 | Comparison of protein extraction of human gut samples of new and old workflow. For protein separation a 12% SDS-PAGE with 1 mm gel thickness was carried out and stained with colloidal coomassie. Proteins were extract by the old workflow **(A)** and new workflow **(B)**. Peptide electrophoresis **(C)** was carried out after FASP digest according to Schägger (2006) using a 10 and a 16% acrylamide gel. (STD) molecular weight standard; (Hgut 1–3) 100 µg of human fecal sample 1–3 resp. 90 µg for peptide electrophoresis Quality and purity of protein extracts was examined by SDS-PAGE (**Supplementary Presentation S1**).

3.5 mL phenol solution (10 g phenol dissolved in 1 mL ultrapure water) were added to a 15 mL reaction tube. Subsequently, the falcon was transferred into a ball mill (FastPrep-96, MP Biomedicals, Eschwege, Germany) and shaken for 5 min (A1) resp. 30 min (B1) at RT and 1,800 rpm. After centrifugation (10 min, RT, 10,000 *g*), the upper phenol phase was collected into a new 15 mL reaction tube and washed with the same volume of 1 M sucrose solution for 10 min at RT and 120 rpm on a shaker/ball mill. Finally, the sample was centrifuged again (10 min, RT, 10,000 g), and proteins were precipitated by ammonium acetate in methanol precipitation.

Ammonium Acetate in Methanol Precipitation (A2, B2)

Addition of the fourfold volume of ice-cold 100 mM ammonium acetate in methanol for 20 min (A2) resp. 60 min (B2) at −20°C precipitated proteins in the phenol phase. Afterward, the sample was centrifuged (10 min, 4°C, 10,000 g), and the supernatant was discarded. This precipitation step was repeated once.

Further Washing Steps (B3)

In order to remove remaining impurities, the precipitated protein pellet was washed four times with a threefold volume of ice-cold 80% acetone, 70% ethanol, 80% acetone, and 70% ethanol. Between the washing steps, the sample was incubated at −20°C, centrifuged (10 min, RT, 10,000 *g*) and the supernatant was discarded.

Dissolution of the Sample in Urea (A3, B4)

Finally, the protein pellet was dried at 60°C for 15 min and dissolved in 1 mL urea buffer (7 M urea, 2 M thiourea, 1% DTT). After 10 min shaking in a ball mill at (RT, 1,800 rpm), non-dissolved particles were removed by centrifugation (10 min, 4°C, 10,000 *g*). Protein extracts were stored at −20°C for later use.

Protein Quantification Using Amido Black Assay (X1)

For quantification of protein concentration (**Supplementary Data Sheet S1**) 50 μL of the sample were precipitated with 300 μL amido black staining solution. Afterward, the sample was centrifuged (5 min, RT, 16,400 *g*) and the supernatant was discarded. Two washing steps with 10% acetic acid in methanol and two centrifugation steps (5 min, RT, 16,400 *g*) removed unbound dye. Finally, the pellet was dissolved in 1 mL 0.1 M sodium hydroxide and absorption was measured at wavelength 615 nm using a photometer (Spectrophotometer Genesys 10S UV-Vis, Thermo Scientific, Waltham, United States).

SDS-PAGE (B5, X2)

For SDS-PAGE (**Supplementary Data Sheet S1**), 100 μg protein extract was diluted with the same volume of ultrapure water and precipitated by the same volume of ice-cold 100% acetone. After incubation at −20°C overnight, samples were centrifuged (30 min, 4°C, 16,400 *g*), the supernatant was discarded, and the pellet was dried. Subsequently, the protein pellet was dissolved in 20 μL SDS sample buffer, the sample was centrifuged (30 min, 4°C, 16,400 *g*), and the supernatant was loaded on the SDS-PAGE. In parallel to sample preparation, a 1 mm SDS-PAGE gel was prepared using a 12% separation and a 4% stacking gel. Subsequently, SDS-PAGE gels were inserted into the SDS-PAGE chamber (Mini-Protean Tetra System, BioRad, Hercules, United States), and the samples were loaded. Finally, 10 mA current was applied until proteins entered the separation gels, then 20 mA until the end of the gel. For subsequent in-gel digestion, the electrophoresis was stopped after the dye front entered into the separation gel for 5 mm. For visualization, proteins were incubated for 1 h in fixation solution (40% ethanol, 10% acetic acid) and then stained with Coomassie staining solution.

Peptide Electrophoresis (X3)

Peptide electrophoresis (**Supplementary Data Sheet S1**) was conducted in a standard electrophoresis chamber (Mini-Protean Tetra System, BioRad, Hercules, United States) (Schägger, 2006). In brief, 90 μg peptides were precipitated with acetone, diluted in 10 μL sample buffer, and incubated in a thermomixer for 60 min at 37°C and 1,400 rpm. Afterward, samples were centrifuged (10 min, 4°C, 16,400 *g*) and the supernatant was loaded on the gel. The gel comprised a 4% stacking gel as well as a 10% and a 16% separation gel. For separation, a voltage of 30 V was applied until the running front entered the 10% separation gel and increased subsequently to 90 V until it reached the end of the gel. Protein staining with Coomassie was carried out analogously to the staining of SDS-PAGEs, but the fixation solution contained methanol instead of ethanol.

FASP Digestion (A4)

For the FASP digestion (**Supplementary Data Sheet S1**), 100 μg protein extract in 200 μL urea buffer were loaded onto the FASP filter (Pall Nanosep 10K Omega, MWCO 10 kDa) and centrifuged (10–20 min, RT, 10,000). Note: Soil and human fecal samples required longer centrifugation times until all liquid passed through the FASP filter (about 20 min). Reduction and alkylation of proteins were carried out by addition of 100 μL DTT (20 min, 56°C, 300 rpm) and 100 μL IAA (20 min, RT, 300 rpm, in the dark). After each of these steps the liquid was removed by centrifugation (5 min, RT, 10,000 *g*) and the flow through was discarded. Subsequently, the proteins were washed once for 2 min with 100 μL 8 M urea, three times with 100 μL 50 mM ammonium bicarbonate, and centrifuged afterward (5 min, RT, 10,000 *g*). After removal of the flow through, trypsin was added onto the FASP filter (2 h, 37°C, 300 rpm) in an enzyme to protein ratio of approximately 1–100. Subsequently, the sample was centrifuged (5 min, RT, 10,000 *g*). Remaining peptides were rinsed through the filter by addition of 50 μL 50 mM ammonium bicarbonate and 50 μL ultrapure water (Millipore Q-POD Merck, Darmstadt, Germany) followed by another centrifugation step (5 min, RT, 10,000 *g*). Finally, 30 μL were acidified by addition of 3 μL 0.5% TFA, centrifuged (10 min, 4°C, 10,000 *g*), and transferred into an HPLC vial.

In-Gel Digestion (B6)

The single protein fraction after early stopping SDS-PAGE was cut into cubes of approx. 1 mm side length and transferred into a 2 mL reaction tube. For removal of the Coomassie dye, the gel cubes were incubated in 900 μL washing solution (50% methanol, 45% ultrapure water, 5% acetic acid) twice, once overnight and once the next day for 1 h in a shaker (RT, 150 rpm). After a further washing step with 900 μL acetonitrile (10 min, RT, 150 rpm), gel cubes were dried in a vacuum centrifuge (Digital Series SpeedVac SPD121P, Thermo Scientific, Waltham, United States). Reduction and alkylation of proteins were carried out by addition of 900 μL DTT (30 min, RT, 150 rpm) and 900 μL IAA (30 min, RT, 150 rpm, in the dark). After each of these steps, gel cubes were incubated in 900 μL acetonitrile (10 min, RT, 150 rpm). Subsequently, the gel cubes were washed with 50 mM ammonium bicarbonate (10 min, RT, 150 rpm) and acetonitrile (10 min, RT, 150 rpm). For tryptic digestion of proteins, 200 μL trypsin buffer (enzyme to substrate ratio: 1:100) was added over night (37°C, 150 rpm). The next day, the supernatant was collected into a new 2 mL reaction tube. Remaining peptides were washed out of the gel by incubation in extraction buffer 1 (90% ultrapure water, 10% formic acid; 30 min, RT, 150 rpm) and extraction buffer 2 (50% ultrapure water, 49% ACN, 1% TFA; 30 min, RT, 150 rpm). Both extracts were collected in a new reaction tube. Finally, the peptide solution was dried in the vacuum centrifuge and stored at −20°C. For LC-MS/MS measurements, dried peptides were dissolved in 300 μl solvent A (98% ultrapure water, 2% acetonitrile, 0.05% TFA), centrifuged (30 min, 4°C, 13,000 *g*) and transferred into a HPLC-vial.

LC-MS/MS Measurements (A5, B7)

Peptides were analyzed by LC-MS/MS using an UltiMate 3000 RSLCnano splitless liquid chromatography system coupled online to an Orbitrap Elite^{TM} Hybrid Ion Trap-Orbitrap MS/MS (MS) (both from Thermo Fisher Scientific, Bremen, Germany). After injection, peptides were loaded isocratically on a trap column (Dionex Acclaim, nano trap column, 100 μm i.d. × 2 cm, PepMap100 C18, 5 μm, 100 Å, nanoViper) with a flow rate of 7 μL/min chromatographic liquid phase A (98% ultrapure water, 2% acetonitrile, 0.05% TFA) for desalting and concentration.

Chromatographic separation was performed on a Dionex Acclaim PepMap C18 RSLC nano reversed phase column (2 μm particle size, 100 Å pore size, 75 μm inner diameter, and 250 mm length) at 40°C column temperature. A flow rate of 250 nL/min was applied using a binary A/B-solvent gradient (solvent A: 98% ultrapure water, 2% acetonitrile, 0.1% formic acid; solvent B: 80% acetonitrile, 10% ultrapure water, 10% trifluorethanol, 0.1% formic acid). 5 μl sample were injected. Separation started with 4% B for 5 min, continued with a linear increase to 55% B within 120 min, followed by a column wash with 90% B for 5 min, and re-equilibration with 4% B for 25 min. For mass spectrometry acquisition, a data-dependent MS/MS method was chosen. For the conducted measurements the MS was operated in positive ion mode and precursor ions were acquired in the orbital trap of the hybrid MS at a resolution of 30,000 and an *m/z* range of 350–2,000. Subsequently, fragment ion scans were produced in the linear ion trap of the hybrid MS with mass range and a scan rate at "normal" parameter settings for the top 20 most intense precursors selected for collision-induced dissociation.

Protein Identification Using the MPA (A7)

Orbitrap Elite^{TM} Hybrid Ion Trap-Orbitrap MS/MS measurements raw data files (raw file format) were processed by the Proteome Discoverer Software 1.4 (version 1.4.1.14, Thermo Fisher Scientific, Bremen, Germany), and converted into the Mascot Generic File format (mgf). Subsequently, mgf files were uploaded into the MPA software in the new version 2.12 and the release version 1.0.5 that was published previously (Muth et al., 2015a).

Three different types of software were used for peptide spectral matching: X!Tandem (Craig and Beavis, 2004), OMSSA (Geer et al., 2004) and MASCOT (version 2.5, Matrix Science, London, England) (Perkins et al., 1999). The MASCOT search was managed by the ProteinScape software (Bruker Daltonics, Bremen, Deutschland, (version 4.0.3 315) (Chamrad et al., 2007). All protein database searches used the following parameters: enzyme trypsin, one missed cleavage, monoisotopic mass, carbamidomethylation (cysteine) as fixed modification, oxidation (methionine) as variable modifications, ±10 ppm precursor and ± 0.5 Da MS/MS fragment tolerance, $1^{13}C$ and +2/+3 charged peptide ions. The Mascot search results (dat file format) were uploaded to the MPA software (only version 2.12). The MPA was designed to do the ensemble search (multiple search engines). Results were combined by uniquely identifying spectra and peptides throughout data processing. Therefore, spectra and peptides were not duplicated when multiple search engines reported the same match. In the rare case that two different peptides were found for a single spectrum both results were written into the database. This is not accurate with respect to spectral counting for quantification but kept as much information as possible.

Four protein databases – one for each sample type – were used for protein database searches (**Table 1**). These databases were created by combining UniProtKB/SwissProt (release November 2017) with an appropriate metagenome. Peptides found by X!Tandem and OMSSA searches were associated with all proteins containing them using a dedicated peptide database generated from the four protein databases prior to searches (peptide database lookup).

A false discovery rate (FDR) was applied at the PSM level. With the exception of soil and compost samples, an FDR of 1% was applied to all other samples. The old laboratory workflow did not report any proteins for soil and compost if the FDR was set to 1%. Therefore, the FDR of 5% was chosen for soil samples to allow for a fair comparison between the old and new workflows. In MPA version 2.12, identified proteins without taxonomic and functional classification were annotated with UniProtKB metadata by using protein BLAST [NCBI-Blast-version 2.6.0 (Altschul et al., 1990; Camacho et al., 2009)] against the UniProtKB/SwissProt database using an *e*-value cutoff of 10^{-4}. Subsequently, all protein BLAST proposals with the best identity were merged and used to annotate a protein.

TABLE 1 | Source and size of protein sequence databases.

Database	Protein sequences	Source/Reference	Used for samples
Biogas + SwissProt	2,349,714	Schluter et al., 2008; Rademacher et al., 2012; Hanreich et al., 2013; Stolze et al., 2016	BGP
Human Gut + SwissProt	6,159,039	Qin et al., 2010 https://www.ebi.ac.uk/metagenomics/studies/ERP000108	Hgut
Soil + SwissProt	684,487	JGI sequencing project; https://gold.jgi.doe.gov/study?id=Gs0085736	Soil compost
WWTP + SwissProt	2,243,839	Albertsen et al., 2012	WWTP
SwissProt	556,196	SwissProt downloaded in November 2017 www.uniprot.org	

Proteins were grouped into metaproteins using the shared peptide rule. The shared peptide rule adds a protein to the metaprotein if it has at least one distinct peptide in common with any other protein that belongs to this metaprotein. This did not require that all proteins of a metaprotein shared the same peptide. Metaproteins generated in this way were given a merged annotation. The taxonomy and UniRef Cluster of the metaprotein is set as the common ancestor of its proteins, while functional keywords and KEGG orthologies are compiled into non-redundant lists.

Several statistics for each sample were collected using the MPA software (**Supplementary Table S3**) and the metaproteins as well as metaprotein taxonomies were exported as comma separated value files (version 2.12 and version 1.0.5) (**Supplementary Table S4**). The sample comparison feature of MPA version 2.12 was used to generate metaproteins among all 54 samples and the resulting table was exported for later analysis. For quantification the spectral counts were taken. Finally, all MS data were submitted to PRIDE (Vizcaino et al., 2016) with the accession number PXD010550.

Biostatistics Evaluation

The data collected through the MPA software (**Supplementary Table S4**) were used to calculate the average number of identified spectra, peptides, proteins, and metaproteins. Metaproteins were split into known and unknown proteins depending on the existence of metadata beyond the protein sequence (i.e., taxonomy). The taxonomy distribution was calculated by counting the occurrence of specific taxonomies at all taxonomic ranks (**Supplementary Table S5**). The results of the comparison function were exported as a single csv file (**Supplementary Table S6**), and principle coordinate analysis (PCoA) was carried out using PAST3 (version 3.20).

RESULTS

The evaluation of the new workflow was divided into two steps: (i) improvements of the laboratory workflow and (ii) improvements of the bioinformatic workflow.

Improvements of the Laboratory Workflow

Validation of Protein Extraction

Phenol extraction from 2 g sample material resulted in between 0.55 and 10.94 mg protein per sample (**Supplementary Table S7**). To obtain sufficient protein for soil samples, pooling of seven extracts was required. Protein concentrations of previous and new sample preparations were similar (see **Supplementary Table S7**). Observed variation in protein amounts between sample types indicated that protein quantification of new samples should be performed to guarantee equal protein loading for FASP digestion and MS. For samples with limited availability, less raw material could be extracted because for protein quantification, FASP digestion and mass spectrometry, about 100 μg protein are required.

The old and the new sample preparation protocols resulted in a similar band pattern for every given sample, suggesting successful protein extraction in all cases (**Figure 4**). However, different intensities of the lanes indicated differences in the purity and quantity of the protein extracts. Protein extracts from human feces, WWTP and soil showed higher intensities than protein extracts from the BGP and compost (**Supplementary Presentation S1**). Peptide electrophoresis after FASP digestion yielded complete proteolysis of proteins and showed comparable intensities of peptides for most samples, indicating successful FASP digestion. Furthermore, performing peptide electrophoresis post-FASP digestion could enable researchers to identify problems that might occur during the digestion step. For example, the peptide electrophoresis of sample Hgut 3B showed protein bands at molecular weight of more than 10 kDa indicating incomplete digestion. The increase of the trypsin to protein ratio should be considered for samples of this type.

Validation of Protein Identification

Comparative LC-MS/MS measurements resulted in more identified spectra for the new extraction workflow (**Figure 5B**). For some soil samples extracted with the old workflow, no proteins with FDR 1% were identified. To allow comparison of search results of both workflows, an FDR of 5% was applied for all soil samples although this strategy is questionable regarding the correctness of identifications. The significant increase for BGP, Hgut and soil was related to a higher percentage of identified spectra from accumulated spectra indicating a higher quality of extraction of the new workflow (**Figure 5** and **Supplementary Table S8**). No significant increase was observed for WWTP. In addition, higher numbers of spectra were measured (**Figure 5A**). Probably, the FASP workflow was more efficient or removed more contaminants allowing the measurement of more and qualitatively better spectra. Numerous washing steps before digestion removed low molecular weight contaminants more efficiently. Furthermore, high molecular weight contaminants remained in the retentate while collecting

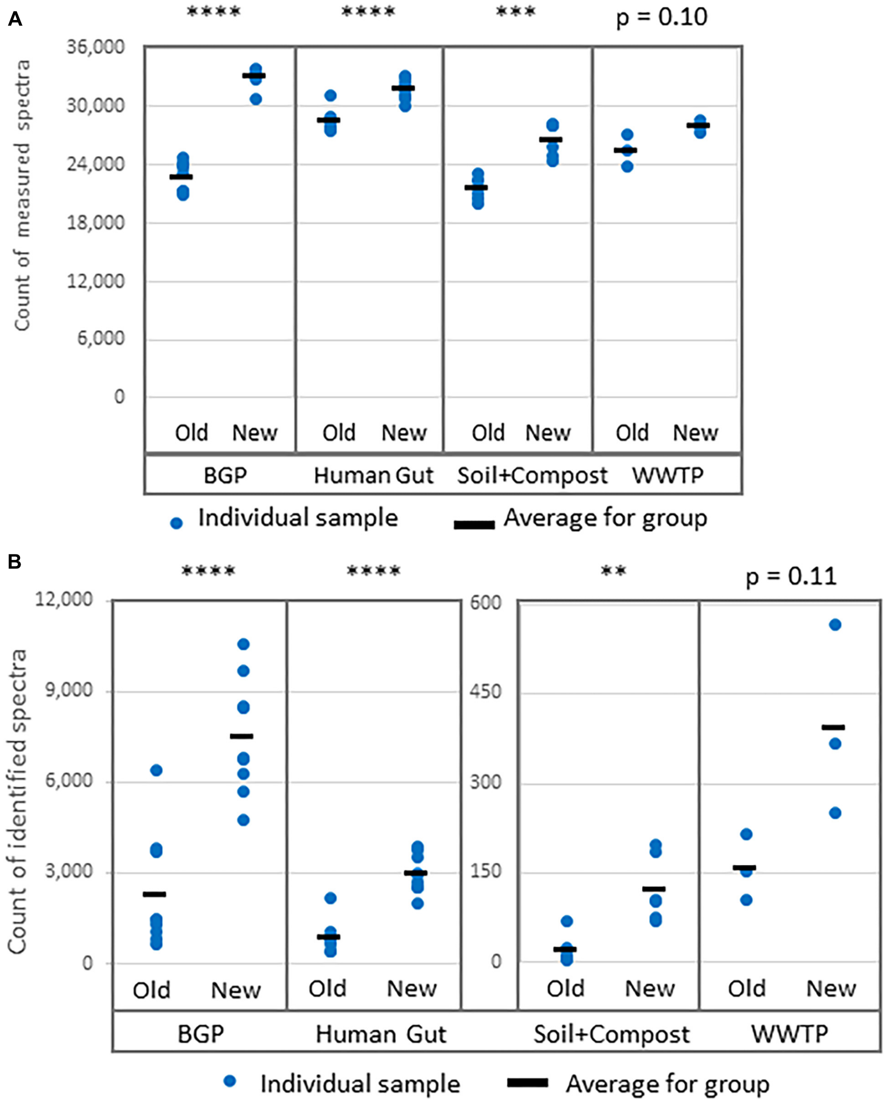
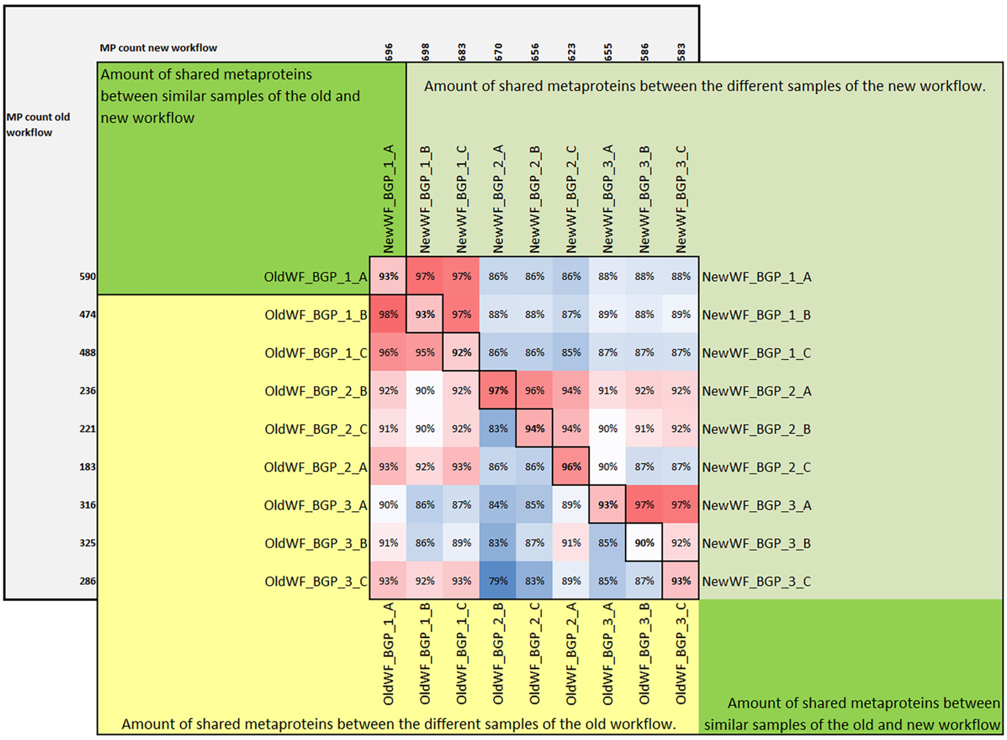


FIGURE 5 | Increase of **(A)** measured spectra and **(B)** identified spectra using the new workflow of sample preparation compared against the old workflow. The data was analyzed with MPA v2. The four types of samples from BGP, human gut, soil, and compost, and WWTP show significant differences regarding spectral counts for old and new workflow (*p*-values of *t*-test are shown in the figure). Similar results were obtained for identified peptides, percentage of identified spectra or identified metaproteins (**Supplementary Table S8**). *P*-values: *p = 0.05, **p = 0.01, ***p = 0.001, ****p = 0.0001.

the peptides in the filtrate. Skipping lyophilization after FASP and direct injection of acidified eluate had no negative impact on the number of identified spectra (**Supplementary Data Sheet S2** and **Supplementary Table S8**). Peptide and metaprotein counts followed the same trend as identified spectra. Furthermore, this increase in identifications was independent of the MPA version used (see **Supplementary Table S8**).

For qualitative evaluation of the new workflow, taxonomy and function were assigned to identified metaproteins of a BGP 1A to C (using the advanced feature of MPAv2.12). Although some function were detected with the old workflow only, the new workflow showed a higher coverage of metabolic pathways in KEGG map 1200 (**Figure 6** and **Supplementary Table S10**). The Krona plots of both samples showed minor differences in the taxonomy profile only (**Supplementary Table S10**). The abundances of orders varied about 1% between old and new workflow. Some minor orders were not shown either for the new or the old workflow due to limitations of this visualization. For further validation of the new laboratory workflow, pairwise Pearson correlation coefficients (**Supplementary Table S6**) based on the abundance of metaproteins and the percentage of identical metaproteins (**Figure 7**) for all pairs of samples and workflow were calculated. Both figures showed the same trends: (i) replicates of one sample were most similar (more than 90% identical metaproteins, Pearson coefficients higher than 0.9), (ii) different groups of samples were clearly separated (less than 70% identical metaproteins, Pearson coefficients lower than 0.7), (iii) identical samples prepared with the old and the new



MP count old workflow	MP count new workflow	696	698	683	670	656	623	655	586	583	
		NewWF_BGP_1_A	NewWF_BGP_1_B	NewWF_BGP_1_C	NewWF_BGP_2_A	NewWF_BGP_2_B	NewWF_BGP_2_C	NewWF_BGP_3_A	NewWF_BGP_3_B	NewWF_BGP_3_C	
590	OldWF_BGP_1_A	93%	97%	97%	86%	86%	86%	88%	88%	88%	NewWF_BGP_1_A
474	OldWF_BGP_1_B	98%	93%	97%	88%	88%	87%	89%	88%	89%	NewWF_BGP_1_B
488	OldWF_BGP_1_C	96%	95%	92%	86%	86%	85%	87%	87%	87%	NewWF_BGP_1_C
236	OldWF_BGP_2_B	92%	90%	92%	97%	96%	94%	91%	92%	92%	NewWF_BGP_2_A
221	OldWF_BGP_2_C	91%	90%	92%	83%	94%	94%	90%	91%	92%	NewWF_BGP_2_B
183	OldWF_BGP_2_A	93%	92%	93%	86%	86%	96%	90%	87%	87%	NewWF_BGP_2_C
316	OldWF_BGP_3_A	90%	86%	87%	84%	85%	89%	93%	97%	97%	NewWF_BGP_3_A
325	OldWF_BGP_3_B	91%	86%	89%	83%	87%	91%	85%	90%	92%	NewWF_BGP_3_B
286	OldWF_BGP_3_C	93%	92%	93%	79%	83%	89%	85%	87%	93%	NewWF_BGP_3_C
		OldWF_BGP_1_A	OldWF_BGP_1_B	OldWF_BGP_1_C	OldWF_BGP_2_B	OldWF_BGP_2_C	OldWF_BGP_2_A	OldWF_BGP_3_A	OldWF_BGP_3_B	OldWF_BGP_3_C	

FIGURE 6 | Amount of shared metaproteins between the old and new workflow. The upper triangular matrix shows the amount of shared metaproteins of the different BGP samples using the new workflow. The lower triangular matrix shows the amount of shared metaproteins of the different BGP samples using the old workflow. The diagonal shows the amount of shared metaproteins of the same sample analyzed by the old and the new workflow. For the calculation of the amount of shared metaproteins, the number of shared metaproteins was divided by the smaller number of metaproteins from both samples. For this analysis only metaproteins were considered which had in at least one sample a spectral count of 4. MP, metaprotein.

workflow showed also high similarity (more than 90% identical metaproteins, Pearson coefficients higher than 0.8), and (iv) sample groups with overall lower number of metaproteins (soil, WWTP) show heterogeneous results. These values are in the range of the observed reproducibility (70% identical proteins) of technically replicated LC-MS runs for protein identification (Tabb et al., 2010). For further validation of the reproducibility, spectral counts of identified metaproteins were compared between the two replicates of sample NewWF BGP_1. The scatterplot showed a good correlation between both replicates (**Figure 8** and **Supplementary Table S14**). No changes in abundances (more than twofold) were detected for metaprotein present with at least 10 spectral counts in one of the replicates. In contrast the comparison of the samples NewWF_BGP_1_A and NewWF_BGP_2_A showed 116 metaproteines (present with at least 10 spectral counts in one of the replicates) with more than twofold changes in abundance that could be related to differences in the microbial community of both samples.

Improvements of the Bioinformatic Workflow

BLAST of Metagenomes for Better Protein Annotation

The upgraded MPA integrates a convenient fully automated protein BLAST for user defined metagenomes. It gives the user the choice to use multiple BLAST hits and to combine them into a single entry, if they have the same *e*-value, sequence identity or bit score. A common entry uses the common ancestor taxonomy, chooses the common UniRef clusters and combines different ontologies, EC-numbers, KO-numbers between BLAST hits.

The protein databases used for protein identification consisted of UniProtKB/SwissProt combined with an appropriate metagenome for the four sample types (Schluter et al., 2008; Qin et al., 2010; Albertsen et al., 2012; Rademacher et al., 2012; Hanreich et al., 2013). MPAv1 did not support the integrated BLAST resulting in lower numbers of annotated proteins. For the BGP, and Hgut, the portion of annotated proteins was doubled applying the integrated BLAST of MPAv2 (**Figure 9** and **Supplementary Table S11**). For soil, and WWTP, the increase was not significant. The increase of annotated proteins was also reflected in the increase in the number of assigned KO numbers allowing better reconstruction of metabolic pathways or cellular functions. The low increase for soil and compost was related to the small size of soil metagenome supplementing UniProtKB/SwissProt.

Effect of Peptide Database Lookup for Metaprotein Generation

The new MPA version creates an index peptide database (since version 1.12) for uploaded protein databases (FASTA format).

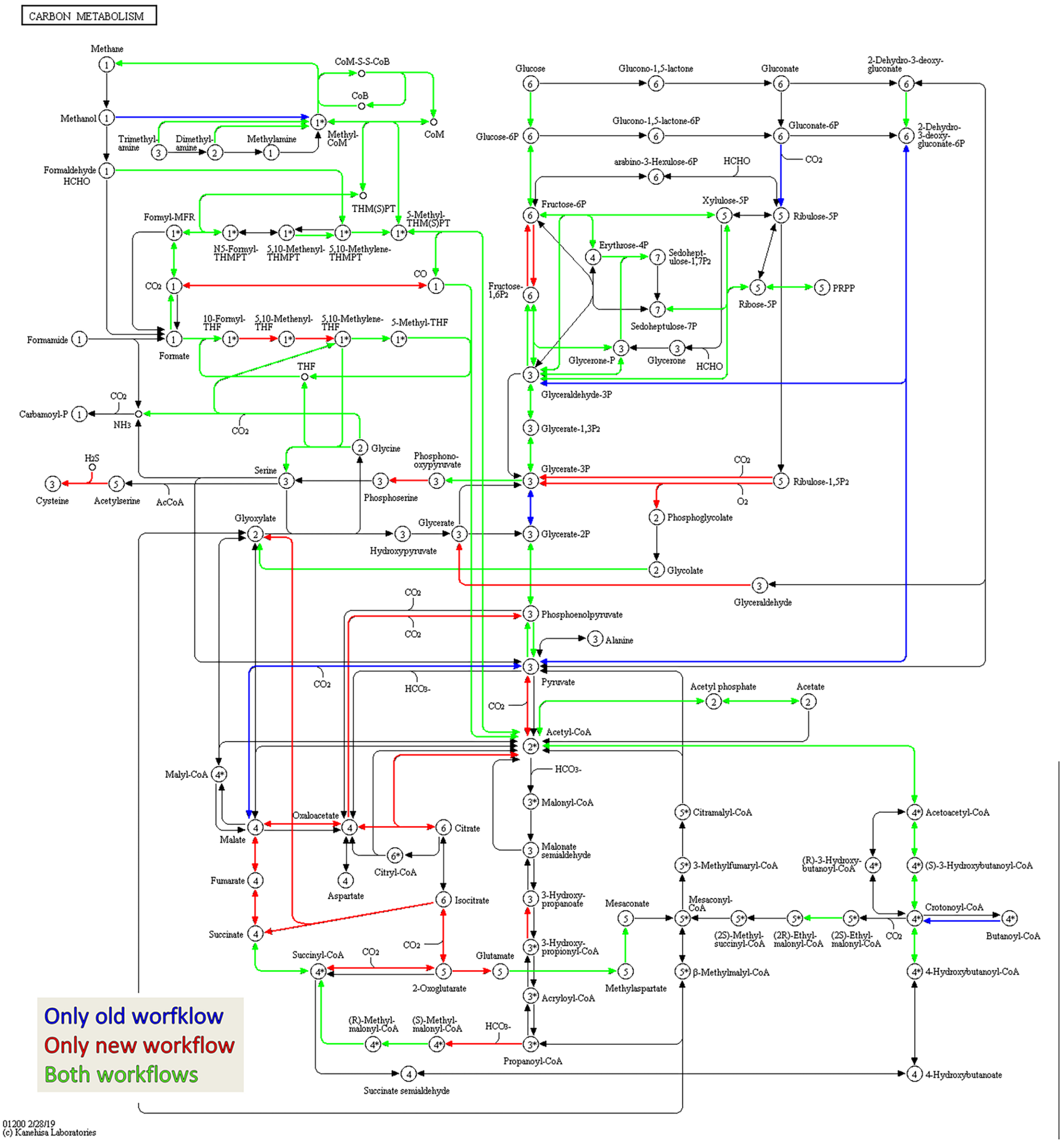


FIGURE 7 | Amount of shared metaproteins between the old and new workflow. KEGG map for the carbon metabolism showing enzymes in the sample BGP_1 (three technical replicates combined, analyzed with MPAv2.12). The map is colored to highlight differences between functional annotation, where blue are KO numbers exclusively found in the analysis with old workflow, red are KO numbers exclusively found in the analysis with the new workflow and green are KO numbers found with both. The maps are also hosted on: http://www.mpa.ovgu.de/review/kegg_carbonmetabolism_BGP_1.png.

After database searches are finished, a lookup in this peptide index collects all proteins that contain the identified peptides. This strategy works in conjunction with the metaprotein generation, which aims to accurately represent homologous proteins across multiple species.

The result of using the peptide database lookup in the new MPA version was an increase of reported proteins by a factor of up to 16, while the number of reported metaproteins remained approximately the same or slightly decreased (**Figure 10** and **Supplementary Tables S12, S13**). This was in line with expectations: since no new PSMs were added, the number of identified metaproteins should remained equal.

The integration of a peptide database lookup increased the ambiguity of metaprotein annotations, in particular the taxonomy. If more proteins were grouped together into a single metaprotein, the taxonomic specificity decreased applying shared peptides for metaprotein calculation and the lowest common ancestor for taxonomic assignment (Muth et al., 2015a; for further options regarding metaprotein generation see **Supplementary Table S2**). This negative effect was counteracted by increased number of protein annotations from BLAST

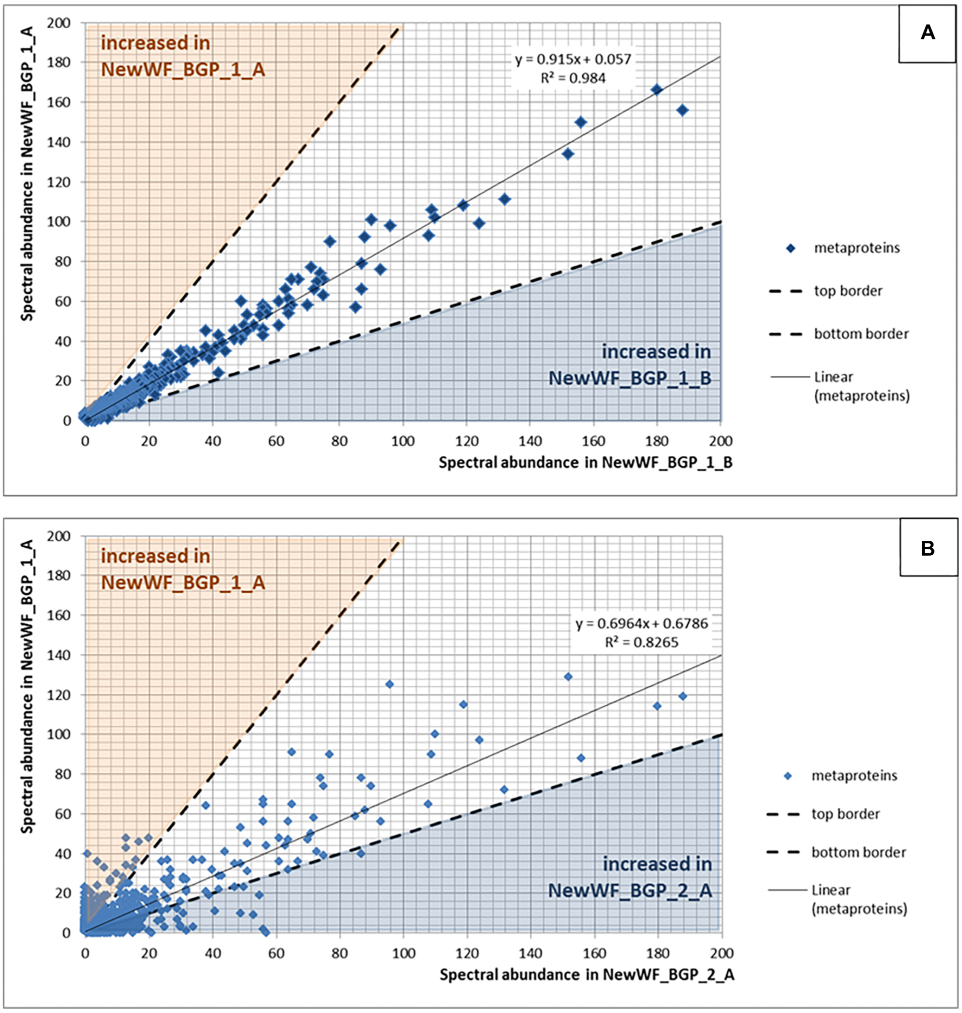


FIGURE 8 | (A) Reproducibility using replicated samples. The spectral counts of the metaproteins from the sample NewWF_BGP_1_A were plotted against the spectral counts of the metaproteins from the sample NewWF_BGP_1_B. The points in the blue or the orange area are at least doubled in the corresponding sample. **(B)** Differences between samples. The spectral counts of the metaproteins from the sample NewWF_BGP_1_A were plotted against the spectral counts of the metaproteins from the sample NewWF_BGP_2_A. The points in the blue or the orange area are at least decreased (blue) or increased (orange) twofold.

(**Figure 9**) providing taxonomic annotations of previously non-annotated metaproteins.

Compare Function for Fast Quantitative Analysis of Multiple Datasets

Another feature of the new MPA is the sample comparison function, which allows a quantitative comparison between metaproteins, peptides, taxonomies, and functional ontologies for large number of samples (highest number so far: 200). A comparison between multiple samples at the protein or peptide level is straightforward, since the protein accession or peptide sequence serve as unique identifiers. This is more complicated for metaproteins, taxonomies and functional ontologies, because these more abstract groupings are highly variable and dependent on the underlying data. For instance, using the shared peptide rule for metaprotein generation, a metaprotein will only be created if one peptide belongs to two proteins. If this shared peptide is absent in sample A, but present in sample B, sample A will contain two metaproteins and sample B will contain only one metaprotein, distorting a quantitative comparison. Therefore, the new sample comparison function of the MPA performs the metaprotein generation over any number of samples, enabling an accurate comparison of different experiments (for details regarding metaprotein generation see **Supplementary Table S2**).

To demonstrate its functionality, we compared all 54 samples on the metaprotein level using the spectral count of a metaprotein as quantitative measure. The comparison table of MPAv2 (**Supplementary Table S6**) was exported as a comma separated value file and used as direct input for a PCoA (**Figure 11**). A clear separation between the human fecal samples, the BGP samples and the soil, compost and WWTP samples was visible. The quality of grouping the technical replicates seemed to depend on the sample types. On the one hand, the observed scattering of replicates was related to the

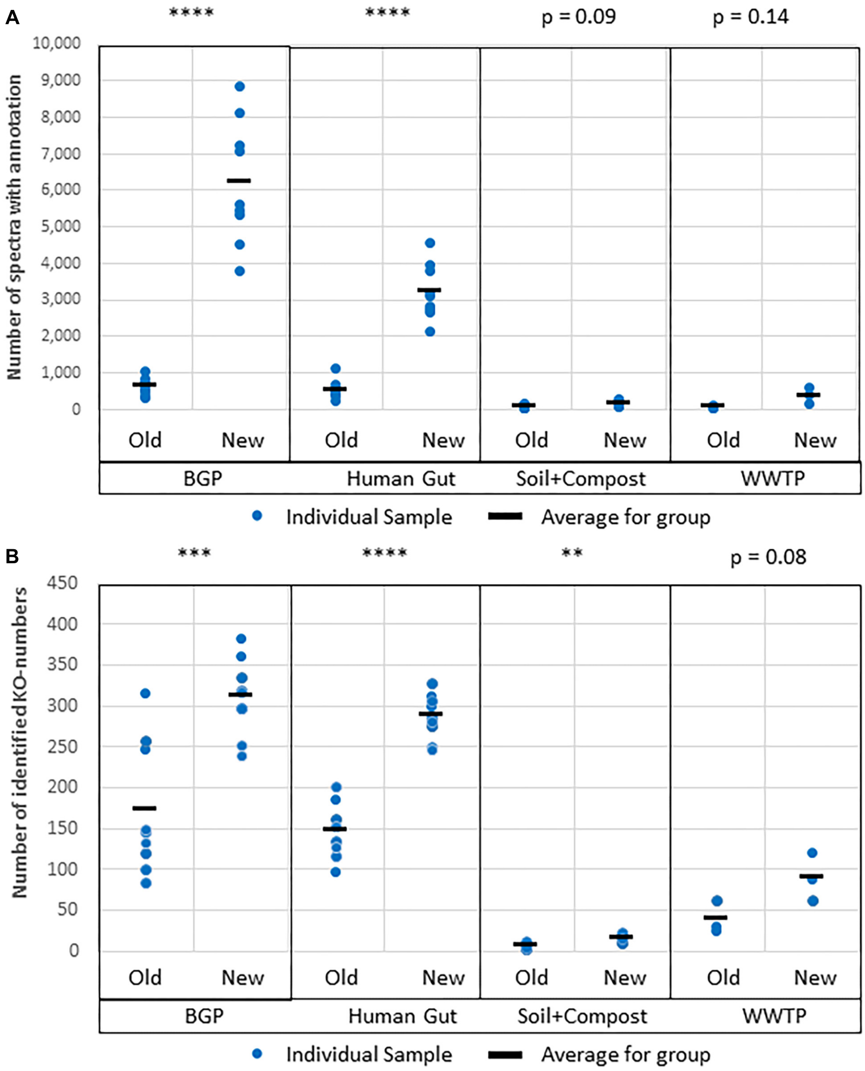


FIGURE 9 | Improved protein annotation via BLAST using MPAv2 in comparison to MPAv1. **(A)** Increase of annotated spectra. **(B)** Identified KO-numbers. Significance values calculated by Student's *t*-test for differences between the old and the new workflow are shown above the plots. The comparison was carried out with data obtained with the new laboratory workflow. The samples BGP, human gut, soil, and compost, and WWTP as well as their averages (black line) are shown separately. For further detail see **Supplementary Table S15**. *P*-values: **p* = 0.05, ***p* = 0.01, ****p* = 0.001, *****p* = 0.0001.

quality of data. WWTP and soil samples with low numbers of identifications showed a higher scattering than BGP and human gut samples. The higher scattering in PCoA was also related to higher distances in the clustering (**Figure 12**). On the other hand, the scattering of samples with high quality (human gut, BGP) visualized the error of replicates (low distances in the clustering).

Chord Diagrams for Visualization of the Relation Between Taxonomy and Function

One major question in microbiome research is how taxonomy is linked to function. Metaproteome data contains both levels of information. The previously published tool for connecting both levels into a single interactive figure (Zoun et al., 2017) is supported by a special export function of MPAv2 (**Figure 13**). The interactive figure can be adapted to the requirements by simply switching on and off certain taxonomies and functions allowing fast visualization of taxonomy-function-relationships according to user requirements (**Figure 13** and **Supplementary Table S10**). This new export supplemented other valuable visualizations available for MPA users internally (pie charts) and externally (KEGG maps, Krona plot).

DISCUSSION

In this study, we proposed and evaluated a new robust and fast workflow for metaproteomics of microbial community samples for routine application. The advantages over the previous workflow (Heyer et al., 2013; Muth et al., 2015a)

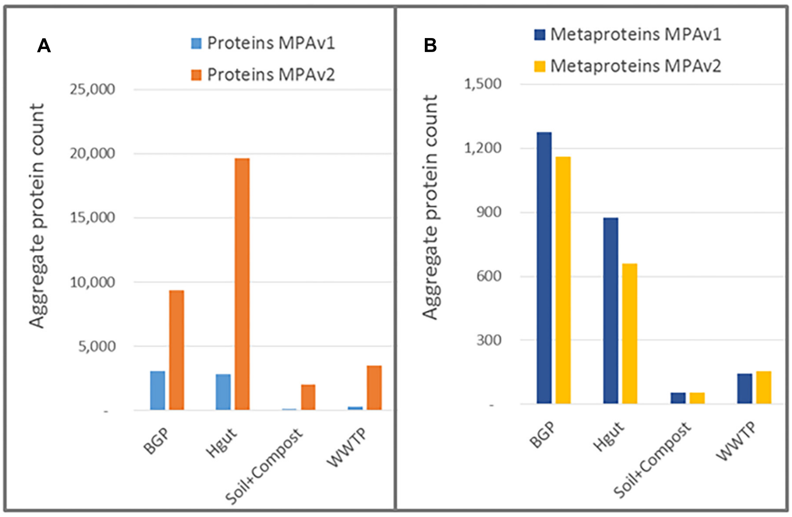


FIGURE 10 | Impact of peptide database lookup on reported proteins **(A)** and metaproteins **(B)** for MPAv1 and MPAv2.The comparison was carried out with data obtained with the new laboratory workflow. The bars represent the accumulated number of proteins/metaproteins for each sample group.

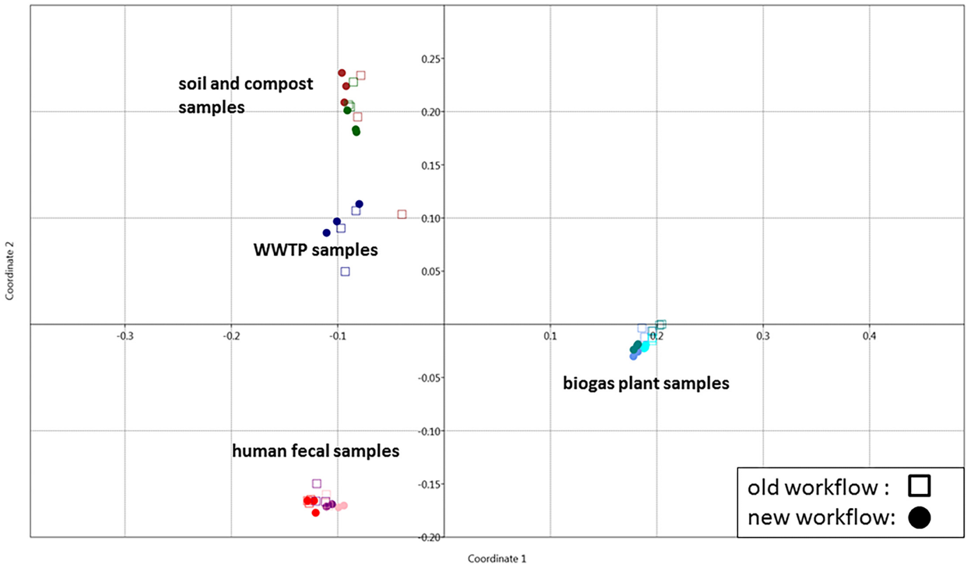


FIGURE 11 | Grouping of samples using PCoA. Principle coordinate analysis of all samples extracted with the previous (square) and the new (dots) workflow using the Past 3 tool and the Bray–Curtis distance as parameter. For analysis, all metaproteins that represented at least one percent of the identified spectra in at least one sample were considered. The samples comprise the three BGP samples 1–3 (aqua, cornflower blue, teal), the three human fecal samples 1–3 (light pink, purple, red), the WWTP samples (navy), the soil sample (brown) and the compost sample (dark green).

included performance improvements in both sample preparation and bioinformatics data processing. The objectives of our new protocols were speed, simplicity, high throughput, reproducibility, and robustness.

Advantages of the New Laboratory Workflow

The new laboratory workflow combined phenol extraction (Heyer et al., 2013), FASP (Wisniewski et al., 2009) and LC-MS/MS measurement (Link et al., 1999). Phenol extraction combined with cell lysis in a ball mill was previously applied to numerous environmental samples (Jia et al., 2017; Thorn et al., 2018; Heyer et al., 2019). For simplicity and robustness, the new workflow omitted sophisticated and time-consuming enrichment of biomass from environmental matrices by centrifugation or filtration (Xiong et al., 2015). Furthermore, fractionation, which was frequently applied in sample preparation (Hinzke et al., 2019), was sacrificed for speed of the final workflow. The final workflow enabled an investigation with a throughput of up to 5 samples in only 24 h, only limited by the throughput of the

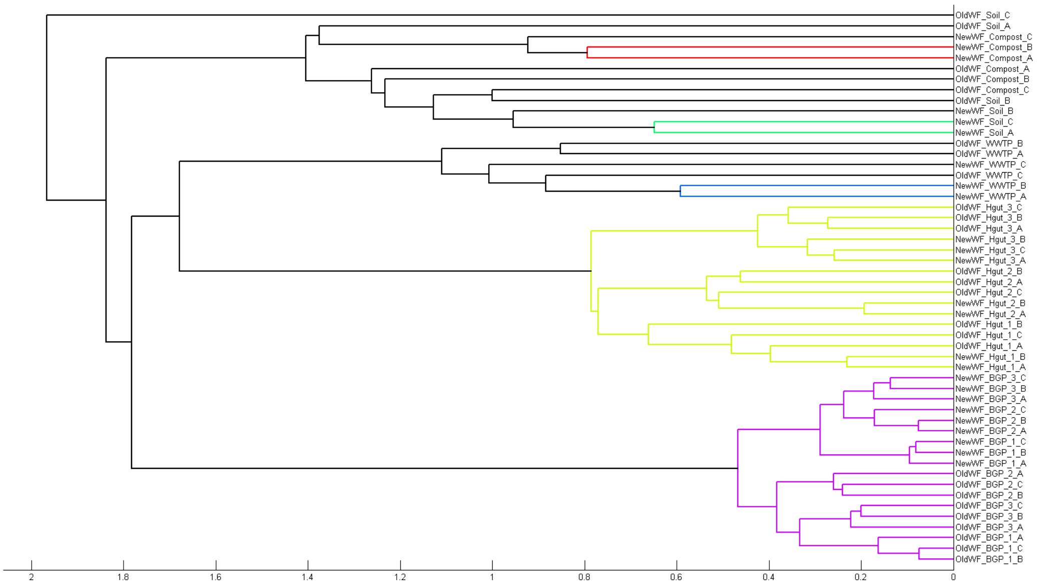


FIGURE 12 | Separation of samples in cluster tree. Cluster analysis of all samples extracted with the previous and the new workflow using Matlab and the "cityblock" distance and the "average" linkage as parameter was carried out. For analysis, all metaproteins that represented at least one percent of the identified spectra in at least one sample were considered. The samples comprise the three BGP samples, the three human fecal samples 1–3, the WWTP samples, the soil sample, and the compost sample.

MS. The throughput could be improved even further by parallel sample preparation in micro titer plates (Switzar et al., 2013), or the use of more mass spectrometers.

The evaluation of the new workflow confirmed that FASP digestion increased the number of identifications by at least a factor of two in comparison to the well-established in gel digestion protocol (Shevchenko et al., 2006). The most probable explanation for this large difference was a decreased efficiency of trypsin in in-gel digestion, because proteins inside the gel matrix were partially inaccessible to trypsin or the recovery of peptides from the gel matrix was poor (Leon et al., 2013). Furthermore, FASP was considered to remove contaminations: (i) low molecular weight contaminations were removed by filtrations before digestion and (ii) high molecular weight contaminations remained in the retentate after digestion. However, the number of identifications was heavily dependent on the sample type. First, a literature comparison (**Supplementary Table S9**) confirmed that soil metaproteome studies (Keiblinger et al., 2012; Bagnoud et al., 2016; Bastida et al., 2016; Thorn et al., 2018) identified less proteins and peptides than studies of Hgut (Tanca et al., 2016; Brown et al., 2018; Zhang et al., 2018a; Rechenberger et al., 2019) and BGP (Bize et al., 2015; Hagen et al., 2017; Joyce et al., 2018). Second, it became obvious that sacrificing the fractionation before or after (Hinzke et al., 2019) tryptic digestion resulted in lower number of identifications. Considering the speed for measuring the samples without fractionation, the number of identified proteins was still competitive in most cases, for BGPs even better. Despite the increased efficiency achieved with the new FASP protocol, the number of identifications was still influenced strongly by the sample type. Poor protein abundance could be overcome by collecting higher sample volumes and pooling of multiple extracts of the same sample. When a higher metaproteome coverage is required to derive meaningful results for more scientific projects, supplementary fractionation techniques such as isoelectric focusing (Kohrs et al., 2014) or ion exchange chromatography (Erickson et al., 2012; Kleiner et al., 2017) could be applied. However, these solutions would come at the expense of throughput. Since low protein abundance and poor extraction from sample matrices might occur with any new sample, the recommended strategy for new samples is to control the quality of extraction and digestion using SDS-PAGE and peptide electrophoresis beforehand.

The reproducibility of the workflow was demonstrated by high numbers of identical metaproteins and high Pearson correlation coefficients for replicated samples or for sample types. Considering the number of identical metaproteins, the reproducibility cannot exceed the limits of replicated LC-MS/MS measurements for protein identification (Tabb et al., 2010). High reproducibility was confirmed further by similar spectral counts for identified metaproteins of two technical replicates of a BGP sample, whereas the quantitative comparison of two different BGP samples revealed numerous metaproteins with different abundance.

Robustness of the workflow was related to repeated assignment of replicates to each other using statistical data analysis. Grouping of replicates and separation of different sample types was observed by PCoA and clustering. Therefore, single replicates appeared to be sufficient for future studies. The specificity of the workflow should enable the separation of different samples as shown for BGP and Hgut (different patients). For soil and WWTP, reproducibility and robustness were lower due to low numbers of identified metaproteins. These results

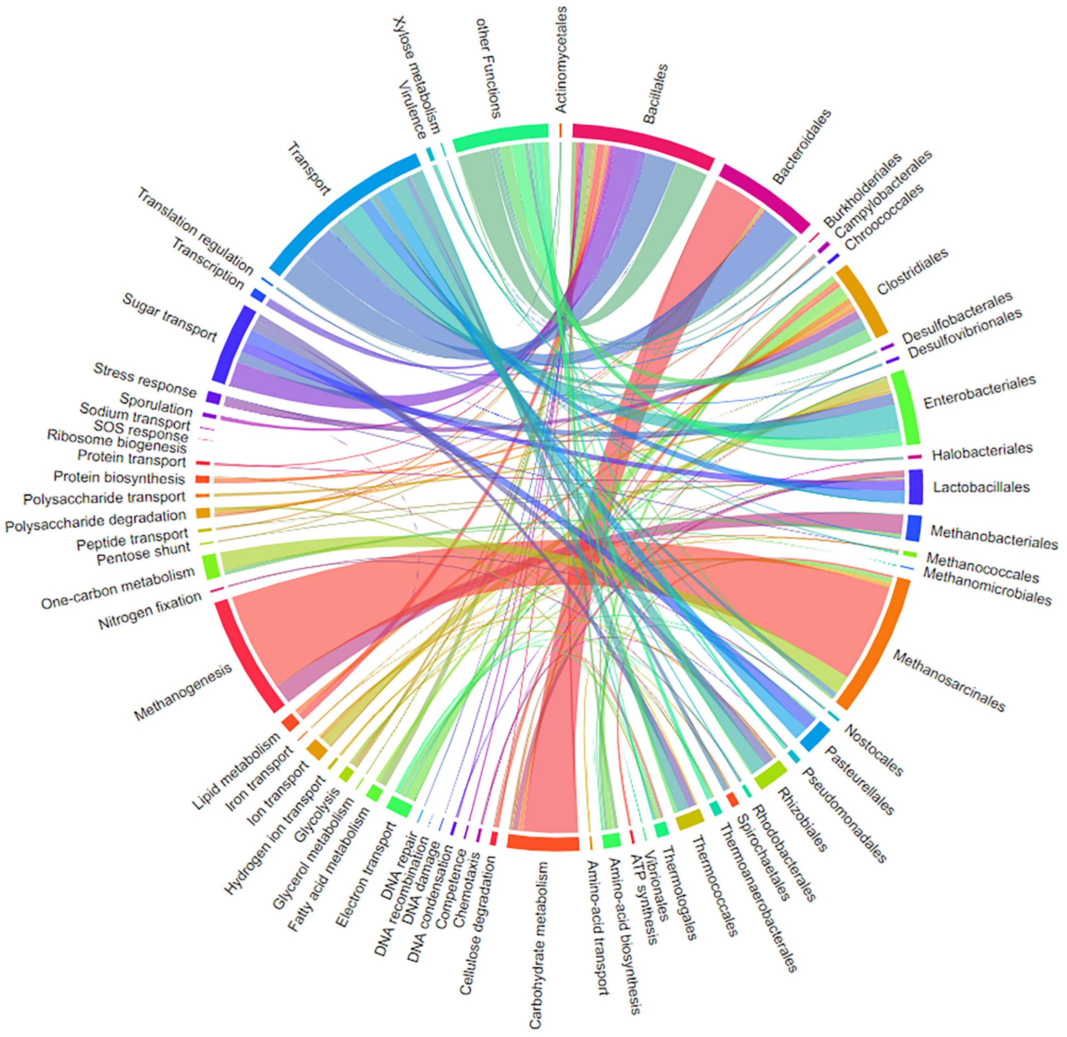


FIGURE 13 | Chord-diagram visualizing the taxonomy-function-relationships for samples BGP 1A–C. Data was exported from MPA. All taxonomies except bacterial and archaeal orders were removed in the diagram (chord diagram for a Hgut sample is found in **Supplementary Table S10**).

indicated that at least several hundred metaproteins are required for statistical data analysis.

Advantages of the New MPA

Another focus of this study was the improvement of the bioinformatics workflow by further development of the MPA software. Several tools for metaproteomics are available and provide valuable problem-specific solutions (e.g., Prophane, iMetaLab 1.0, UniPept) (Schneider et al., 2011; Cheng et al., 2017; Mesuere et al., 2018). None of these tools, however, offers the user a full workflow beginning with MS data and ending with protein reports and visualizations. Major advantages of the previous MPA were the dynamic metaprotein generation and the flexibility in taxonomic as well as functional filtering.

In contrast to the recently published MPA Portable (Muth et al., 2018), which fits well into a research context, where data science experts and computing resources are more easily available, the MPA 2.12 enables users with little or no background in computer science to conduct metaproteomics experiments with ease. While both options – local deployment or central solution – are available to users, central solutions (Cheng et al., 2017; Afgan et al., 2018; Liao et al., 2018) can keep up with the ever increasing data generated by high-throughput MS and the associated computational demands for broad application in routine analyses.

The newly implemented peptide database lookup and the integrated protein BLAST doubled the number of metaproteins annotated on the taxonomic and functional level. Together with the previously implemented metaprotein generation, the MPA now provides a unique workflow of functions that are available separately by other tools, e.g., Unipept or Prophane. The unique workflow within a single software speeds up the data analysis by omitting the file-based transfer of data between different tools. For further improvement, binned metagenomes containing taxonomic and functional data of high quality (Junemann et al., 2017) could be used. Assignment of metaproteins to genome bins would allow a more specific reconstruction of metabolic pathways based on additional information from the context of the genome bin. Furthermore, the concatenation of metagenomes from a similar sample and UniProtKB/SwissProt could improve the identification rate even more (Heyer et al., 2016). In addition, metapeptide databases based on raw metagenomes have been shown to increase protein identification too (May et al., 2016). The issue of correct selection of databases requires attention of users but is discussed elsewhere (Muth et al., 2015b; Timmins-Schiffman et al., 2017; Schiebenhoefer et al., 2019).

Building on these strengths, the new quantitative comparison function provides an overall metaprotein generation unifying single datasets for final export into other software. The exported CSV-files allowed a fast subsequent analysis of multiple sample data with Excel, MatLab, Past3 or R. The simple and fast combination of multiple datasets by MPA is a precondition for quantitative and statistical analysis of data from high-throughput-studies. It needs to be mentioned that due to the application of multiple search engines more than one peptide could be assigned to a spectrum. Due to high mass accuracy of precursor spectra with orbitrap instruments this ambiguity is a very rare event. Therefore, it was decided to keep both results when developing the first version of MPA. The minor risk of failures in counting should be considered for diagnostic applications. We strongly suggest the validation of potential markers peptides and quantification based on multiple peptides.

In addition, the chord diagram is a smart interactive tool visualizing the relation between taxonomy and functions that could be used for primary exploration of data or for preparing interactive visualization of data for publications.

Steps Toward the Application of Metaproteomics in Applied Research and Diagnostics

The new metaproteomic workflow was substantially improved regarding speed, throughput and simplicity. Reproducibility, and robustness were shown by statistical analysis of the provided data. In contrast to these strengths, its resolution was limited due to sacrificing additional fractionation steps in sample preparation. However, it could be easily upgraded for fundamental science by adding fractionation on the peptide level (e.g., MudPIT; Schirmer et al., 2003), at the expense of speed. Next steps for its application in applied research and diagnostics are: (i) validation using more samples, (ii) further exploration of its strengths and limitations, and (iii) approval of its sensitivity and specificity in real projects from researchers in biotechnology and medicine.

Related to the exploration of strengths and limitations, the depth of data required for valuable data analysis needs to be considered. Instead of deep exploration of microbiomes by achieving as many identifications as possible, proteotyping of microbial communities (Heyer et al., 2016; Kohrs et al., 2017) aims to detect single marker proteins or process (disease) specific protein signatures. It is questionable, whether metaproteins are the preferred level of data. Metaproteins contain a high level of information (taxonomy and function), but merging peptides of multiple proteins could hinder correlations with patient/process data. Therefore, single peptides should also be correlated to the state of the samples. Based on such results, multiple reaction monitoring (Yao et al., 2013) could be applied as a more specific and more quantitative approach for diagnostic applications. Furthermore, the specificity of selected marker peptides needs to be crosschecked by bioinformatic analysis (e.g., the tryptic peptide analysis of Unipept 4.0; Mesuere et al., 2018)[2]. However, Unipept is based on UniProt database and does probably not contain all peptides detected in the samples.

The main dilemma is that further development and validation of the workflow for diagnosis requires its extensive application producing comprehensive datasets for subsequent correlation to patient/process data, but in comparison to conventional diagnostic tools the effort still appears to be very high at this stage. The samples analyzed in this paper exemplify potential applications. In order to justify further comprehensive studies, selected results are discussed referring to recent literature. Omitting extensive sample preparation enabled also the detection of "contaminating" non-microbial proteins from host (Lehmann et al., 2019) or from feed (Heyer et al., 2015) that could be valuable for understanding disease or technical processes. For instance, the disease marker calprotectin is commonly monitored in stool samples through ELISA to discriminate between inflammatory bowel syndrome and inflammatory bowel disease (Caccaro et al., 2012). Calprotectin was easily found using our metaproteomics workflow alongside many other potential disease markers of human and microbial origin (**Supplementary Table S6**; Lehmann et al., 2019). Whereas ELISA is restricted to a single protein and relies on antibodies that may bind unspecifically, metaproteomics can detect a multitude of protein alterations for disease specific pattern recognition and thus enable a more comprehensive and robust diagnosis. This will be particularly useful if the impact of the microbiome on certain diseases such as diabetes, several autoimmune diseases, obesity and depression is better understood and microbial marker proteins for these diseases are known. For BGP, the supporting effect of annotating hits from non-annotated metagenome data by BLAST was obvious. Key enzymes for all major pathways of anaerobic digestion were detected. The abundance of methyl-coenzyme M reductase has been identified previously as a predictive biomarker for performance of BGP (Munk et al., 2012). Whereas the suggested RT-PCR assay focussed only on a single function, metaproteome data provides additional data that discriminated between the acetoclastic and hydrogentrophic pathways of methanogenesis (Heyer et al., 2016, 2019).

CONCLUSION

In conclusion, the new metaproteomics workflow presented in this study combines robust and fast sample preparation with improved data processing in a single standardized workflow. The evaluation of the workflow showed a significant increase

[2] https://unipept.ugent.be/

in quality and quantity of generated results compared to our previously reported workflows. Performance and processing time provide a basis for establishing metaproteome based diagnostics in clinical settings and routine analysis of technical and environmental samples in the future. Further steps to explore the potential of the workflow are necessary and should be a major focus of future research.

DATA AVAILABILITY

The raw data and the FASTA database are available for download from PRIDE (PXD010550) (Vizcaino et al., 2016).

ETHICS STATEMENT

Fecal samples were collected from three healthy, omnivorous male subjects (A, B, and C) in the age-range of 30–33 as part of the proof-of-principle study. The study was approved by the ethical committee of the Otto von Guericke University Magdeburg (Number 99/10). All healthy volunteers provided written informed consent. The samples were stored at −20°C.

AUTHOR CONTRIBUTIONS

The improvement of the laboratory workflow was carried out by AnB, CS, and RH. The further development of the MPA was done by KS, RZ, RH, TM, and SD. The manuscript was written by RH, DB, KS, and UR. FK and SP tested the software and provided user feedback for development. TM, SD, AlB, and GS contributed with the valuable advice and by editing the manuscript. All authors read and approved the final manuscript.

FUNDING

RH was supported by a grant of the Federal Ministry of Food, Agriculture and Consumer Protection (BMELV) communicated by the Agency for Renewable Resources (FNR), Grant No. 22404115 ("Biogas Measurement Program III"). KS and RZ were funded by the de.NBI network ("MetaProtServ de-NBI-039").

ACKNOWLEDGMENTS

The authors thank the Information Technology Group of the Max Planck Institute for Dynamics of Complex Technical Systems Magdeburg, in particular, Dr. Gerrit Danker.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01883/full#supplementary-material

TABLE S1 | Digestion time.

TABLE S2 | Manual MPA.

TABLE S3 | Sample metadata.

TABLE S4 | Metaprotein lists.

TABLE S5 | Taxonomic annotation.

TABLE S6 | Comparison table and reproducibility.

TABLE S7 | Protein quantification.

TABLE S8 | Identifications.

TABLE S9 | Literature comparison.

TABLE S10 | Additional visualizations.

TABLE S11 | Effect of BLAST.

TABLE S12 | PeptideDB lookup.

TABLE S13 | Metaprotein annotations.

TABLE S14 | Calculation of scatterplots.

TABLE S15 | Number of identified KOs and ECs.

DATA SHEET S1 | Collection SOPs.

DATA SHEET S2 | Chromatograms.

PRESENTATION S1 | Quality control gels.

REFERENCES

Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Cech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544. doi: 10.1093/nar/gky379

Albertsen, M., Hansen, L. B. S., Saunders, A. M., Nielsen, P. H., and Nielsen, K. L. (2012). A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. *ISME J.* 6, 1094–1106. doi: 10.1038/ismej.2011.176

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1006/jmbi.1990.9999

Bagnoud, A., Chourey, K., Hettich, R. L., De Bruijn, I., Andersson, A. F., Leupin, O. X., et al. (2016). Reconstructing a hydrogen-driven microbial metabolic network in Opalinus Clay rock. *Nat. Commun.* 7:12770. doi: 10.1038/ncomms12770

Bastida, F., and Jehmlich, N. (2016). It's all about functionality: how can metaproteomics help us to discuss the attributes of ecological relevance in soil? *J. Proteom.* 144, 159–161. doi: 10.1016/j.jprot.2016.06.002

Bastida, F., Jehmlich, N., Lima, K., Morris, B. E. L., Richnow, H. H., Hernandez, T., et al. (2016). The ecological and physiological responses of the microbial community from a semiarid soil to hydrocarbon contamination and its bioremediation using compost amendment. *J. Proteom.* 135, 162–169. doi: 10.1016/j.jprot.2015.07.023

Benndorf, D., Balcke, G. U., Harms, H., and Von Bergen, M. (2007). Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *ISME J.* 1, 224–234. doi: 10.1038/ismej.2007.39

Benndorf, D., Vogt, C., Jehmlich, N., Schmidt, Y., Thomas, H., Woffendin, G., et al. (2009). Improving protein extraction and separation methods for investigating the metaproteome of anaerobic benzene communities within sediments. *Biodegradation* 20, 737–750. doi: 10.1007/s10532-009-9261-3

Bize, A., Cardona, L., Desmond-Le Quemener, E., Battimelli, A., Badalato, N., Bureau, C., et al. (2015). Shotgun metaproteomic profiling of biomimetic anaerobic digestion processes treating sewage sludge. *Proteomics* 15, 3532–3543. doi: 10.1002/pmic.201500041

Brown, C. T., Xiong, W., Olm, M. R., Thomas, B. C., Baker, R., Firek, B., et al. (2018). Hospitalized premature infants are colonized by related bacterial strains with distinct proteomic profiles*. *Mbio* 9.

Brum, J. R., Ignacio-Espinoza, J. C., Kim, E. H., Trubl, G., Jones, R. M., Roux, S., et al. (2016). Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc. Natil. Aca. Sci. U.S.Am.* 113, 2436–2441. doi: 10.1073/pnas.1525139113

Caccaro, R., D'inca, R., Pathak, S., and Sturniolo, G. C. (2012). Clinical utility of calprotectin and lactoferrin in patients with inflammatory bowel disease: is there something new from the literature? *Exp. Rev Clin Immunol* 8, 579–585. doi: 10.1586/eci.12.50

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Chamrad, D. C., Blueggel, M., Koerting, G., Glandorf, J., Vagts, J., Hufnagel, P., et al. (2007). P5-M Proteinscape—Software Platform for Managing Proteomics Data. *J. Biomol. Tech.* 18, 2–3.

Cheng, K., Ning, Z. B., Zhang, X., Li, L. Y., Liao, B., Mayne, J., et al. (2017). MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* 5, 157.

Chourey, K., Jansson, J., Verberkmoes, N., Shah, M., Chavarria, K. L., Tom, L. M., et al. (2010). Direct cellular lysis/protein extraction protocol for soil metaproteomics. *J. Proteom. Res.* 9, 6615–6622. doi: 10.1021/pr100787q

Colatriano, D., and Walsh, D. A. (2015). An aquatic microbial metaproteomics workflow: from cells to tryptic peptides suitable for tandem mass spectrometry-based analysis. *J. Vis. Exp.* 103:52827. doi: 10.3791/52827

Craig, R., and Beavis, R. C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20, 1466–1467. doi: 10.1093/bioinformatics/bth092

Deusch, S., Camarinha-Silva, A., Conrad, J., Beifuss, U., Rodehutscord, M., and Seifert, J. (2017). A structural and functional elucidation of the rumen microbiome influenced by various diets and microenvironments. *Front. Microbiol.* 8:1605. doi: 10.3389/fmicb.2017.01605

Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., and Pan, C. L. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of crohn's disease. *Plos One* 7:e49138. doi: 10.1371/journal.pone.0049138

Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., et al. (2004). Open mass spectrometry search algorithm. *J. Proteom Res.* 3, 958–964.

Hagen, L. H., Frank, J. A., Zamanzadeh, M., Eijsink, V. G. H., Pope, P. B., Horn, S. J., et al. (2017). Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. *Appl. Environ. Microbiol.* 83:e1955–16.

Hanreich, A., Schimpf, U., Zakrzewski, M., Schluter, A., Benndorf, D., Heyer, R., et al. (2013). Metagenome and metaproteome analyses of microbial communities in mesophilic biogas-producing anaerobic batch fermentations indicate concerted plant carbohydrate degradation. *Syst. Appl. Microbiol.* 36, 330–338. doi: 10.1016/j.syapm.2013.03.006

Heyer, R., Benndorf, D., Kohrs, F., De Vrieze, J., Boon, N., Hoffmann, M., et al. (2016). Proteotyping of biogas plant microbiomes separates biogas plants according to process temperature and reactor type. *Biotechnol. Biofuels.* 9:155. doi: 10.1186/s13068-016-0572-4

Heyer, R., Kohrs, F., Benndorf, D., Rapp, E., Kausmann, R., Heiermann, M., et al. (2013). Metaproteome analysis of the microbial communities in agricultural biogas plants. *N. Biotechnol.* 30, 614–622. doi: 10.1016/j.nbt.2013.01.002

Heyer, R., Kohrs, F., Reichl, U., and Benndorf, D. (2015). Metaproteomics of complex microbial communities in biogas plants. *Microb. Biotechnol.* 8, 749–763. doi: 10.1111/1751-7915.12276

Heyer, R., Schallert, K., Siewert, C., Kohrs, F., Greve, J., Maus, I., et al. (2019). Metaproteome analysis reveals that syntrophy, competition, and phage-host interaction shape microbial communities in biogas plants. *Microbiome* 7:69. doi: 10.1186/s40168-019-0673-y

Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* 261, 24–36. doi: 10.1016/j.jbiotec.2017.06.1201

Hinzke, T., Kouris, A., Hughes, R.-A., Strous, M., and Kleiner, M. (2019). More Is not always better: evaluation of 1D and 2D-LC-MS/MS methods for metaproteomics. *Front. Microbiol.* 10:238. doi: 10.3389/fmicb.2019.00238

Jia, X., Xi, B. D., Li, M. X., Yang, Y., and Wang, Y. (2017). Metaproteomics analysis of the functional insights into microbial communities of combined hydrogen and methane production by anaerobic fermentation from reed straw. *Plos One* 12:e0183158. doi: 10.1371/journal.pone.0183158

Joyce, A., Ijaz, U. Z., Nzeteu, C., Vaughan, A., Shirran, S. L., Botting, C. H., et al. (2018). Linking microbial community structure and function during the acidified anaerobic digestion of grass. *Front. Microbiol.* 9:540. doi: 10.3389/fmicb.2018.00540

Junemann, S., Kleinbolting, N., Jaenicke, S., Henke, C., Hassa, J., Nelkner, J., et al. (2017). Bioinformatics for NGS-based metagenomics and the application to biogas research. *J. Biotechnol.* 261, 10–23. doi: 10.1016/j.jbiotec.2017.08.012

Keiblinger, K. M., Fuchs, S., Zechmeister-Boltenstern, S., and Riedel, K. (2016). Soil and leaf litter metaproteomics-a brief guideline from sampling to understanding. *FEMS Microbiol. Ecol.* 92:fiw180. doi: 10.1093/femsec/fiw180

Keiblinger, K. M., Wilhartitz, I. C., Schneider, T., Roschitzki, B., Schmid, E., Eberl, L., et al. (2012). Soil metaproteomics - Comparative evaluation of protein extraction protocols. *Soil Biol. Biochem.* 54, 14–24. doi: 10.1016/j.soilbio.2012.05.014

Kleiner, M., Thorson, E., Sharp, C. E., Dong, X. L., Liu, D., Li, C., et al. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* 8:1558. doi: 10.1038/s41467-017-01544-x

Kohrs, F., Heyer, R., Bissinger, T., Kottler, R., Schallert, K., Puttker, S., et al. (2017). Proteotyping of laboratory-scale biogas plants reveals multiple steady-states in community composition. *Anaerobe* 46, 56–68. doi: 10.1016/j.anaerobe.2017.02.005

Kohrs, F., Heyer, R., Magnussen, A., Benndorf, D., Muth, T., and Behne, A. (2014). Sample prefractionation with liquid isoelectric focusing enables in depth microbial metaproteome analysis of mesophilic and thermophilic biogas plants. *Anaerobe* 29, 59–67. doi: 10.1016/j.anaerobe.2013.11.009

Kolmeder, C. A., De Been, M., Nikkila, J., Ritamo, I., Matto, J., Valmu, L., et al. (2012). Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. *PLoS One* 7:e29913. doi: 10.1371/journal.pone.0029913

Lehmann, T., Schallert, K., Vilchez-Vargas, R., Benndorf, D., Puttker, S., and Sydor, S. (2019). Metaproteomics of fecal samples of crohn's disease and ulcerative colitis. *J. Proteom.* 201, 93–103. doi: 10.1016/j.jprot.2019.04.009

Leon, I. R., Schwammle, V., Jensen, O. N., and Sprenger, R. R. (2013). Quantitative assessment of in-solution digestion efficiency identifies optimal protocols for unbiased protein analysis. *Mol. Cell Proteom.* 12, 2992–3005. doi: 10.1074/mcp.M112.025585

Liao, B., Ning, Z., Cheng, K., Zhang, X., Li, L., Mayne, J., et al. (2018). iMetaLab 1.0: a web platform for metaproteomics data analysis. *Bioinformatics* 34, 3954–3956. doi: 10.1093/bioinformatics/bty466

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., et al. (1999). Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682.

Mann, M., and Wilm, M. (1994). Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399. doi: 10.1021/ac00096a002

May, D. H., Timmins-Schiffman, E., Mikan, M. P., Haryey, H. R., Borenstein, E., Nunn, B. L., et al. (2016). An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *J. Proteom. Res.* 15, 2697–2705. doi: 10.1021/acs.jproteome.6b00239

Mesuere, B., Van Der Jeugt, F., Willems, T., Naessens, T., Devreese, B., Martens, L., et al. (2018). High-throughput metaproteomics data analysis with unipept: a tutorial. *J. Proteom.* 171, 11–22. doi: 10.1016/j.jprot.2017.05.022

Munk, B., Bauer, C., Gronauer, A., and Lebuhn, M. (2012). A metabolic quotient for methanogenic Archaea. *Water Sci. Technol.* 66, 2311–2317. doi: 10.2166/wst.2012.436

Muth, T., Behne, A., Heyer, R., Kohrs, F., Benndorf, D., Hoffmann, M., et al. (2015a). The metaproteomeanalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *J. Proteom. Res.* 14, 1557–1565. doi: 10.1021/pr501246w

Muth, T., Kolmeder, C. A., Salojarvi, J., Keskitalo, S., Varjosalo, M., Verdam, F. J., et al. (2015b). Navigating through metaproteomics data: a logbook of database searching. *Proteomics* 15, 3439–3453. doi: 10.1002/pmic.201400560

Muth, T., Benndorf, D., Reichl, U., Rapp, E., and Martens, L. (2013). Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol. Biosyst.* 9, 578–585. doi: 10.1039/c2mb25415h

Muth, T., Kohrs, F., Heyer, R., Benndorf, D., Rapp, E., Reichl, U., et al. (2018). MPA portable: a stand-alone software package for analyzing metaproteome samples on the go. *Anal. Chem.* 90, 685–689. doi: 10.1021/acs.analchem.7b03544

Nesvizhskii, A. I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data - the protein inference problem. *Mol. Cell. Proteom.* 4, 1419–1440. doi: 10.1074/mcp.r500012-mcp200

Patient, S., Wieser, D., Kleen, M., Kretschmann, E., Martin, M. J., and Apweiler, R. (2008). UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics* 24, 1321–1322. doi: 10.1093/bioinformatics/btn122

Perkins, D. N., Pappin, D. J. C., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567. doi: 10.1002/(sici)1522-2683(19991201)20:18<3551::aid-elps3551>3.0.co;2-2

Püttker, S., Kohrs, F., Benndorf, D., Heyer, R., Rapp, E., and Reichl, U. (2015). Metaproteomics of activated sludge from a wastewater treatment plant - A pilot study. *Proteomics* 15, 3596–3601. doi: 10.1002/pmic.201400559

Qin, J. J., Li, R. Q., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–U70. doi: 10.1038/nature08821

Rademacher, A., Zakrzewski, M., Schluter, A., Schonberg, M., Szczepanowski, R., Goesmann, A., et al. (2012). Characterization of microbial biofilms in a thermophilic biogas system by high-throughput metagenome sequencing. *FEMS Microbiol. Ecol.* 79, 785–799. doi: 10.1111/j.1574-6941.2011.01265.x

Rechenberger, J., Samaras, P., Jarzab, A., Behr, J., Frejno, M., and Djukovic, A. (2019). Challenges in clinical metaproteomics highlighted by the analysis of acute leukemia patients with gut colonization by multidrug-resistant *Enterobacteriaceae*. *Proteomes* 7, E2. doi: 10.3390/proteomes7010002

Schägger, H. (2006). Tricine-SDS-PAGE. *Nat. Protoc.* 1, 16–22. doi: 10.1038/nprot.2006.4

Schiebenhoefer, H., Van Den Bossche, T., Fuchs, S. Y., Renard, B., Muth, T., and Martens, L. (2019). Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Exp. Rev. Proteom.* 16, 375–390. doi: 10.1080/14789450.2019.1609944

Schirmer, E. C., Yates, J. R. III, and Gerace, L. (2003). MudPIT: a powerful proteomics tool for discovery. *Discov. Med.* 3, 38–39.

Schluter, A., Bekel, T., Diaz, N. N., Dondrup, M., Eichenlaub, R., and Gartemann, K. H. (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J. Biotechnol.gy* 136, 77–90. doi: 10.1016/j.jbiotec.2008.05.008

Schneider, T., Schmid, E., De Castro, J. V. Jr., Cardinale, M., Eberl, L., Grube, M., et al. (2011). Structure and function of the symbiosis partners of the lung lichen (Lobaria pulmonaria L. Hoffm.) analyzed by metaproteomics. *Proteomics* 11, 2752–2756. doi: 10.1002/pmic.201000679

Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* 1, 2856–2860. doi: 10.1038/nprot.2006.468

Stolze, Y., Bremges, A., Rumming, M., Henke, C., Maus, I., and Puhler, A. (2016). Identification and genome reconstruction of abundant distinct taxa in microbiomes from one thermophilic and three mesophilic production-scale biogas plants. *Biotechnol. Biofuels* 9, 156. doi: 10.1186/s13068-016-0565-3

Switzar, L., Van Angeren, J., Pinkse, M., Kool, J., and Niessen, W. M. A. (2013). A high-throughput sample preparation method for cellular proteomics using 96-well filter plates. *Proteomics* 13, 2980–2983. doi: 10.1002/pmic.201300080

Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A. J., and Bunk, D. M. (2010). Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteom. Res.* 9, 761–776. doi: 10.1021/pr9006365

Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., et al. (2016). The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* 4:51.

Tanca, A., Palomba, A., Pisanu, S., Deligios, M., Fraumene, C., Manghina, V., et al. (2014). A straightforward and efficient analytical pipeline for metaproteome characterization. *Microbiome* 2:49. doi: 10.1186/s40168-014-0049-2

Thorn, C. E., Bergesch, C., Joyce, A., Sambrano, G., Mcdonnell, K., Brennan, F., et al. (2018). A robust, cost-effective method for DNA, RNA and protein co-extraction from soil, other complex microbiomes, and pure cultures. *Mol. Ecol. Resour.* 19, 439–455. doi: 10.1111/1755-0998.12979

Timmins-Schiffman, E., May, D. H., Mikan, M., Riffle, M., Frazar, C., Harvey, H. R., et al. (2017). Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J.* 11, 309–314. doi: 10.1038/ismej.2016.132

Vizcaino, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., et al. (2016). 2016 update of the PRIDE database and its related tools (vol 44, pg D447, 2016). *Nucleic Acids Res.* 44, 11033–11033. doi: 10.1093/nar/gkw880

Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., et al. (1995). Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* 16, 1090–1094.

Wenzel, L., Heyer, R., Schallert, K., Löser, L., Wünschiers, R., Reichl, U., et al. (2018). 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. doi: 10.1002/elsc.201800062

Wilmes, P., and Bond, P. L. (2006). Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* 14, 92–97. doi: 10.1016/j.tim.2005.12.006

Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat. Methods* 6, 359–362. doi: 10.1038/nmeth.1322

Xiong, W. L., Abraham, P. E., Li, Z., Pan, C. L., and Hettich, R. L. (2015). Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota. *Proteomics* 15, 3424–3438. doi: 10.1002/pmic.201400571

Yao, X., Mcshane, A. J., and Castillo, M. J. (2013). "Chapter 17 - quantitative proteomics in development of disease protein biomarkers," in *Proteomic and Metabolomic Approaches to Biomarker Discovery*, eds H. J. Issaq, and T. D. Veenstra, (Boston: Academic Press), 259–278. doi: 10.1016/b978-0-12-394446-7.00017-0

Zhang, X., Deeke, S. A., Ning, Z. B., Starr, A. E., Butcher, J., Li, J., et al. (2018a). Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.* 9, 2873. doi: 10.1038/s41467-018-05357-4

Zhang, X., Li, L., Mayne, J., Ning, Z., Stintzi, A., and Figeys, D. (2018b). Assessing the impact of protein extraction methods for human gut metaproteomics. *J. Proteom.* 180, 120–127. doi: 10.1016/j.jprot.2017.07.001

Zoun, R., Schallert, K., Broneske, D., Heyer, R., Benndorf, D., and Saake, G. (2017). "Interactive chord visualization for metaproteomics," in *Database and Expert Systems Applications (DEXA), 2017 28th International Workshop on*, (France), 79–83.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Heyer, Schallert, Büdel, Zoun, Dorl, Behne, Kohrs, Püttker, Siewert, Muth, Saake, Reichl and Benndorf. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

TaxIt: An Iterative Computational Pipeline for Untargeted Strain-Level Identification Using MS/MS Spectra from Pathogenic Single-Organism Samples

doi: 10.1021/acs.jproteome.9b00714

Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for Seamless End-to-End Metaproteomics Data Analysis

doi: 10.1021/acs.jproteome.0c00136

A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and Prophane

doi: 10.1038/s41596-020-0368-7

doi: 10.1093/nargab/lqaa058

gNOMO: a multi-omics pipeline for integrated host and microbiome analysis of non-model organisms

Maria Muñoz-Benavent [1,*], Felix Hartkopf [2], Tim Van Den Bossche [3,4], Vitor C. Piro [2], Carlos García-Ferris [1,5], Amparo Latorre [1,6], Bernhard Y. Renard [2] and Thilo Muth [2,*]

[1] Institute for Integrative Systems Biology (I2SysBio), Universitat de València/CSIC, Paterna (València) 46980, Spain, [2] Bioinformatics Unit (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin 13353, Germany, [3] VIB - UGent Center for Medical Biotechnology, VIB, Ghent 9000, Belgium, [4] Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent 9000, Belgium, [5] Departament de Bioquímica i Biologia Molecular, Universitat de València. Burjassot (València) 46100, Spain and [6] Área de Genómica y Salud, Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO), València 46020, Spain

Received January 31, 2020; Revised June 19, 2020; Editorial Decision July 08, 2020; Accepted August 03, 2020

ABSTRACT

The study of bacterial symbioses has grown exponentially in the recent past. However, existing bioinformatic workflows of microbiome data analysis do commonly not integrate multiple meta-omics levels and are mainly geared toward human microbiomes. Microbiota are better understood when analyzed in their biological context; that is together with their host or environment. Nevertheless, this is a limitation when studying non-model organisms mainly due to the lack of well-annotated sequence references. Here, we present gNOMO, a bioinformatic pipeline that is specifically designed to process and analyze non-model organism samples of up to three meta-omics levels: metagenomics, metatranscriptomics and metaproteomics in an integrative manner. The pipeline has been developed using the workflow management framework Snakemake in order to obtain an automated and reproducible pipeline. Using experimental datasets of the German cockroach *Blattella germanica*, a non-model organism with very complex gut microbiome, we show the capabilities of gNOMO with regard to meta-omics data integration, expression ratio comparison, taxonomic and functional analysis as well as intuitive output visualization. In conclusion, gNOMO is a bioinformatic pipeline that can easily be configured, for integrating and analyzing multiple meta-omics data types and for producing output visualizations, specifically designed for integrating paired-end sequencing data with mass spectrometry from non-model organisms.

INTRODUCTION

Symbiosis is a widespread relationship present in all groups of organisms but intensely developed between animals and bacteria that benefit from each other in order to survive. Consequently, both acquire an evolutionary advantage in comparison to individuals lacking this relationship. Two different types of symbiosis can be distinguished: ectosymbiosis, in which bacteria are attached to the surface of the host, and endosymbiosis, which usually is a mutualistic relationship, where bacteria live intracellularly in the host and are transmitted vertically (1,2). To understand these evolutionary relationships host and symbionts are best studied together. In mutualistic symbiosis, the eukaryotes provide a safe environment for endosymbiotic bacteria that live in close interaction with the host. In return, the endosymbionts provide nutrients and metabolites (such as essential amino acids or vitamins) to the host that cannot be obtained in any other way. For example, it has been estimated that around 15% of insect species maintain endosymbiotic associations with bacteria that supply the host with the nutrients that are lacking in their diets (3) On the other hand, most insects possess a gut microbiome that affects the physiology of the host by, for example, contributing to metabolic and nutritional needs, and the immune system development

[*] To whom correspondence should be addressed. Tel: +49 30 8104 1943; Fax: +49 30 8104 1943; Email: thilo.muth@bam.de
Correspondence may also be addressed to Maria Muñoz-Benavent. Tel: +34 963 54 48 14; Fax: +34 963 54 48 14; Email: Maria.Munoz-Benavent@uv.es
Present addresses:
Felix Hartkopf, Section eScience (S.3), Federal Institute for Materials Research and Testing, Berlin 12205, Germany.
Vior C. Piro, Hasso-Plattner-Institute, Faculty of Digital Engineering, University of Potsdam, Potsdam 14482, Germany.
Bernhard Y. Renard, Hasso-Plattner-Institute, Faculty of Digital Engineering, University of Potsdam, Potsdam 14482, Germany.
Thilo Muth, Section eScience (S.3), Federal Institute for Materials Research and Testing, Berlin 12205, Germany.

© The Author(s) 2020. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

(4). Recently, many studies have been performed in humans to study the gut microbiota (5), but non-model organisms require further investigations to better understand this specific type of symbiosis. In this context, cockroaches are a suitable model, because they have two symbiotic systems, i.e. an endosymbiont (*Blattabacterium cuenoti*) in the fat body and a rich and complex gut microbiota (6,7). The German cockroach *Blattella germanica* is a hemimetabolous insect (it has an incomplete metamorphosis) with three developmental stages. Regarding its symbionts, genome analysis demonstrated that the endosymbiont *Blattabacterium* contributes to the nitrogen (N) recycling and the synthesis of essential amino acids (8), but the function of the gut microbiota in cockroaches still has to be elucidated. It has been shown that the gut microbiome of cockroaches shows much overlap with the one in humans probably reflecting a similar omnivorous diet (6,9–10).

Recently, research interests in microbial communities have been strongly increased due to findings on the impact of the microbiome on human health (11,12). Microbiome studies often employ meta-omics techniques such as metagenomics (13) that aims to analyze the genetic material from all members in a microbial community sample. Despite many advantages, metagenomics still presents a static gene-centric approach that cannot assess temporal dynamics and functional activities of complex microbial populations (14). To gain insights into the dynamic functional repertoire of microbial communities, further techniques such as metatranscriptomics and metaproteomics have been established in recent years (15,16). Beyond the genome level, these meta-omics analysis approaches allow studying complex microbial systems and their host interactions at the gene expression level (transcripts and proteins, respectively). Used separately, metagenomics, metatranscriptomics and metaproteomics are already powerful because they complement and mutually support each other. However, the bioinformatics analysis still faces various specific challenges that concern, for example, the identification of genes and proteins, the construction of multi-organism databases, the database selection process influencing the taxonomic and functional assignment (17), and the use of different sample extraction or data analysis protocols making the results comparison difficult (18). Finally, the lack of properly annotated reference genomes and proteomes is also a typical overseen issue in this context (19). These challenges must be overcome to design optimized and standardized meta-omics pipelines for analysing microbiome data.

In the past, powerful tailored bioinformatic solutions have been developed for the individual meta-omics analysis levels (13,15–16). However, the true strength unfolds when these analysis techniques are integrated (20,21). As a holistic approach, a complete meta-omics integration can extend the capabilities of microbiome and host-related studies in various ways. Most importantly, integrating multiple meta-omics levels allows to expand the possibilities of biological interpretation and to investigate biological pathways from a more comprehensive perspective. Compared to single-omics strategies, an integrative approach provides a deeper and more thorough understanding of how the key players of microbial communities regulate underlying pathway mechanisms (22).

While the integration of meta-omics has been described in previous studies (23), its potential has not been fully exploited so far. In particular, the data analysis is challenging, because studies often present customized in-house workflows that cannot be fully automated or are not reproducible. In general, automated multi-omics analysis pipelines are rare and limited to few meta-omics levels (24) and are not tailored for host and microbiome analyses of non-model organisms.

Here, we present gNOMO, a meta-omics software pipeline that allows integrating three different levels of omics analyses, derived from metagenomics, metatranscriptomics and metaproteomics experiments. It provides two different, optionally iterative operating modes: (i) each of the three omics levels can be analyzed separately and independently of each other and subsequently, (ii) up to three omics layers can be analyzed in a fully integrated fashion. The workflow of gNOMO starts from raw data to essential processing steps and finally provides output visualizations for taxonomic classification, functional metabolic pathway profiling and differential sample analysis. The integration of metagenomics, metatranscriptomics and metaproteomics data is possible due to the production of a tailored proteogenomic database, which optimizes the identification and quantification of peptides in metaproteomics data (25,26). As microbiota needs to be analyzed in its context, the host is also studied together with the microbiome. Host data can be analyzed without a reference database, which allows to study non-model organisms, and proteins of the host are also identified with a tailored host database obtained from genomics and transcriptomic sequences. The pipeline has been implemented using the Python-based Snakemake (27) framework to perform fully automated and reproducible multi-omics analyses of host and microbiome samples. So far, gNOMO has been developed and optimized for data from non-model organism samples, but it is fully executable on generic sample types, for example, from human or mouse microbiomes. With gNOMO, we aim to fill the gap of barely existing multi-omics pipelines for microbial community samples being able to compare and integrate data at the genome, transcriptome and proteome level.

MATERIALS AND METHODS

gNOMO is a pipeline that integrates multiple bioinformatic methods and software tools to analyze metagenomics, metatranscriptomics and metaproteomics data and to provide the results with an easily readable final output. One of the main purposes of integrating such different kinds of multi-omics data is to directly improve the analysis of microbial populations and to investigate their function in poorly characterized environments, such as non-model organisms. At the genome and transcriptome level, our pipeline includes both quality control and data preparation steps, of which parameters can be adjusted depending on the quality of the input data. In addition, gNOMO allows to directly create a proteogenomic database from metagenomics and metatranscriptomics data. This important processing step makes it possible to connect the metagenomics and metatranscriptomics analysis to the protein identification at the metaproteomics level. In particular, the proteoge-

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

nomic database generation step leads to the full integration of all three omics levels.

The complete gNOMO pipeline is built in Snakemake (27), a management system for bioinformatic workflows, that allows obtaining standardized and reproducible output data. The input data and parameters of programs that are used in Snakemake are defined by editing a single configuration file. Further, the gNOMO pipeline including all dependencies is available at the BioConda channel (28). Tools added to BioConda provide a user-friendly installation because the required tools and libraries are easily incorporated and automatically installed with the use of Snakemake environments. Due to the high computational needs of some parts of the workflow, we recommend a system with at least 16 available cores and at least 200 GB RAM. The storage requirements are data-dependent and were in our case about 1 TB of free storage. The runtime highly depends on the number of available cores because Snakemake is able to parallelize non-dependent tasks and decreases the runtime this way substantially. On a cluster node with 16 cores and 200 GB RAM the analysis of the *B. germanica* microbiome took about 72 h. The runtime of gNOMO can vary from run to run as it not only depends on CPU power but network speeds used, for example, for database updates as well. In addition, it should be stated that the Snakemake workflow engine is compatible and scalable in cluster environments (e.g. using the SLURM Workload Manager). The gNOMO pipeline typically consists of five main steps (Figure 1): (i) pre-processing, (ii) metagenomics and metatranscriptomics data analysis, (iii) proteogenomic database creation, (iv) metaproteomics data analysis and (iv) data integration. In the following paragraphs, these individual steps are described in more detail.

Pre-processing

The first step includes various pre-processing mechanisms improving metagenomics and metatranscriptomics read quality, including: (i) FastQC (29) for reviewing the quality of the reads, (ii) PrinSeq (30) for cleaning and for trimming the sequences, (iii) a second quality control with FastQC and Fastq-join (31) for binning the pair-end reads. This binning step is included because our workflow is designed for paired-end reads.

Metagenomic and metatranscriptomic analysis

In the metagenomic and metatranscriptomic analysis step, the pre-processed paired-end sequences are analyzed using pre-configured tools. These tools include (i) a genome mapping against the NCBI non-redundant (nr) database (accessed 5 July 2019) using Kaiju (32), (ii) an assembly using Ray, (iii) and protein prediction using both Prodigal (33) for bacterial proteins and (iv) Augustus (34) for host proteins. The contigs obtained through the genome assembly are used to increase the accuracy of the protein predictions. Bacterial proteins are predicted using Prodigal, a program specifically designed to predict bacterial open reading frames. Host proteins are predicted, with an engine (Augustus, (34)), from the same samples as bacterial proteins, because our pipeline is designed to analyze mixtures of host hindgut cells and bacterial cells. In this experiment, the vivisection process has been performed to ensure the only acquisition of hindgut tissue, essential to properly integrate bacterial data in its context, which is the hindgut of the host. Functional annotation of these predicted proteins is performed using EggNOG (version 1.0 accessed 5 June 2019) (35) to obtain KEGG Orthology (KO) identifiers. An optional step is included that requires the installation of InterProScan (36). This software is not implemented in BioConda but will be automatically installed locally with the snakemake script and allows a TIGRFAM (37) functional annotation. Details regarding the quality of the annotation in metagenomics and metatranscriptomics are available in the Supplementary Table S1.

Proteogenomic database generation

The output of the previous bacterial prediction from the metagenomics and metatranscriptomics data is used to create a proteogenomic database. This database includes bacterial and host proteins from metagenomics, metatranscriptomics or both kinds of data. A database with both kinds of information provides a comprehensive reference for peptide and protein identification (see next paragraph). The proteogenomic database obtained from the validation data has been built with the sequences resulting from the bacterial protein prediction performed with Prodigal. This database (data of creation: 19 November 2019) contains 1 014 200 sequences, of which 850 455 are unique (i.e. occur only once in the database).

Metaproteomic data analysis

For peptide and protein identification, MS-GF+ (38) is used as database search engine, employing the custom proteogenomic database as reference for peptide-to-spectrum matching. Both taxonomic and functional annotations of the peptides are performed with Unipept version 4.0 (39). The output obtained from this step is a taxonomic annotation at three different levels (phylum, family and genus) and the Enzyme Commission (EC) number associated with each peptide. To assess the performance of our tailored database, we compared the peptide identification yield with a very complete human gut microbial protein database: NIH Human Microbiome Project Gastrointestinal database (accessed 25 November 2019) (Supplementary Table S2). With our tailored database we obtained four times more peptides identified than using the NIH Gastrointestinal database. The search parameters are available in the modifications file for msgf plus (mods) and the config file. These results are consistent with previous studies on the use of metagenomic sequences for constructing proteogenomics databases (40).

Meta-omics data integration and visualization

The final step concerns the integration and visualization of all three-level meta-omics data and results. The integration of all three meta-omics data levels is performed in the following stages: (i) parallelized meta-omics analysis, (ii) proteogenomic database construction and (iii) pathway visualization.

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

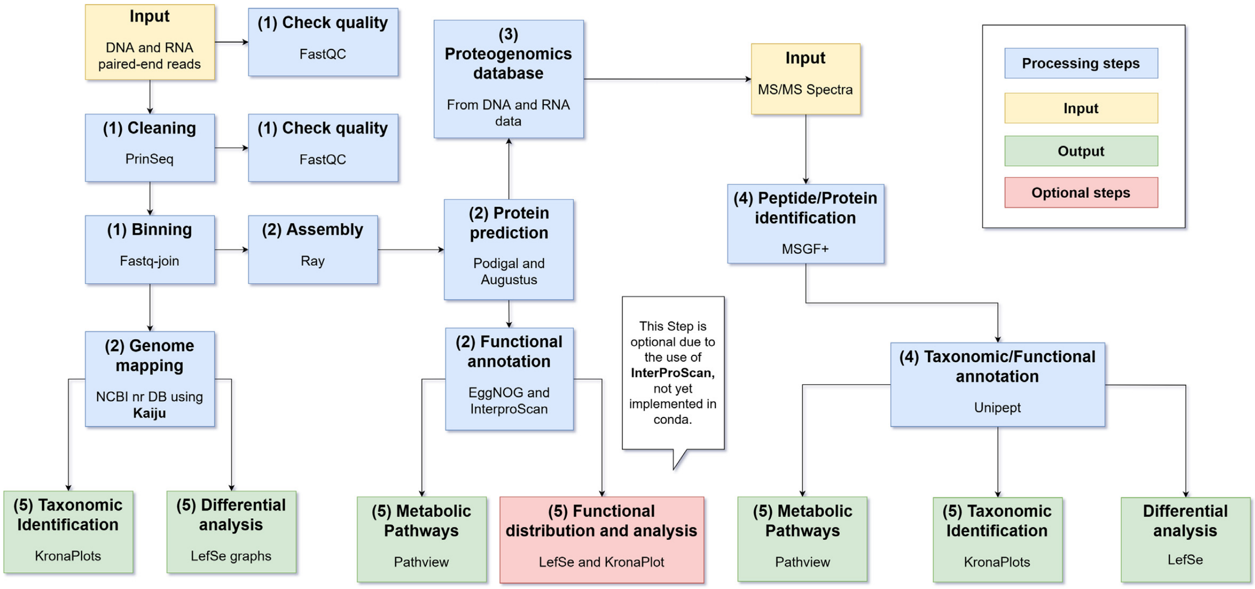


Figure 1. Workflow overview of the gNOMO pipeline. Each box represents a processing step in the pipeline. Box colors indicate the types of steps: input (orange), processing step (blue), optional step (red) and output (green). A legend with the colors is also incorporated. In each step, the process is indicated as well as the program used (in blue, red or green boxes), or which kind of input is required (in yellow boxes). Each blue, green and red box is marked with a number in parentheses indicating to which pipeline step it belongs: (1) pre-processing, (2) metagenomic and metatranscriptomic data analysis, (3) proteogenomics database construction, (4) metaproteomics data analysis, (5) final output visualizations based on the meta-omics integration.

First, both metagenomics and metatranscriptomics data are analyzed in parallel, which allows a reliable integration of them. The taxonomic annotation of the microbiome is visualized with KronaPlots (41). These plots show the taxonomic distribution in each sample reads for metagenomics and metatranscriptomics data. To analyze this information further, linear discriminant analysis (LDA) effect size (LEfSe) (42) is used that performs a statistical analysis on the microbiome data. LEfSe identifies features most likely to explain differences between conditions by coupling standard statistical tests with additional tests encoding biological consistency and effect relevance. The statistics performed are Kruskal-Wallis rank-sum test on classes, Wilcoxon rank-sum test among subclasses and LDA score on relevant features. Taking account of the effect size is essential to properly analyze microbiomes. The outcome of the statistical analysis is depicted in a graph with up to two levels of classification, and only the features with an LDA score over 2 are shown. This allows visualizing different conditions and different data within the same graph.

For the functional annotation, the representation of the metabolic pathways is included using Pathview (43), which allows pathway integration. The Pathview plots represent the log2 ratio of the means of the different conditions and data compared (i.e. 10d and 20d, metagenomic, metatranscriptomic and metaproteomic data, see below), after a fold change normalization. These log2 ratios are calculated for the proteins predicted from the contigs assembled from each sample. The database used to identify the peptides in the metaproteomics data is based on the protein prediction from the metagenomics and metatranscriptomics data. This proteogenomics approach creates a sample-specific protein database and therefore optimizes the peptide and protein identification at the metaproteome level, and provides a full integration of three datasets: metagenomics, metatranscriptomics and metaproteomics. The log2 ratio of the means of the peptides identified are then included in the Pathview visualization. When integrating all three datasets (metagenomics, metatranscriptomics and metaproteomics), the log2 ratios are compared between pairs of datasets (transcripts/gene, protein/gene, protein/transcript). Pathview shows these ratios as a color gradient, indicating which dataset is over-represented in the comparison. We can interpret if the transcriptional activity is high (transcripts over-represented among genes), or if the protein production is low (genes over-represented among proteins). This R-based tool shows the differential expression of the enzymes on graphs visualizing the selected metabolic pathways. Pathview itself uses functional pathway information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (44).

Validation data

Blattella germanica population originated from a stable laboratory population housed by Dr X. Bellés' group at the Institute of Evolutionary Biology (CSIC-UPF, Barcelona). It was reared in chambers at the Institute for Integrative Systems Biology (University of Valencia) at 25°C, 60% humidity and a photoperiod of 12L:12D. Cockroaches were fed dog-food pellets (Teklad global 21% protein dog diet 2021C, Envigo, Madison, WI, USA) and water *ad libitum*. Samples were taken at 10 days and 20 days after becoming adults, conditions names 10d and 20d, respectively. Vivisections of CO_2-anesthetized females were performed to obtain the hindgut of each individual. DNA and RNA sam-

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

ples were obtained from the same hindgut, with a total of 12 samples (six replicates per condition). Protein samples were obtained from individuals of the same age and population, with a total of eight samples (with four replicates per condition). Hindgut was ground with a sterile plastic pestle. DNA and RNA extraction of each hindgut was performed using Nucleospin RNA XS and Nucleospin DNA/RNA Buffer Set (Macherey-Nagel, France). Protein extraction of each hindgut was performed solubilizing the ground hindgut with lysis buffer (7 M urea, 2 M thiourea, 4% (w/v) CHAPS). Metagenomic sequencing using the Illumina MiSeq (2 × 300 bp) technology was done at the FISABIO (Valencia, Spain). Metaproteomics shotgun sequencing was performed by the Proteomics Unit of the Servei Central de Suport a la Investigació Experimental (SCSIE) at the University of Valencia.

A small subset of human data has been also analyzed in order to show the plasticity of the pipeline. The dataset consisted of two samples of metagenomics and metaproteomics data from the study of Tanca *et al*. (45). Both samples correspond to faecal samples from healthy Sardinian individuals: a female and a male.

RESULTS

To illustrate the outputs and analysis that can be obtained from this pipeline, we used a complex gut microbiota dataset from the non-model organism *B. germanica,* which genome has been sequenced (without being fully annotated) (46). This dataset consists of metagenomics, metatranscriptomics and metaproteomics data of two different adult conditions: 10d and 20d.

Comparison of metagenomics and metatranscriptomics/metaproteomics datasets for one-condition sample (multi-meta-omic approach)

Assessing bacterial composition from metagenomics and metatranscriptomics data. The analysis of microbial community samples often raises the question of which bacteria form a given population. To answer this question, we performed two different types of analysis using gNOMO. First, we processed and analyzed metagenomics data to investigate the taxonomic composition of a given sample. Second, we analyzed and compared samples of two different conditions: 10d and 20d.

For the first analysis, the output was visualized using a Krona plot that is produced for each metagenomics and metatranscriptomics sample automatically within the gNOMO pipeline. For the first-condition (10d) sample, we observed that the main phyla present in this population were *Bacteroidetes, Firmicutes* and *Proteobacteria* (Figure 2). After analyzing the taxonomic distribution differences between the 10d and 20d samples, we observed no significant abundance differences in a preliminary analysis (Supplementary Tables S3 and 4). In this analysis, the relative abundance of the main phyla and families was calculated in relation to the mean abundance of the two conditions. We observed that the four most abundant phyla distributions match our previous published studies based on 16S gene sequencing, while others (e.g. *Planctomycetes, Deferribacteres and Actinobacteria*) do not match exactly previous studies on this topic (10) (Supplementary Table S3). We made similar observations regarding taxonomic abundances at the family level (Supplementary Table S4). In general, this can be explained by the difference concerning the method and annotation between 16S rRNA gene sequencing analysis and metagenomics. 16S rRNA gene sequencing focuses on bacterial data and can be useful in environmental studies due to the lack of fully sequenced bacterial genomes in these kinds of scenarios. In contrast, metagenomics offers higher resolution, enabling a more specific taxonomic classification of sequences as well as the detection of new bacterial genes and genomes (47).

As described previously, our first analysis provided no clearly visible abundance differences between the two conditions, as we were expecting when studying such a stable situation (both are adult individuals differing in 10 days of development). However, we decided to validate this finding by a more sensitive statistical approach. To investigate this issue further, we used LEfSe (42) as a well-established statistical method for comparing the taxonomic distribution at genus level between 10d and 20d conditions. LEfSe has the advantage of recognizing the hierarchy of the taxonomic classification and accurately calculate statistically significant differences (represented as LDA scores) between different conditions.

Using LEfSe, we found, for example, that *Fusobacterium* (*Fusobacteriaceae* family), was more abundant at 10 days (LDA score > 3) in both metagenomics and metatranscriptomics data (Figure 3). The role of *Fusobacterium* on cockroaches' gut microbiome deserve a detailed study due to these results and some interesting findings about this groups' role in other organisms: *Fusobacterium* has been related to disease and stress situations in the human gut microbiota (48), but is has also been related to the infants gut microbiota (49). Conversely, an unidentified genus belonging to the family *Ruminococcaceae,* has been found more abundant in 20d than 10d condition (LDA score > 3) in metagenomics data (Figure 3A), but no differences between conditions have been found in metatranscriptomics data (Figure 3B). Various genera belonging to the family *Ruminococcaceae* have been related to a healthy gut microbiota, like *Ruminococcus* and *Faecalibacterium.* These have been linked to degradation of starch in the human colon making it available for other bacteria in the gut (50), and degradation of cellulose in herbivorous mammals (51). These differences between 10d and 20d conditions could suggest that, even if the population is very stable along adult stages, it is being rearranged to its final composition. This rearrangement would imply a reduction in *Fusobacterium* and an increase of *Ruminococcaceae* along time (10d against 20d, Figure 3A). On the other hand, *Pseudomonas* genus and an unclassified genus belonging to the family *Pelagibacteraceae* are more abundant only in metatranscriptomics analysis at 20d against 10d (Figure 3B). *Pelagibacteraceae* has been described as a bacterial family localized in marine and freshwater environments (52), but has also been detected in the mouse gut microbiome (53) *Pseudomonas* genus has been related to pathogenicity in animals and plants, and is a commonly detected taxa in the gut of cockroaches (54). These results suggest that these taxa

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

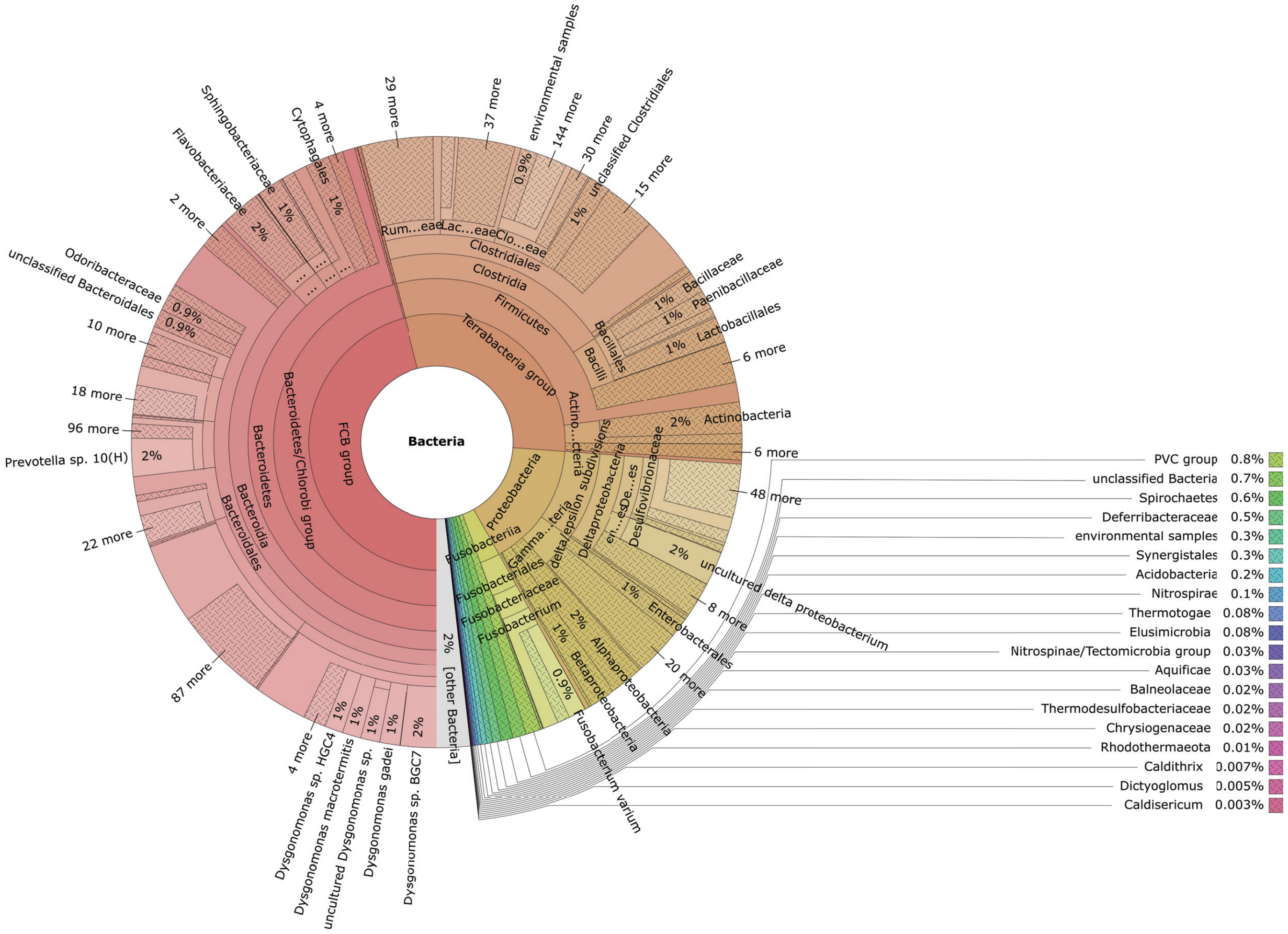


Figure 2. KronaPlot of the taxonomic annotation of a metagenomics sample (condition 10d). Bacterial taxa distribution of metagenomics data, corresponding to condition 10d. The bacterial taxa are classified by taxonomic hierarchy levels, from higher levels in the center of the chart (Kingdom *Bacteria*) progressing outward until genus level.

increase their transcriptional activity but not their abundance in the population along time. By the same reason the unidentified genus of *Ruminococcaceae* reduce its transcriptional activity (is over-represented at metagenomics level but not at metatranscriptomics level in 20d sample). More importantly, for the present work is the integration of this level of comparison that allows detection of particular taxa that differ significantly in their abundance in different conditions.

Functional analysis from integrated metagenomics and metatranscriptomics data for one-condition sample. Next steps concern the functional analysis of each microbiome dataset and the qualitative and quantitative differences of assigned functional annotations. To assess the level of transcriptional activity of the population, we compare the metagenomics data (gene pool) and the metatranscriptomics data (transcripts) corresponding to the microbiota of the 10d condition. Integrating metagenomics and metatranscriptomics allows calculating transcript/gene ratios that indicate gene transcriptional activation or repression. For this purpose, we applied LEfSe based on the functional role (or subrole) assignment using TIGRFAM (Figure 4 and Supplementary Table S5). We observed that energy metabolism (both anaerobic and aerobic metabolisms) and protein production are the most active metabolic pathways (Figure 4), which indicates that the bacterial population is active.

Alternatively, a pathway analysis enables discovering differences between states by using the Pathview R package. An analysis with Pathview shows which specific metabolic pathways (KEGG pathways) have statistically significant correlations between sample types and/or conditions and thereby complements the information provided by LEfSe. In a Pathview graph, an increase of the gene activity involved in a certain pathway can be observed. Our exemplary analysis using Pathview here focuses on the tricarboxylic acid cycle (TCA cycle) of the gut microbiota, comparing again gene pool (metagenomics data) against transcripts (metatranscriptomics data) (Figure 5). The TCA cycle consists of a series of oxidative reactions to finally obtain energy (adenosine triphosphate) from oxidative degradation of the acetyl group, in the form of acetyl-CoA, to carbon dioxide. The full cycle can be performed by bacteria in aerobic conditions, but some autotrophic bacteria are also able to per-

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

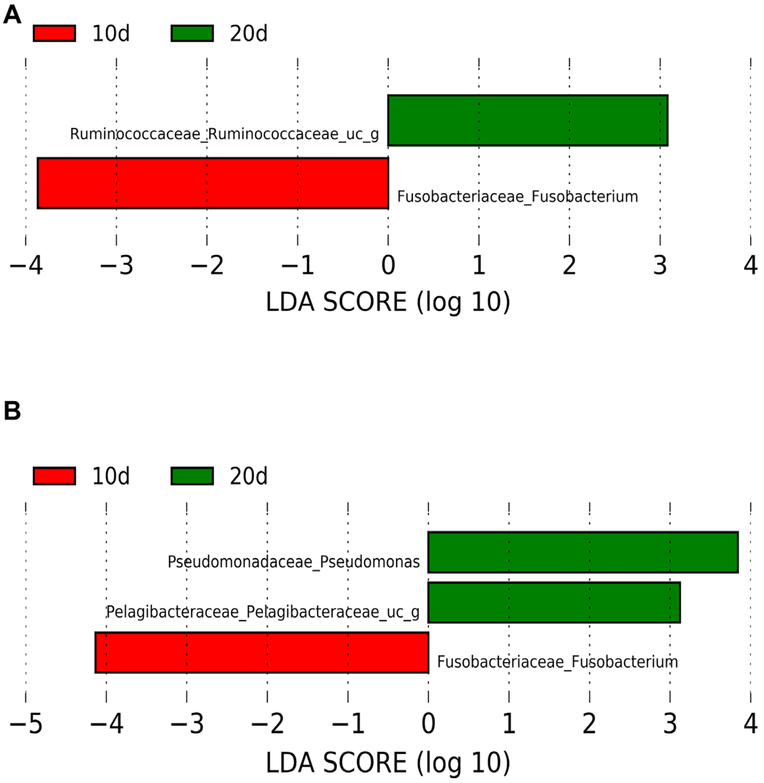


Figure 3. LEfSe graph of taxonomic annotation of metagenomics (top) and metatranscriptomics (bottom) data comparing the two conditions: 10d and 20d. Taxa with significant different distribution among the two conditions are identified. Only taxa with LDA scores over 2 are shown. Positive LDA scores are assigned to the taxa over-represented in the condition 20d (green), and negative LDA scores to the taxa over-represented in the condition 10d (red). Metagenomics data (**A**) and metatranscriptomics data (**B**) are represented.

form the reverse TCA cycle (rTCA), and even some anaerobic bacteria are able to carry out an incomplete TCA cycle, defining the pan-metabolic capabilities for this pathway of the gut microbiota.

We have found that most enzymes that take part in the TCA cycle are over-represented at the transcript level. This confirms our previous observations related to energy metabolism (Figure 4). With both analysis methods and their visualizations, we were able to study different levels of complexity of the pan-metabolism of all bacterial populations. We observed that the microbiome actively produces energy and proteins to grow and maintain a very complex population. Beyond the use case shown above, depending on the particular study, other pathways could be analyzed.

Meta-omics integration: comparing metagenomics, metatranscriptomics and metaproteomics data at the functional pathway level

Each meta-omics level data provides unique information in various ways, but their integration is crucial to gain a complete overview of the metabolic capabilities of the studied bacterial populations. Metaproteomics data incorporation to the integrated analysis of microbiomes is essential to have a realistic overview of the functional capabilities of the bacterial populations. For this purpose, we analyzed these meta-omics data together, as an example, focusing on the N metabolism pathway, corresponding to the N cycle, the set of reactions by which different inorganic N compounds are transformed into ammonia, a biologically reduced form of N that can be mainly introduced into synthesis of amino acids (glutamine and glutamate). We were interested in this pathway due to previous findings related to N metabolism of the host (*B. germanica*) and the endosymbiont *Blattabacterium*. As explained previously, *Blattabacterium* participates in the N recycling from stored urates to ammonia that can be used to synthesize glutamine and glutamate, connecting with the amino acid biosynthesis pathway (6). Here, the aim was to study N metabolism in the host gut microbiome and then to assess if the bacterial population has the metabolic capability to produce a form of usable N.

In this analysis, we investigated how variable or stable the overall N metabolism is at the gene, transcript and protein level along time (10d against 20d) in the investigated pathway (Figure 6). While metagenomics and metatranscriptomics show almost complete coverage of the N metabolism pathways and very variable along time, only a few enzymes were observed in the metaproteomics data and very stable along time. These results suggest that while the gene pool (the population) can be variable, the final transcripts and at least the four detected proteins remain stable, which could point in the direction of a functional redundancy at the protein level, as has been previously described for human gut

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

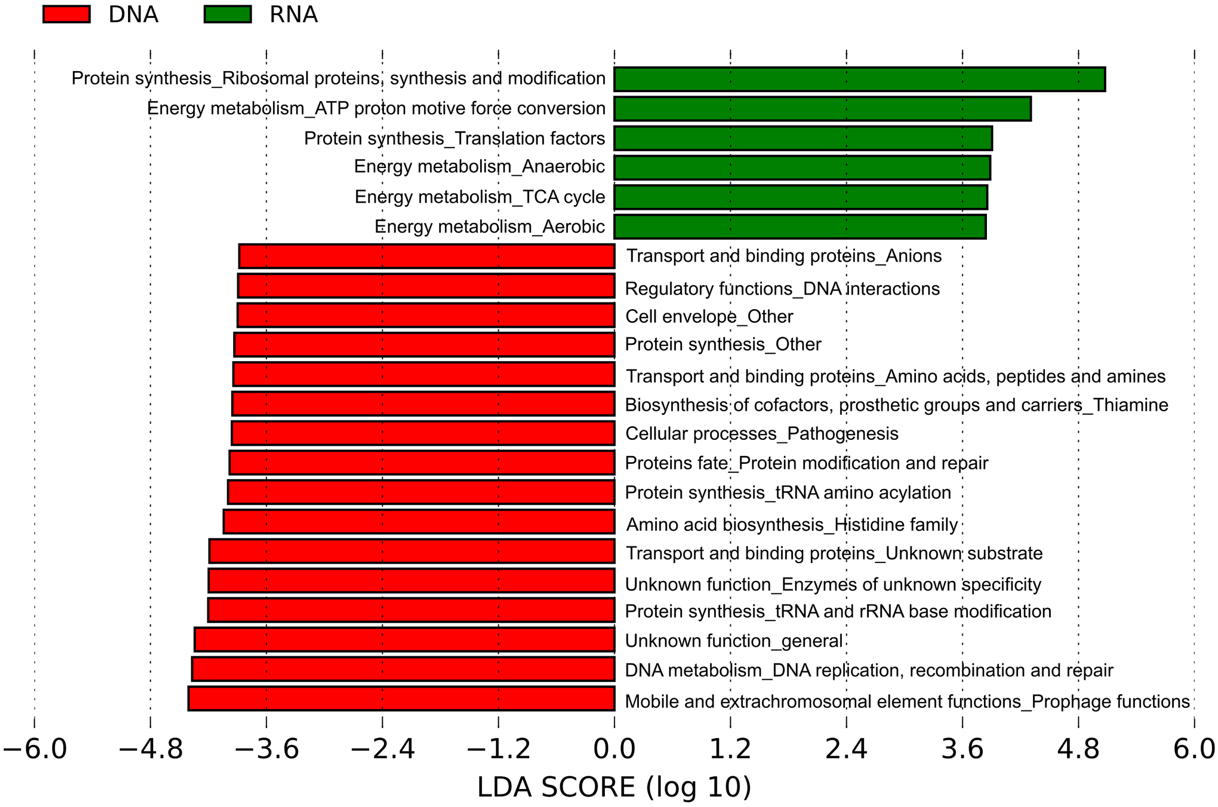


Figure 4. LEfSe graph comparing metagenomics and metatranscriptomics data of TIGRFAM annotation (role and subrole levels) of condition 10d. Taxa with significant different distribution among metagenomics and metatranscriptomics data are identified. Only taxa with LDA scores above 2 are shown. Positive LDA scores are assigned to the functional categories over-represented in the metatranscriptomics data (RNA, green), and negative LDA scores to the functional categories over-represented in metagenomics data (DNA, red).

microbiota (55). However, deeper coverage of the metaproteomics data would be necessary to confirm these findings.

Comparison of host and microbiome data

Microbiota metabolism and functions are better understood when studied together with its host. gNOMO includes the analysis of the host data in parallel with its microbiome, so we can integrate and compare the metabolic pathways of host and microbiome. In the case of *B. germanica,* we have studied the N metabolism pathway that we had analyzed before with the focus on the microbiota data (Figure 6) integrating the host data (Figure 7). We have observed which enzymes can be found in the bacterial population data and which ones can be explained by the host data (Figures 6 and 7).

We expected to find a maximum of four enzymes in the host data, as in most eukaryotes only four enzymes of this pathway are present, and we could detect those in the host pathway. While these four enzymes were the only ones detected in the host, its gut microbiome possesses most of the enzymes present in the N metabolism pathway.

If we study these four enzymes present in the host data in detail, it can be observed that all of them are over-represented at 10d against 20d condition in metaproteomics data, and in metagenomics and metatranscriptomics data, they are almost undetectable (Figure 7). When looking at the microbiome metatranscriptomics data, these proteins have a stable abundance over the whole time (Figure 6). These findings could indicate that the production of these proteins in the hindgut of the host is reduced along time, but its production by the microbiome remains stable.

After analyzing the bacterial and the host capabilities together regarding this metabolic pathway, we find that the N metabolism corresponding to the N cycle is mostly performed by the microbiome. These data show the importance of the meta-omics integration, as different levels of cell function are represented, each of them with different implications. DNA (in metagenomics) is more stable and can represent the gene pool of a population, but it can be misunderstood as also dead bacteria and genes which are not active are being represented with this methodology. RNA (in metatranscriptomics) shows the levels of active transcription, essential to understand the activity of a microbiome, which can differ substantially from the gene pool, both in bacterial and eukaryotic cells (Figures 6 and 7). The identified proteins for both microbial and host data, have been decisive to conclude that the N cycle is active in the German cockroaches' hindgut due to its microbiome (Figure 6). This conclusion is reinforced by the host data, as it has been proven that the host is not actively taking part of the N cycle (Figure 7). The importance of these findings should be analyzed in the future, including other path-

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

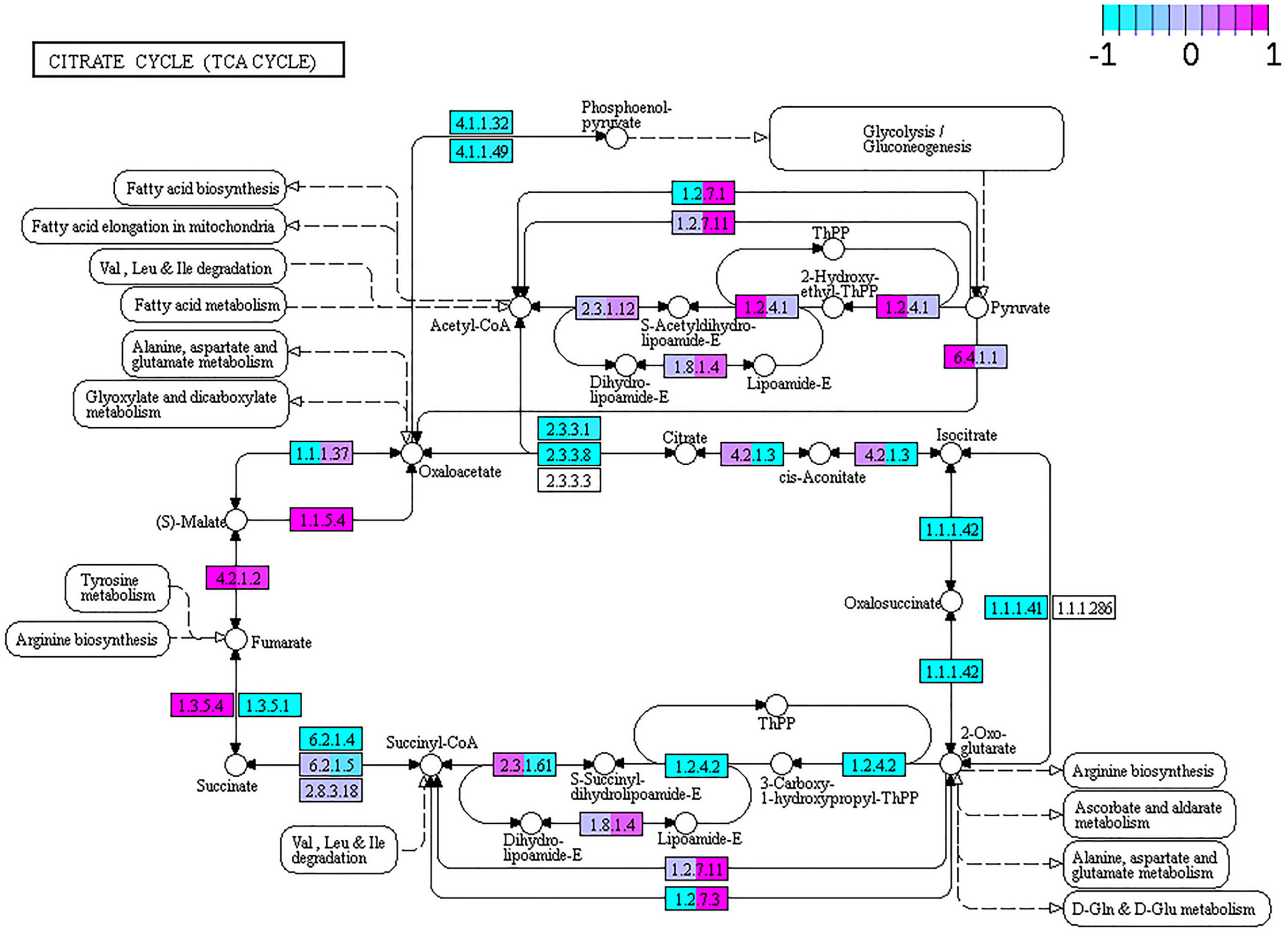


Figure 5. KEGG Pathview graph of the TCA cycle metabolism route comparing metagenomics versus metatranscriptomics data of the microbiota of 10d and 20d conditions. Some nodes are split between two colors, indicating 10d (left) and 20d (right) conditions. Light blue (−1) depicts genes under-represented in metagenomics (but over-represented in metatranscriptomics), while those marked in pink (1) depicts over-represented genes in metagenomics (but under-represented in metatranscriptomics). In purple, values close to 0 in the ratio metatranscriptomics/metagenomics, indicating no differences in frequency.

ways and improving the metaproteomics coverage of the microbiota.

Human microbiome dataset

In order to evaluate the applicability of gNOMO to other microbiome data, we performed an analysis on human microbiome data. In this analysis, we processed metagenomics and metaproteomics data of two healthy Sardinian individuals gut microbiota (45). The results of this exemplary analysis are included as two tables and two figures in the Supplementary File.

The basic statistics of the metagenomics data used are available in Supplementary Table S6. The output of the human dataset analysis includes the average taxonomic distribution of the metagenomics data of these samples in Supplementary Table S7. Our taxonomic identification at levels of phylum and family corresponds with the ranges obtained in the original study.

To exemplify the functional annotation output in the human dataset, we have included two Pathview graphs of the glycolysis/gluconeogenesis KEGG pathway. In Supplementary Figure S1, the two chosen conditions (male/female) are compared in both metagenomics and metaproteomics data. In this figure, the metatranscriptomics data possible spot in blank, which implies that the pipeline works even with the lack of one of the meta-omics data, and in general, the pipeline also works with all three meta-omics levels (as shown in the previous text). It should be noted that these exemplary data cannot be directly compared to the results of the original study, as their authors had not compared the microbiota between sexes. The results indicate an overall similar behavior of both bacterial populations, but with punctual strong divergences between individuals, which is in line with the results from the original study.

Finally, the ratio between metagenomics and metaproteomics data are studied in both conditions. The results show very different abundances between metagenomics and

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

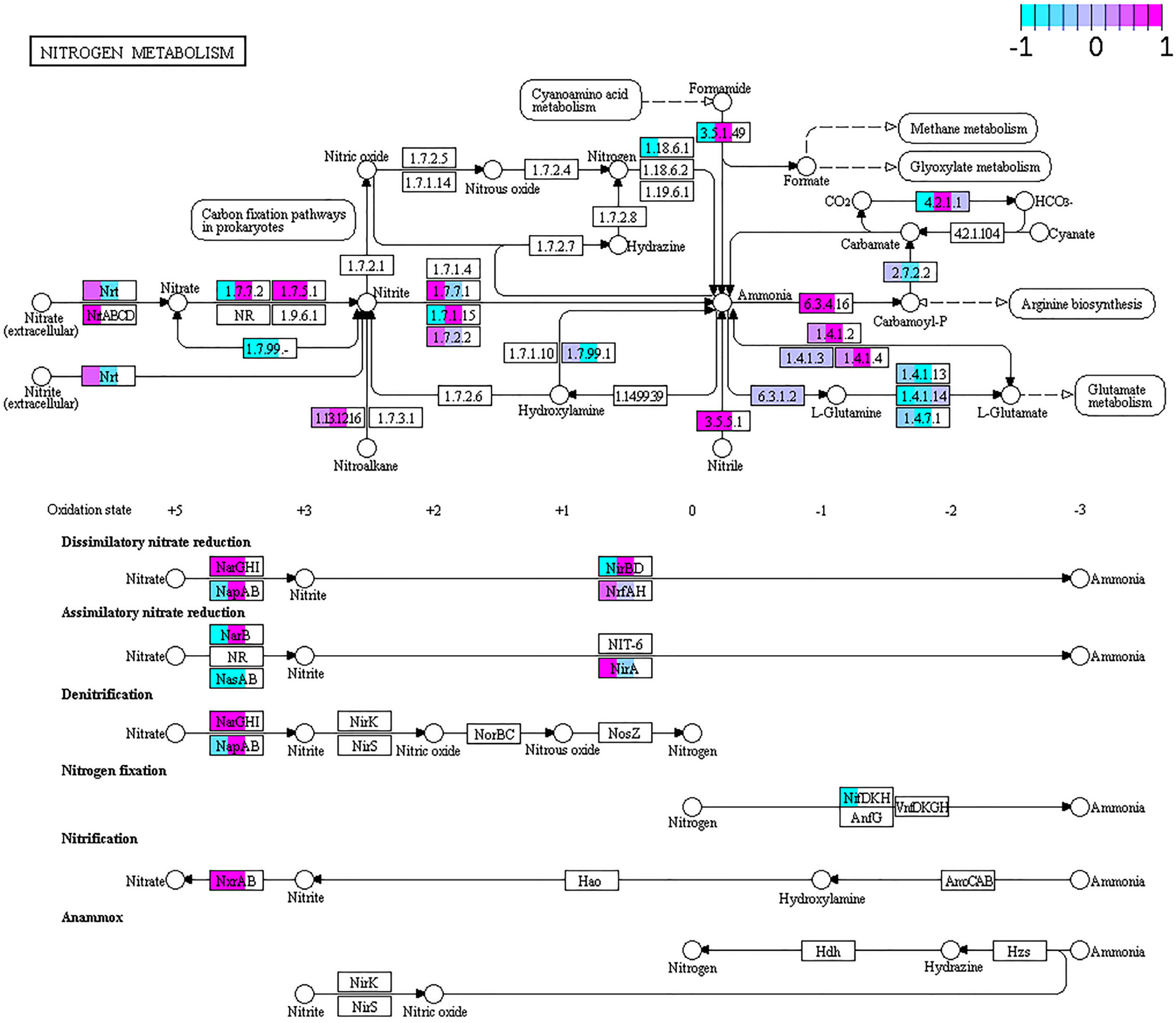


Figure 6. KEGG Pathview graph of the N metabolism route comparing metagenomics/metatranscriptomics/metaproteomics data of the microbiome at 10d and 20d. Some nodes are split between different colors, indicating metagenomics (left), metatranscriptomics (middle) and metaproteomics (right) data. Light blue (−1) depicts genes/transcripts/proteins over-represented in 10d (but under-represented in 20d), while those marked in pink (1) depicts genes/transcripts/proteins over-represented in 20d (but under-represented in 10d). In purple, values close to 0 in the ratio 10d/20d, indicating no differences in frequency.

metaproteomics data, which indicates high or low translational activity, depending on the positive or negative value of the ratio (Supplementary Figure S2). These findings also confirm the results obtained from the Sardinian cohort study.

DISCUSSION

The aim of our software design and implementation was to provide a complete pipeline to analyze omics data from a non-model host and its microbiome. Based on these requirements, we developed the gNOMO software that presents an end-to-end workflow covering all the required data analysis steps starting from the processing of raw omics data to the final output visualization of the results. gNOMO performs the analysis of up to three different meta-omics data: metagenomics, metatranscriptomics and metaproteomics, and their integration.

gNOMO is designed for paired-end sequencing of metagenomics and metatranscriptomics data, the pipeline includes a preprocessing and binning step designed for this type of datasets. A tailored proteogenomic database is generated to perform a highly efficient database search for protein identification in the metaproteomics data analysis without a reference microbiome. To obtain this database metagenomics and metatranscriptomics data are assembled

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

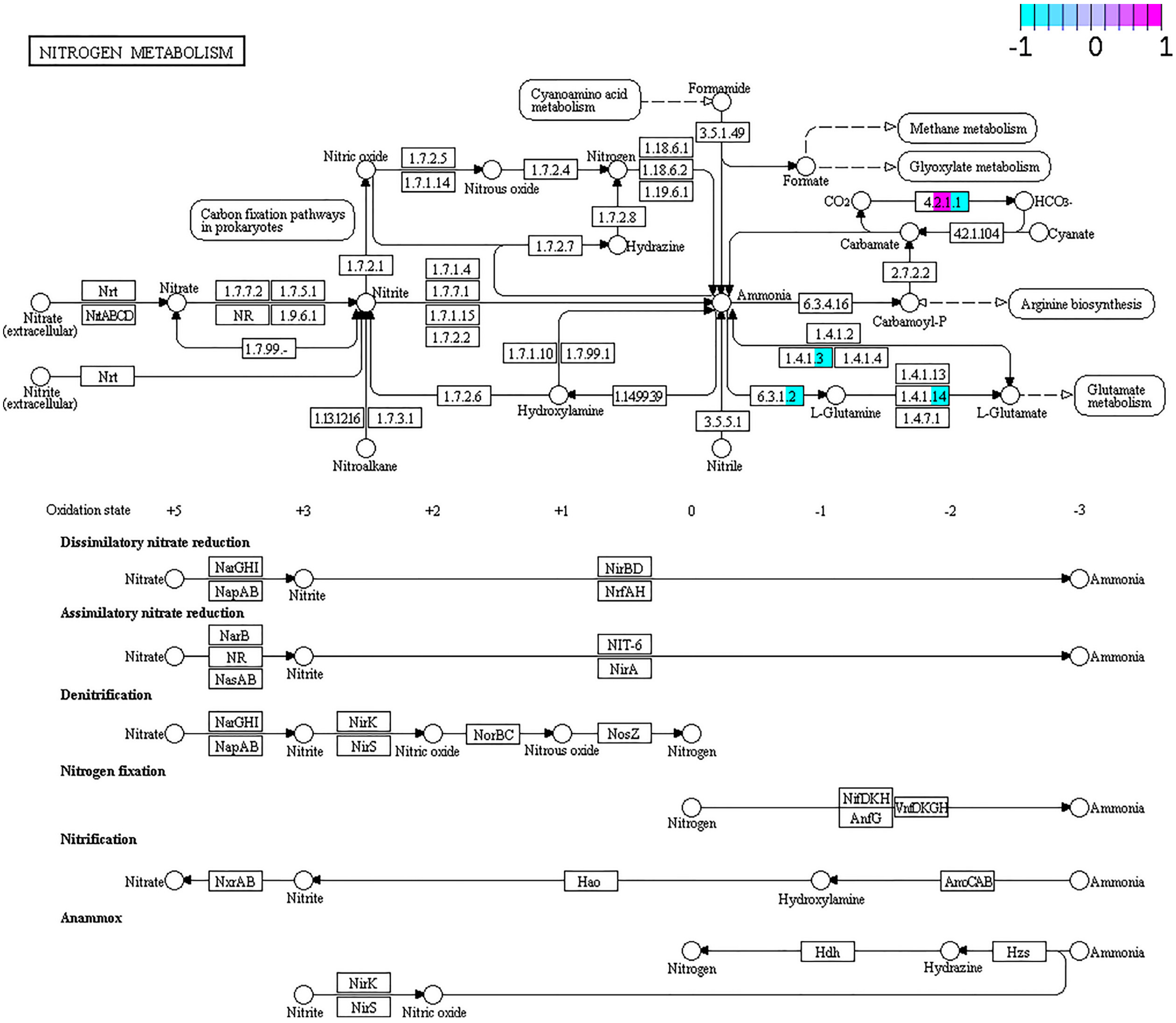


Figure 7. KEGG Pathview graph of the N metabolism pathways comparing metagenomics/metatranscriptomics/metaproteomics data of the host between 10d and 20d conditions. Some nodes are split between different colors, indicating metagenomics (left), metatranscriptomics (middle) and metaproteomics (right) data. Light blue (−1) depicts genes/transcripts/proteins over-represented in 10d (but under-represented in 20d), while those marked in pink (1) depicts genes/transcripts/proteins over-represented in 20d (but under-represented in 10d). In pruple, values close to 0 in the ratio 10d/20d, indicating no differences in frequency.

into contigs, which are then used to predict the proteins present in the samples. Together with the microbiome data, host data is obtained from the same samples and analyzed *de novo* in order to be able to analyze microbiota of non-model organisms integrated with the host information. Host databases can also be provided to analyze human or other model organisms data.

The pipeline is developed using the modular Snakemake framework that allows to incorporate software tools and libraries with different requirements. These tools are available at the BioConda channel and their installation is incorporated in the workflow. Snakemake makes use of programming languages Python and Bash, which are commonly used in bioinformatics. Parameters can be specified in the configuration file provided to Snakemake, so it can be adapted to any kind of host or microbiome analyzed. The use of Snakemake makes gNOMO fully automated, efficient and reproducible.

Previously published meta-omics workflows such as IMP (24) incorporate two layers of meta-omics information by integrating metagenomics and metatranscriptomics data. Such workflows focus on the analysis of the microbiome and often consider host information as contaminant reads: thus, instead of providing a host data analysis, the host genome is only used to remove the host information from the microbiome data. To overcome this issue, gNOMO of-

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

fers the possibility to analyze host data in parallel to microbiome data and both datasets can be studied simultaneously. gNOMO includes the analysis of metaproteomics data and creates a tailored proteogenomic database to achieve better and more efficient protein identification. The incorporation of the metaproteomics data to the study of the microbiome gives another dimension to the analysis of the microbiome because the proteome provides the functional profile and thereby gives insights on the actual interaction between microbial populations and their host.

The visualization output provided by gNOMO pipeline includes krona charts for taxonomic distribution, and KO categories are plotted using Pathview graphs. The functional distribution represented with Pathview permits to investigate two different aspects: first, the completeness of the metabolic pathways by visualizing each enzyme in the route, and second, the differences in abundance of each enzyme by comparing datasets (metagenomics, metatranscriptomics and metaproteomics) or conditions. This integration in gNOMO is highly useful, for example, when information regarding the presence and abundance of specific enzymes is needed. The integration is developed at three different stages: the parallelization of the meta-omics datasets, the integration of the functional annotation in Pathview pathways, and the construction of a proteogenomics database with metagenomics and metatranscriptomics information to identify peptides and proteins from the metaproteomics dataset.

With the study of a small human dataset, we can show the plasticity and adaptation capability of the pipeline to any type of dataset. The results obtained from this study validate the results from the paper the exemplary dataset was obtained from (45), which also proves that gNOMO is a robust and reproducible workflow to work with.

In conclusion, gNOMO is a standardized and reproducible bioinformatic pipeline designed to integrate and analyze metagenomics, metatranscriptomics and metaproteomics microbiota data of non-model organisms. It incorporates preprocessing, binning, assembly steps, taxonomic and functional annotations, and the production of a proteogenomic database to improve the metaproteomics analysis. gNOMO also includes the analysis of both microbiota and host data in parallel, which makes it a useful tool to analyze the microbiome of non-model organisms, as it was demonstrated using experimental data of the German cockroach *B. germanica*. In general, gNOMO can also be applied to data from human or other model organism sample types. Finally, gNOMO generates output and visualization of multiple meta-omics results in a single automated pipeline.

DATA AVAILABILITY

gNOMO is an open source software available in the GitHub repository: https://gitlab.com/rki_bioinformatics/gnomo and https://gitlab.com/gaspilleura/gnomo.

The validation data have been deposited with Zenodo under the accession number 3569690 (https://doi.org/10.5281/zenodo.3569690), metagenomics and metatranscriptomics data have been deposited with ENA under the accession number PRJEB37860 (http://www.ebi.ac.uk/ena/data/view/PRJEB37860) and metaproteomics data have been submitted to PRIDE under the accession number (PXD018642).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Dr Nuria Jiménez and FISABIO for the help in the processing and sequencing of the metagenomics and metatranscriptomics samples and the Proteomics Service of the SCSIE for the processing and sequencing of the metaproteomics samples.

FUNDING

European Regional Development Fund (ERDF), Ministerio de Ciencia, Innovación y Universidades [PGC2018-099344-B-I0 to A.L.]; Generalitat Valenciana [PROMETEO/2018/133 to A.L.], Research Foundation Flanders [1S90918N to T.V.D.B.]; Deutsche Forschungsgemeinschaft [RE 3474/2-2 to B.Y.R.]; Ministerio de Ciencia, Innovación y Universidades, FPU Fellowship [FPU15/01203 to M.M.B.]; Federation of European Biochemical Societies [FEBS Summer Fellowship to TVDB].

Conflict of interest statement. None declared.

REFERENCES

1. Gil,R. and Latorre,A. (2019) Unity makes strength: a review on mutualistic symbiosis in representative insect clades. *Life*, **9**, 21.
2. Moya,A., Peretó,J., Gil,R. and Latorre,A. (2008) Learning how to live together: genomic insights into prokaryote–animal symbioses. *Nat. Rev. Genet.*, **9**, 218–229.
3. Douglas,A.E. (2011) Lessons from studying insect symbioses. *Cell Host Microbe*, **10**, 359–367.
4. Moran,N.A., Ochman,H. and Hammer,T.J. (2019) Evolutionary and ecological consequences of gut microbial communities. *Annu. Rev. Ecol. Evol. Syst.*, **50**, 451–475.
5. Heintz-Buschart,A., May,P., Laczny,C.C., Lebrun,L.A., Bellora,C., Krishna,A., Wampach,L., Schneider,J.G., Hogan,A., de Beaufort,C. *et al.* (2017) Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.*, **2**, 16180.
6. Carrasco,P., Pérez-Cobas,A.E., van de Pol,C., Baixeras,J., Moya,A. and Latorre,A. (2014) Succession of the gut microbiota in the cockroach Blattella germanica. *Int. Microbiol.*, **17**, 99–109.
7. López-Sánchez,M.J., Neef,A., Peretó,J., Patiño-Navarrete,R., Pignatelli,M., Latorre,A. and Moya,A. (2009) Evolutionary convergence and Nitrogen metabolism in Blattabacterium strain Bge, Primary endosymbiont of the cockroach Blattella germanica. *PLoS Genet.*, **5**, e1000721.
8. Patiño-Navarrete,R., Piulachs,M.D., Belles,X., Moya,A., Latorre,A. and Peretó,J. (2014) The cockroach Blattella germanica obtains nitrogen from uric acid through a metabolic pathway shared with its bacterial endosymbiont. *Biol. Lett.*, **10**, 20140407.
9. Pérez-Cobas,A.E., Gosalbes,M.J., Friedrichs,A., Knecht,H., Artacho,A., Eismann,K., Otto,W., Rojo,D., Bargiela,R., Von Bergen,M. *et al.* (2013) Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut*, **62**, 1591–1601.
10. Rosas,T., García-Ferris,C., Domínguez-Santos,R., Llop,P., Latorre,A. and Moya,A. (2018) Rifampicin treatment of Blattella germanica evidences a fecal transmission route of their gut microbiota. *FEMS Microbiol. Ecol.*, **94**, fiy002.
11. Cani,P.D. (2018) Human gut microbiome: hopes, threats and promises. *Gut*, **67**, 1716–1725.

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

12. Mohajeri,M.H., Brummer,R.J.M., Rastall,R.A., Weersma,R.K., Harmsen,H.J.M., Faas,M. and Eggersdorfer,M. (2018) The role of the microbiome for human health: from basic science to clinical applications. *Eur. J. Nutr.*, **57**, 1–14.
13. Piro,V.C., Matschkowski,M. and Renard,B.Y. (2017) MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, **5**, 101.
14. Knight,R., Vrbanac,A., Taylor,B.C., Aksenov,A., Callewaert,C., Debelius,J., Gonzalez,A., Kosciolek,T., McCall,L.-I., McDonald,D. *et al.* (2018) Best practices for analysing microbiomes. *Nat. Rev. Microbiol.*, **16**, 410–422.
15. Martinez,X., Pozuelo,M., Pascal,V., Campos,D., Gut,I., Gut,M., Azpiroz,F., Guarner,F. and Manichanh,C. (2016) MetaTrans: an open-source pipeline for metatranscriptomics. *Sci. Rep.*, **6**, 25447.
16. Muth,T., Behne,A., Heyer,R., Kohrs,F., Benndorf,D., Hoffmann,M., Lehtevä,M., Reichl,U., Martens,L. and Rapp,E. (2015) The MetaProteomeAnalyzer: a powerful Open-Source software suite for metaproteomics data analysis and interpretation. *J. Proteome Res.*, **14**, 1557–1565.
17. Heyer,R., Schallert,K., Zoun,R., Becher,B., Saake,G. and Benndorf,D. (2017) Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.*, **261**, 24–36.
18. Wang,W.-L., Xu,S.-Y., Ren,Z.-G., Tao,L., Jiang,J.-W. and Zheng,S.-S. (2015) Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.*, **21**, 803–814.
19. Shakya,M., Lo,C.-C. and Chain,P.S.G. (2019) Advances and challenges in metatranscriptomic analysis. *Front. Genet.*, **10**, 904.
20. Manzoni,C., Kia,D.A., Vandrovcova,J., Hardy,J., Wood,N.W., Lewis,P.A. and Ferrari,R. (2018) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.*, **19**, 286–302.
21. Hernández-de-Diego,R., Tarazona,S., Martínez-Mira,C., Balzano-Nogueira,L., Furió-Tarí,P., Pappas,G.J. and Conesa,A. (2018) PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.*, **46**, W503–W509.
22. Moya,A. and Ferrer,M. (2016) Functional redundancy-induced stability of gut microbiota subjected to disturbance. *Trends Microbiol.*, **24**, 402–413.
23. Franzosa,E.A., Morgan,X.C., Segata,N., Waldron,L., Reyes,J., Earl,A.M., Giannoukos,G., Boylan,M.R., Ciulla,D., Gevers,D. *et al.* (2014) Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E2329–E2338.
24. Narayanasamy,S., Jarosz,Y., Muller,E.E.L., Heintz-Buschart,A., Herold,M., Kaysen,A., Laczny,C.C., Pinel,N., May,P. and Wilmes,P. (2016) IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.*, **17**, 260.
25. Ruggles,K. V., Krug,K., Wang,X., Clauser,K.R., Wang,J., Payne,S.H., Fenyö,D., Zhang,B. and Mani,D.R. (2017) Methods, tools and current perspectives in proteogenomics. *Mol. Cell. Proteomics*, **16**, 959–981.
26. Schiebenhoefer,H., Van Den Bossche,T., Fuchs,S., Renard,B.Y., Muth,T. and Martens,L. (2019) Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Rev. Proteomics*, **16**, 375–390.
27. Köster,J. and Rahmann,S. (2012) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
28. Grüning,B., Dale,R., Sjödin,A., Chapman,B.A., Rowe,J., Tomkins-Tinch,C.H., Valieris,R. and Köster,J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475–476.
29. Andrews,Si. (2010) FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinformatics*, http://www.bioinformatics.babraham.ac.uk/projects/.
30. Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.
31. Aronesty,E. (2013) Comparison of sequencing utility programs. *Open Bioinform. J.*, **7**, 1–8.
32. Menzel,P., Ng,K.L. and Krogh,A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.*, **7**, 11257.
33. Hyatt,D., Chen,G.-L., LoCascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
34. Stanke,M. and Morgenstern,B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, **33**, W465–W467.
35. Jensen,L.J., Julien,P., Kuhn,M., von Mering,C., Muller,J., Doerks,T. and Bork,P. (2007) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, **36**, D250–D254.
36. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
37. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
38. Kim,S. and Pevzner,P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.*, **5**, 5277.
39. Gurdeep Singh,R., Tanca,A., Palomba,A., Van der Jeugt,F., Verschaffelt,P., Uzzau,S., Martens,L., Dawyndt,P. and Mesuere,B. (2019) Unipept 4.0: Functional analysis of metaproteome data. *J. Proteome Res.*, **18**, 606–615.
40. Tanca,A., Palomba,A., Fraumene,C., Pagnozzi,D., Manghina,V., Deligios,M., Muth,T., Rapp,E., Martens,L., Addis,M.F. *et al.* (2016) The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome*, **4**, 51.
41. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
42. Segata,N., Izard,J., Waldron,L., Gevers,D., Miropolsky,L., Garrett,W.S. and Huttenhower,C. (2011) Metagenomic biomarker discovery and explanation. *Genome Biol.*, **12**, R60.
43. Luo,W., Pant,G., Bhavnasi,Y.K., Blanchard,S.G. and Brouwer,C. (2017) Pathview Web: user friendly pathway visualization and data integration. *Nucleic Acids Res.*, **45**, W501–W508.
44. Kanehisa,M. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
45. Tanca,A., Abbondio,M., Palomba,A., Fraumene,C., Manghina,V., Cucca,F., Fiorillo,E. and Uzzau,S. (2017) Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome*, **5**, 79.
46. Harrison,M.C., Jongepier,E., Robertson,H.M., Arning,N., Bitard-Feildel,T., Chao,H., Childers,C.P., Dinh,H., Doddapaneni,H., Dugan,S. *et al.* (2018) Hemimetabolous genomes reveal molecular basis of termite eusociality. *Nat. Ecol. Evol.*, **2**, 557–566.
47. Jovel,J., Patterson,J., Wang,W., Hotte,N., O'Keefe,S., Mitchel,T., Perry,T., Kao,D., Mason,A.L., Madsen,K.L. *et al.* (2016) Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.*, **7**, 459.
48. Saito,K., Koido,S., Odamaki,T., Kajihara,M., Kato,K., Horiuchi,S., Adachi,S., Arakawa,H., Yoshida,S., Akasu,T. *et al.* (2019) Metagenomic analyses of the gut microbiota associated with colorectal adenoma. *PLoS One*, **14**, e0212406.
49. Rinninella,E., Raoul,P., Cintoni,M., Franceschi,F., Miggiano,G., Gasbarrini,A. and Mele,M. (2019) What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms*, **7**, 14.
50. Flint,H.J., Scott,K.P., Louis,P. and Duncan,S.H. (2012) The role of the gut microbiota in nutrition and health. *Nat. Rev. Gastroenterol. Hepatol.*, **9**, 577–589.
51. Douglas,A.E. (2009) The microbial dimension in insect nutritional ecology. *Funct. Ecol.*, **23**, 38–47.
52. Ortmann,A.C. and Santos,T.T.L. (2016) Spatial and temporal patterns in the Pelagibacteraceae across an estuarine gradient. *FEMS Microbiol. Ecol.*, **92**, fiw133.
53. Dranse,H.J., Zheng,A., Comeau,A.M., Langille,M.G.I., Zabel,B.A. and Sinal,C.J. (2018) The impact of chemerin or chemokine-like receptor 1 loss on the mouse gut microbiome. *PeerJ*, **6**, e5494.
54. Moges,F., Eshetie,S., Endris,M., Huruy,K., Muluye,D., Feleke,T., G/Silassie,F., Ayalew,G. and Nagappan,R. (2016) Cockroaches as a source of high bacterial pathogens with multidrug resistant strains in Gondar Town, Ethiopia. *Biomed. Res. Int.*, **2016**, 2825056.
55. Lozupone,C.A., Stombaugh,J.I., Gordon,J.I., Jansson,J.K. and Knight,R. (2012) Diversity, stability and resilience of the human gut microbiota. *Nature*, **489**, 220–230.

Downloaded from https://academic.oup.com/nargab/article/2/3/lqaa058/5881268 by guest on 23 January 2021

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-27542-8 OPEN

Check for updates

Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows

Tim Van Den Bossche[1,2,23], Benoit J. Kunath[3,23], Kay Schallert[4,23], Stephanie S. Schäpe[5,23], 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], Pedro T. Queiros[3], Udo Reichl[4,11], Bernhard Y. Renard[12,17], Henning Schiebenhoefer[12,17], Alexander Sczyrba[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[3,21], Maximilian Wolf[4], Lennart Martens[1,2,24]✉ & Thilo Muth[22,24]

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 carry out a community-driven, multi-laboratory comparison in metaproteomics: the critical assessment of metaproteome investigation study (CAMPI). Based on well-established workflows, we evaluate the effect of sample preparation, mass spectrometry, and bioinformatic analysis using two samples: a simplified, laboratory-assembled human intestinal model and a human fecal sample. We observe that variability at the peptide level is predominantly due to sample processing workflows, with a smaller contribution of bioinformatic pipelines. These peptide-level differences largely disappear at the protein group level. While differences are observed for predicted community composition, similar functional profiles are obtained across workflows. CAMPI demonstrates the robustness of present-day metaproteomics research, serves as a template for multi-laboratory studies in metaproteomics, and provides publicly available data sets for benchmarking future developments.

A full list of author affiliations appears at the end of the paper.

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 adaptations 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 the shotgun proteomics approach. Here, the proteins are subsequently extracted, isolated, and digested into peptides, after which these are separated and analyzed using liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS). The obtained MS/MS spectra are then matched against in silico generated spectra derived from a protein sequence database, leading to peptide spectrum matches (PSMs). Hereafter, the identified peptides are used to infer the proteins present in the sample. Proteins can then be annotated with taxa and functions, providing information on 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. Sample processing 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. Environmental samples, such as feces or soil, are complex mixtures that can contain microbial cells, host cells, plant-derived fibrous materials, and other abiotic components. Therefore, the composition and abundance of these components must be considered when choosing an appropriate method for cellular lysis and protein extraction. Fortunately, the most commonly used methods nowadays are relatively robust, and generally provide a reasonably representative extraction of proteins found in these complex mixtures. However, because differences exist, methods still need to be optimized for the specific samples and projects[13,14] Besides, apart from different sample processing protocols, different mass spectrometers might also lead to a variation in results.

Moreover, metaproteomics comes with many specific bioinformatic challenges[8,15]. First, the choice of an appropriate sequence database is critical for peptide identification[16,17]. Typically, large databases can strongly impact sensitivity and false discovery rate (FDR) estimation[18], while incomplete reference databases can lead to missing or false positive identifications[19,20]. Second, the protein inference problem[21] is more pronounced in metaproteomics due to many homologous proteins from closely related organisms[22]. As a result, several dedicated bioinformatic tools have been developed or extended for metaproteomic analysis[23–30]. 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[7,22,31–43].

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 comparison of established workflows composition is known a priori) have been used in individual benchmarking studies[44–46]. 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 multilaboratory 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 sample processing 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 as a foundation for future method development for metaproteomics.

Results

At the 3rd International Metaproteome Symposium in December 2018, individual laboratory outcomes of a collaborative, multilaboratory effort to compare metaproteomic workflows were presented. In this study, metaproteomics data was acquired in seven laboratories, 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" section, see Supplementary Data 1 for an overview of all methods).

At the Symposium, the decision was made to re-analyze the acquired data with different bioinformatics pipelines, to obtain a

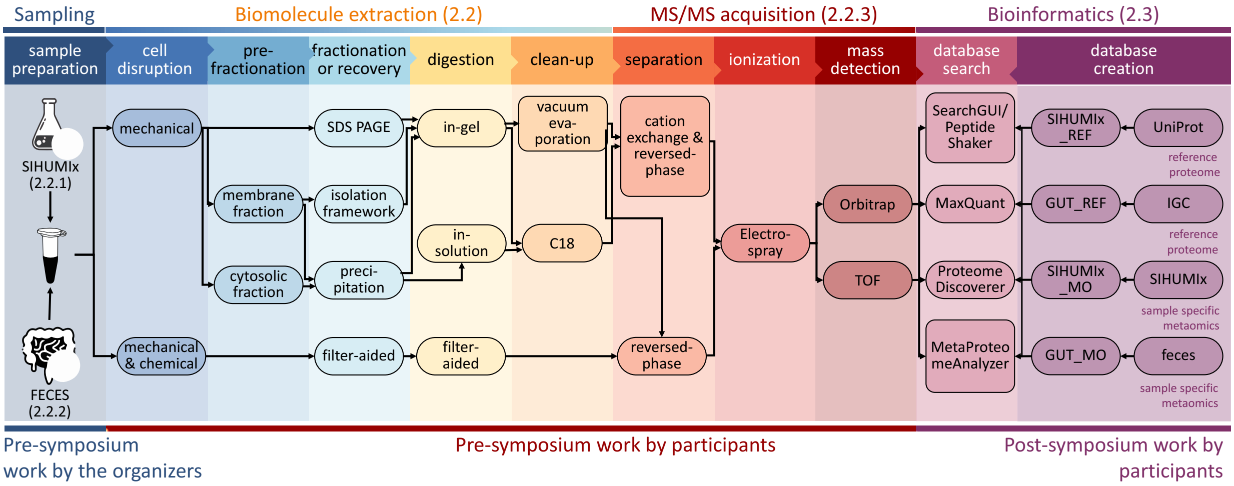


Fig. 1 Schematic representation of the main sample preparation steps and follow-up analyses of the CAMPI study. The figure consists of three parts: (i) Pre-symposium work by the organizers (left panel). The two samples (SIHUMIx and fecal sample) were, prior to the symposium, aliquoted and distributed over the participating laboratories. (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. The stock icons in the leftmost column were obtained from vecteezy.com, flaticon.com, and labicons.net.

multi-laboratory effort in metaproteomics to independently evaluate available methodological and computational approaches, in line with similar community-driven benchmarking studies[47–50]. 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 Fig. 1 for a general overview, and Fig. 2 for the results; see online Methods section 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 sample processing workflows and sample-specific meta-omic search databases lead to more identifications. In order to study the effect of the different sample processing and LC–MS/MS workflows on the identification outcome, we searched all submitted MS files using the widely used X!Tandem search engine[51]. 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 Fig. 2 (raw data in Supplementary Data 2).

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 47k 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 44k and 223k spectra were acquired, with identification rates between 11.99% and 34.79% for GUT_REF, and between 15.70% and 41.94% 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 (S03 and S04: 260 min, S05 and S06: 460 min, S08: 240 min, F01: 210 min, F02: 160 min), fractionation (S11, F07: 4 fractions), and additional separation methods such as MudPIT[52] (F01: 4 fractions) or ion mobility (PASEF)[53] (S13, F09) led to up to eight times more identified spectra, but at the cost of increased time and resources spent[54] (see Supplementary Data 1 for a detailed description, and Supplementary Data 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 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 the 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[16,18,20,44,55]. 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 Data 3).

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 characterized 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

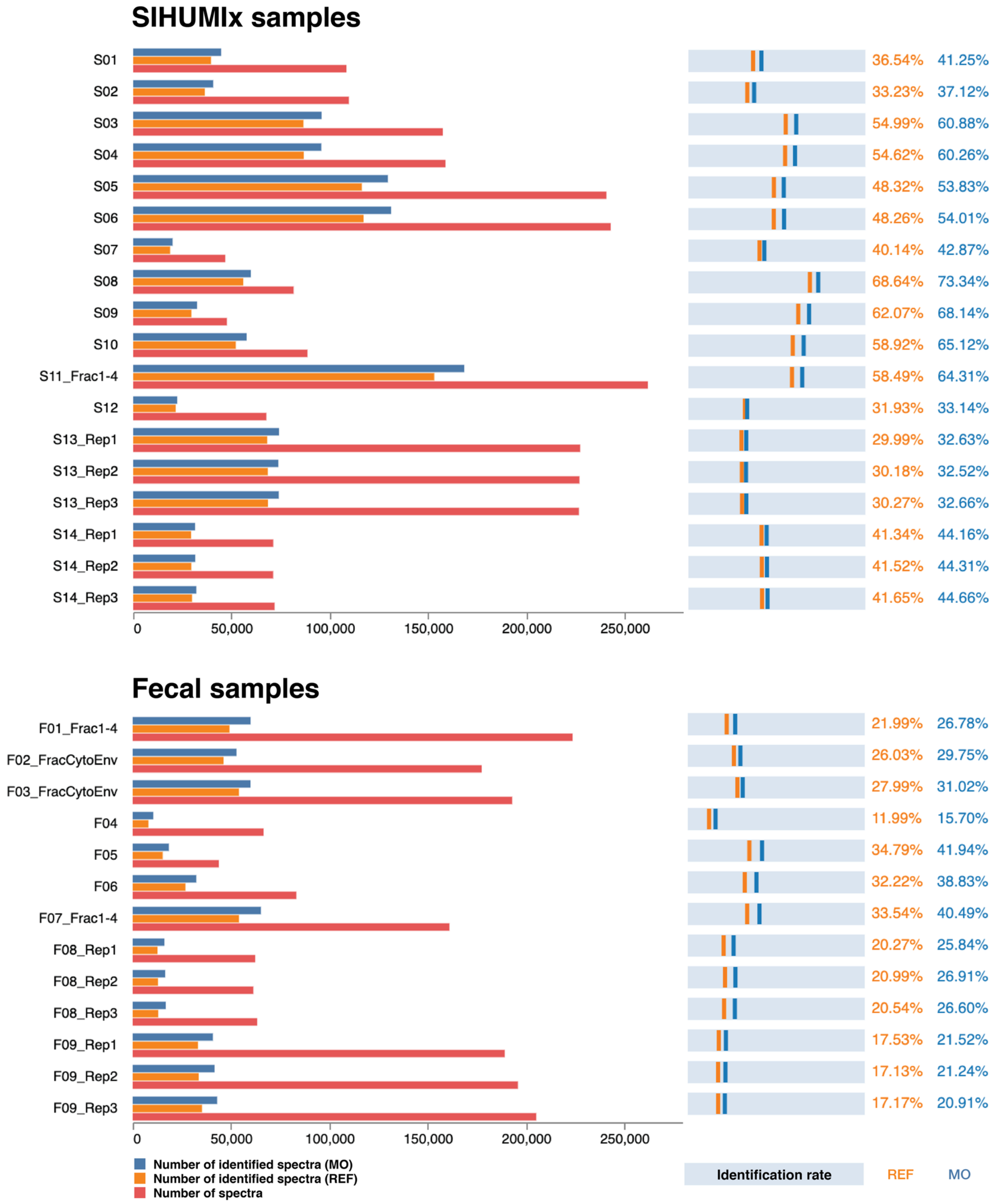


Fig. 2 Comparison of identification rates across all CAMPI workflows. On the left side, the bar charts show the number of identified spectra using the reference (REF) database (orange), the number of identified spectra using the multi-omic (MO) database (dark blue) and total amount of measured spectra (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 Data 1. For database searching, X!Tandem was used as a single search engine. Source data is provided in Supplementary Data 2.

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 analyzed 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) (Fig. 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).

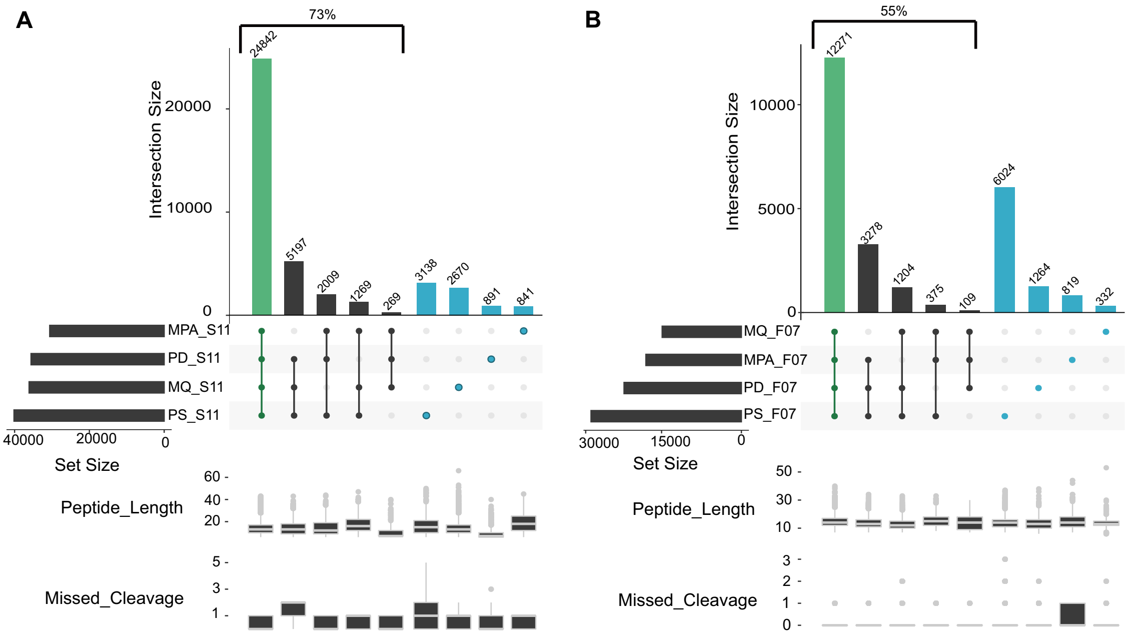


Fig. 3 UpSet plot comparison of identified sets of peptides using different bioinformatic pipelines. The left panel displays the results for the SIHUMIx sample S11 (**A**), while the right panel corresponds to the results for the fecal sample F07 (**B**). The four different bioinformatic pipelines (MetaProteomeAnalyzer (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. Source data is provided as a Source Data file.

For SIHUMIx, the majority of the identified peptides (54.2%) were found by all four bioinformatic pipelines (Fig. 3A), while this ratio dropped to 40% for the more complex fecal F07 sample (Fig. 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 SearchGUI/PeptideShaker, Proteome Discoverer, MPA and MaxQuant pipeline, 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[56].

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[57,58]. This effect may be attributable to algorithms with more sophisticated scoring methods (e.g., MS-GF+[59] 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[60].

Additionally, we compared the pipelines in terms of peptide features using the peptide lengths and the number of missed cleavages (lower panels of Fig. 3A, B). 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 sample processing 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 Data 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 Fig. 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.

At the peptide level (Fig. 4A, B), more complex workflows, such as those with longer gradient length and fractionation, identified the most peptides in general (as shown earlier in Fig. 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 3557 peptides (4.9% of all identified peptides) in the SIHUMIx data sets, and 2186 peptides (3.4% of all identified peptides) in the fecal data set. At

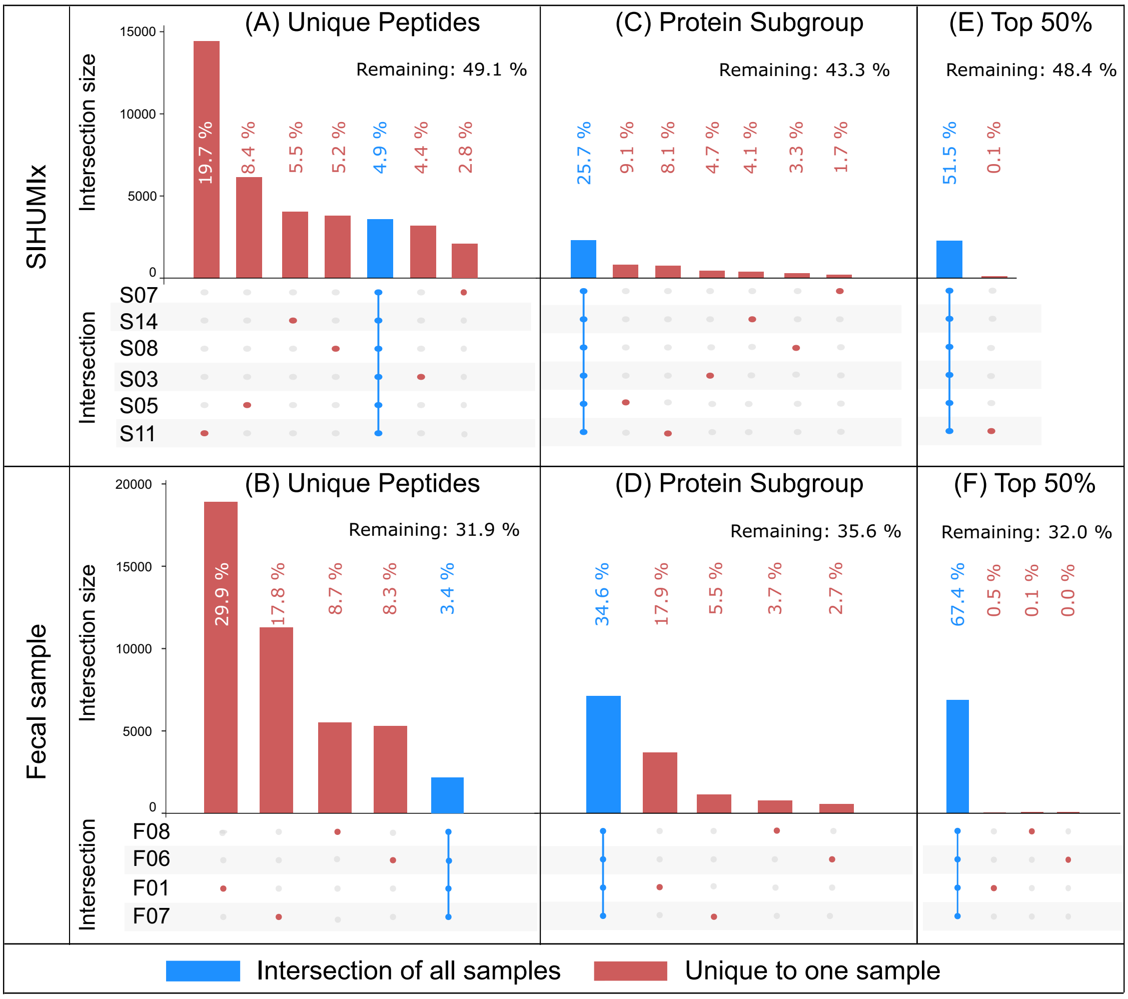


Fig. 4 UpSet plot comparison of sets of identified peptides, protein subgroups, and 50% most abundant protein subgroups. A, **B** Identified peptides, **C**, **D** all identified protein subgroups, and **E**, **F** top 50% subgroups (SIHUMIx and fecal sample, respectively). Top 50% protein subgroups were selected in terms of spectral count per subgroup. 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). Source data is provided as a Source Data file.

the protein subgroup level (Fig. 4C, 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 (Fig. 4E, 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 Figs. 2 and 3). 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 Fig. 4). 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[61] and MPA[28] (Supplementary Note 3). These two methods use different rules for protein inference: PAPPSO uses Occam's razor, and MPA uses anti-Occam's razor[62]. 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 (Supplementary Data 4 and 5). 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 (8802 and 8769) and subgroups (10,132 and 10,134), while substantial differences were found for the fecal samples (protein groups: 10,063 and 9712; subgroups: 17,576 and 21,973, for PAPPSO and MPA, respectively) (Supplementary Data 6). While cross-sample correlation (Supplementary Figs. 5 and 6) confirmed that the impact of bioinformatic pipelines on the analysis here was negligible, little else could be learned from this correlation analysis. To shed some light on these differences between protein grouping methods, we analyzed the agreement between samples for different grouping approaches (Supplementary Figs. 7 and 8). Notably, when applied to the fecal sample, the protein groups resulted in an unusually high number of 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 homologs, it could be more appropriate for single-organism proteomics or for taxonomically well-defined samples like SIHUMIx. In contrast, 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 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 sample processing workflows have an effect on the overall biological conclusions, we quantitatively compared the identified taxa for each selected sample from both data sets using spectral counts, and this at the peptide, the protein subgroup, and the sequencing read level.

We found different trends between the SIHUMIx and fecal samples (Figs. 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 (Fig. 5A and Supplementary Fig 9A, B). Interestingly, samples with more complex sample processing 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 (Fig. 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 Data 7), 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, *Clostridium butyricum* was not found by the read-based method, while *Clostridiales bacterium* and *Bacteroides dorei* were falsely found by the protein-centric method as these are not present in the SIHUMIx sample. However, these last two were both found at very low abundance. For completeness, the comparisons of community composition for SIHUMIx at the genus level were added in Supplementary Fig. 10.

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 (Fig. 6A). While the same families were identified, these had different proportions across methods (Fig. 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 (Fig. 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[45].

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 (Fig. 6A and Supplementary Fig. 11A, 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[63]. 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 only considered SIHUMIx species, therefore 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. Depending on the sample and the method used, the taxonomic resolution will thus vary. To better illustrate that, we compared the resolution across omes and across protein grouping methods (Supplementary Fig. 12A, B). We see that there is usually a drop of resolution either at the species (SIHUMIx) or the genus (Fecal) level and

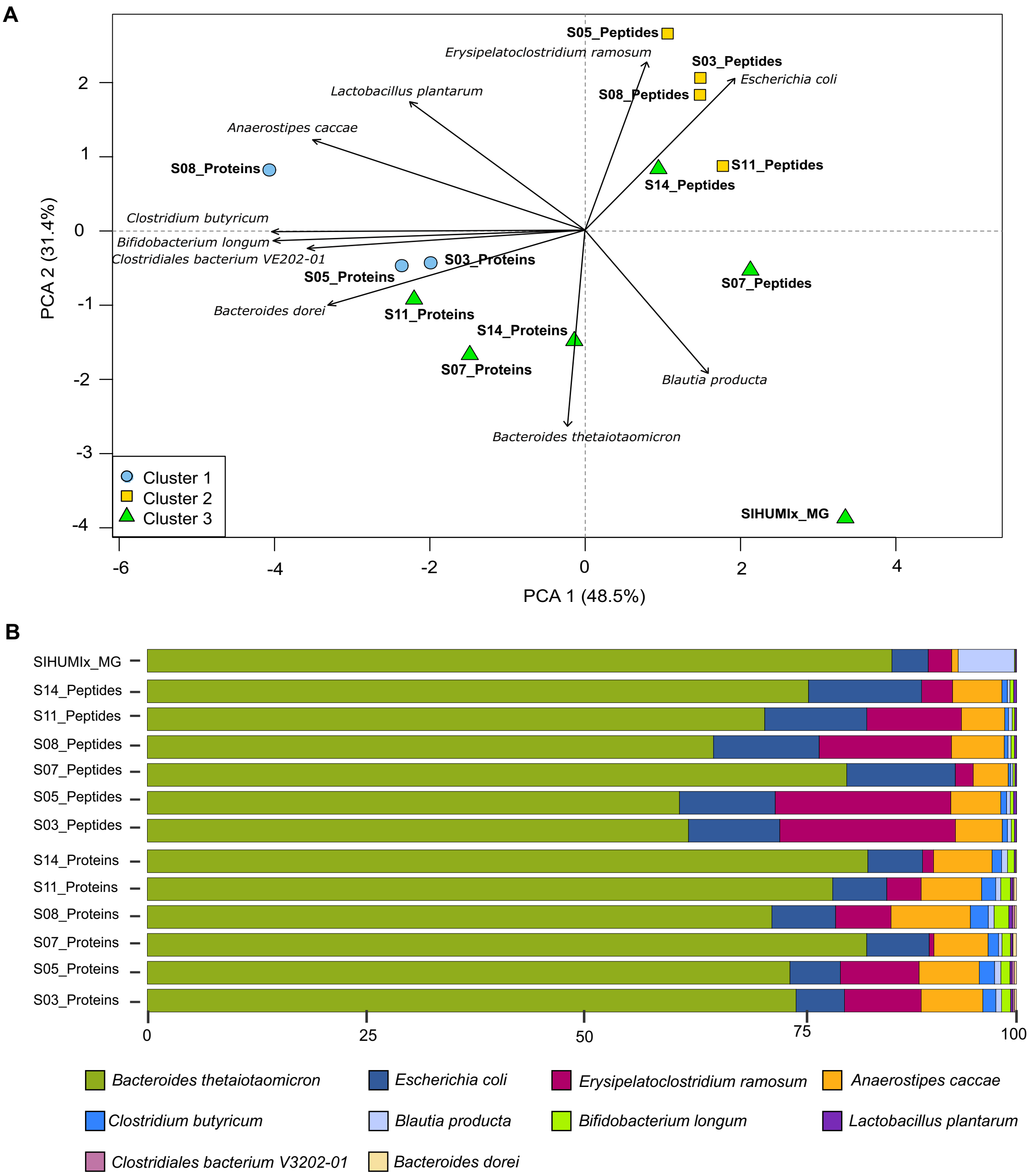


Fig. 5 Comparisons of community composition for SIHUMIx at the species level. The upper panel shows PCA clustering of the results (**A**). Different approaches and tools used for taxonomic annotation (MG - mOTU2, Peptides - Unipept, and Proteins - Prophane) are indicated in the label. Clusters ($k = 3$) 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**). Source data is provided as a Source Data file.

that the PAPPSO grouping method has a higher resolution for complex samples as already discussed in Supplementary Note 3.

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[63]. 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.

The functional profile is similar between different metaproteomics workflows. A major strength of metaproteomics is the ability to provide functional information that reflects the

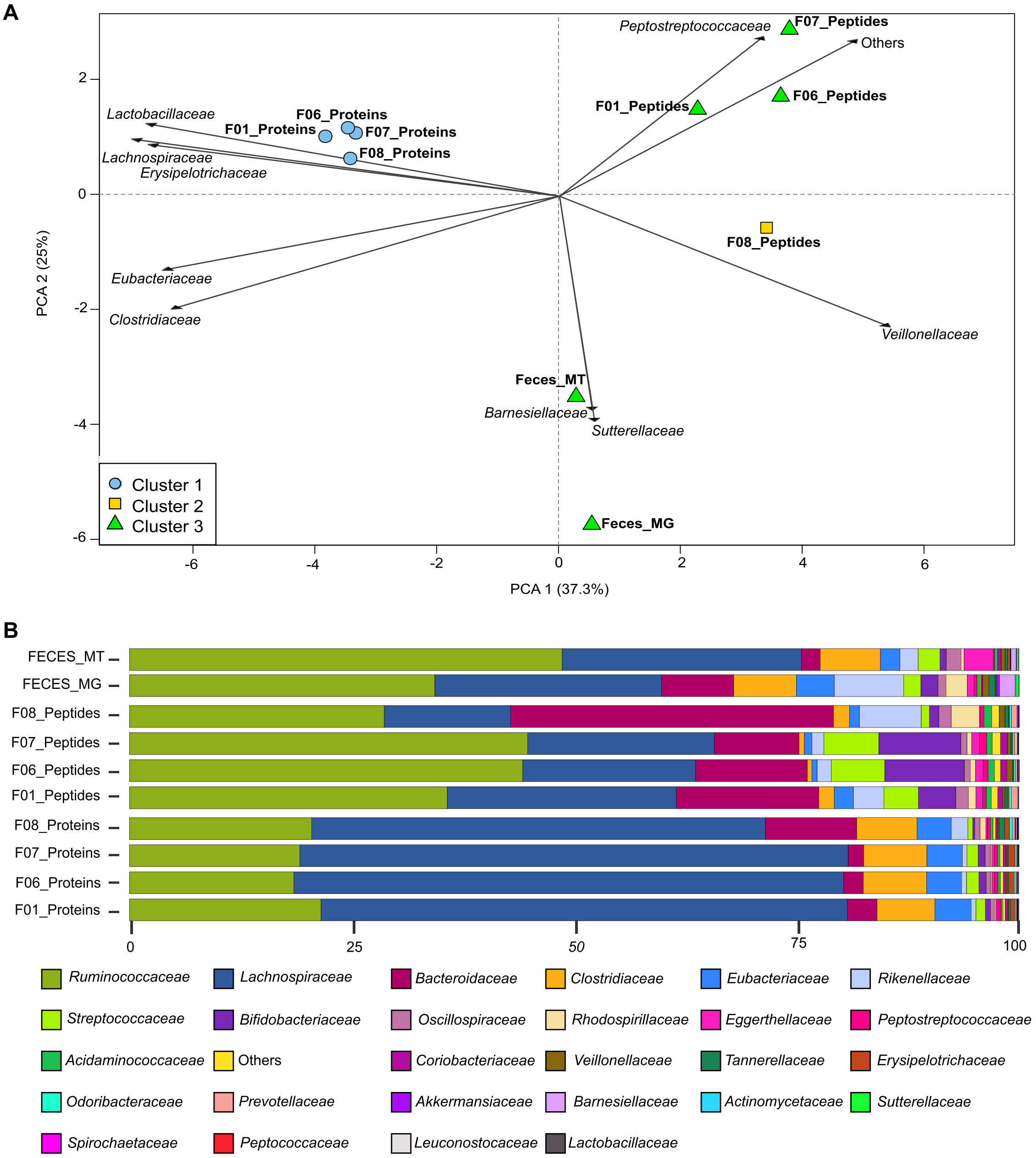
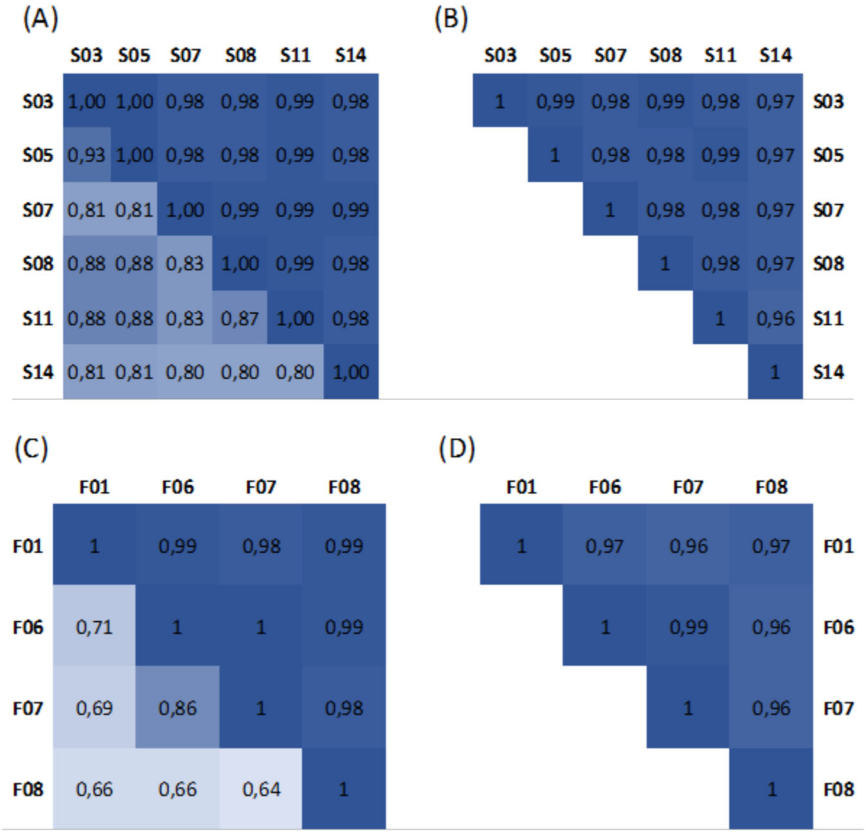


Fig. 6 Comparisons of community composition for fecal data sets. The upper panel shows PCA clustering of the results (**A**). Different approaches and tools used for taxonomic annotation (MG - mOTU2, Peptides - Unipept, and Proteins - Prophane) are indicated in the label. Clusters ($k = 3$) were calculated 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**). Source data is provided as a Source Data file.

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 (Fig. 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[64], 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 Fig. 4: while peptide identifications may differ greatly between samples, the underlying biological meaning reflected by functional annotations are highly similar across different analysis workflows. Moreover, while some more elaborate data measurements yield unique peptides, these peptides do not translate into more functional pathways being identified (Supplementary Fig. 13) and usually correspond to very low abundant proteins, identified with only one peptide (as already shown in Supplementary Fig. 4).

In contrast, a comparison between the different omics domains showed important differences in terms of functional profile.



(A)

	S03	S05	S07	S08	S11	S14
S03	1,00	1,00	0,98	0,98	0,99	0,98
S05	0,93	1,00	0,98	0,98	0,99	0,98
S07	0,81	0,81	1,00	0,99	0,99	0,99
S08	0,88	0,88	0,83	1,00	0,99	0,98
S11	0,88	0,88	0,83	0,87	1,00	0,98
S14	0,81	0,81	0,80	0,80	0,80	1,00

(B)

S03	S05	S07	S08	S11	S14	
1	0,99	0,98	0,99	0,98	0,97	S03
	1	0,98	0,98	0,99	0,97	S05
		1	0,98	0,98	0,97	S07
			1	0,98	0,97	S08
				1	0,96	S11
					1	S14

(C)

	F01	F06	F07	F08
F01	1	0,99	0,98	0,99
F06	0,71	1	1	0,99
F07	0,69	0,86	1	0,98
F08	0,66	0,66	0,64	1

(D)

F01	F06	F07	F08	
1	0,97	0,96	0,97	F01
	1	0,99	0,96	F06
		1	0,96	F07
			1	F08

Fig. 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. Source data is provided as a Source Data file.

Notably, metagenomics and metaproteomics are particularly different from each other, while metatranscriptomics tends to overlap better with metagenomics, highlighting once more the need for integrated meta-omics approaches (Supplementary Figs. 14–16)[32].

Discussion

In this founding edition of CAMPI, we used both a simplified, laboratory-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[45]. 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[65,66]. 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" section). While it is recommended that researchers use well-annotated matched metagenomes for optimal metaproteomics analysis, not all study designs have metagenomics information available. For such studies, iterative search approaches on publicly available repositories are available[25,67–70], some of which address the issue of controlling the false discovery rate of identifications[68]. Moreover, other platforms such as iMetaLab[30] have been widely used for human and mouse gut metaproteomics analysis. We have not used the iterative search approaches or alternative platforms for this study, although the availability of the data should encourage users to evaluate the performance of these approaches.

This CAMPI study has highlighted that there is room for future editions of CAMPI studies. Indeed, based on the issues identified in this study, we can already define interesting future research questions: what is the effect of data set complexity, and how do other sample types such as marine sediments affect the results; 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 the 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. Label-based approaches could also be extremely valuable for the field as it has been shown that stable isotope labeling as a spike-in reference can strongly improve quantification accuracy[71,72]. 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 that aims, among other things, to standardize and accelerate experimental and bioinformatic methodologies in this field. This initiative can combine forces with existing initiatives such as the ABRF iPRG study group, who recently provided a metaproteomics data set to be analyzed by the proteomics informatics community[73]. 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

Ethics. Written informed consent was obtained from the subject enrolled in the study. This study was approved by the ethics committee of the University Magdeburg (reference no. 99/10).

Sample description

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[74]. SIHUMIx was prepared as previously described, with an additional 24 h of cultivation of one control bioreactor, to produce sufficient biomass to be sent out to each participating laboratory[74]. Participants received 3.5 × 10^9 cells/ml of frozen sample (−20 °C) in dry ice.

Human fecal microbiome sample. A natural human fecal microbiome sample was procured from a 33-years-old omnivorous, non-smoking woman. The sample was immediately homogenized, treated with RNA-later, aliquoted, frozen, and stored at −20 °C until aliquots were sent to each participating laboratory.

Biomolecule extraction and nucleotide sequencing

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. Preprocessing of the sequenced reads was performed as part of the Integrated Meta-omic Pipeline (IMP)[75] 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[76] 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[77] and MaxBin 2.2.6[78] with default parameters and minimum contig length of 2500 and 1500 bps, respectively. Bins were refined using DASTool 1.1.2[79] 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[80] as part of the DASTool suite.

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 (Fig. 1). Key characteristics for each workflow can be found in Supplementary Data 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)[81] (workflows 1–3, 5, 7–9, 11, 12, 19–23 in Supplementary Data 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 1/10,000, with digestion times from 2 to 16 h. 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).

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 Data 1. For LC, all laboratories separated peptides using reversed-phase chromatography with a linear gradient length ranging from 60 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[82].

Six groups used an Orbitrap mass spectrometer (4× Q Exactive HF, 1× Q Exactive Plus, 1× Fusion Lumos, ThermoFisher Scientific), while two groups employed a timsTOF mass spectrometer (Bruker Daltonik). All participants used data-dependent acquisition (DDA) with exclusion duration times ranging from 10 to 60 s. All MS proteomics data and X!Tandem results have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[83].

Bioinformatics

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[84] except for *Blautia producta*, for which the whole genus *Blautia* was taken (SIHUMIx_REF). The IGC 9.9 database[85] (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 2719 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).

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[51] (Alanine, 2017.02.01) was used as a 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[86].

The following database search engines were used for the pipeline comparison: (i) MaxQuant[87] (including the search engine Andromeda) (ii) Galaxy-P workflows[88,89] consisting of SearchGUI[90,91] (using OMSSA[92], X!Tandem[51], MS-GF+[59], and Comet[93]) and PeptideShaker[94] to merge the results, (iii) MetaProteomeAnalyzer[28] (server version 3.4, using X!Tandem and OMSSA), and (iv) ProteomeDiscoverer 2.2 (using SequestHT, from ThermoFisher). 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).

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[28] and PAPPSO[61], and analyses were made on protein groups and subgroups (Supplementary Note 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[61], which excludes proteins based on the rule of maximum parsimony, and grouping from MPA[28], 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 3). These algorithms are also implemented in Pout2Prot[95] for independent use.

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[24,27,96]. 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 4). For the taxonomic annotation of the fecal data sets with Unipept, the desktop[96] and CLI[23,97] versions were used. In both analyses for SIHUMIx and the fecal data sets, isoleucine (I) and leucine (L) were equated. The assigned taxonomies for each of the peptides can be found in Supplementary Data 8 and 9.

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[64].

Prophane was used for the protein-centric approach[98,99]. 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)[100], while two functional annotation tasks where performed against the eggNOG (database version 4.5.1)[101] and Pfam-A (db version 32) databases[102] using eggNOG-mapper[103,104] and hmmscan[105], 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 Data 10–12, and the assigned taxonomies for each of the proteins can be found in Supplementary Data 13 and 14.

Metagenomic and metatranscriptomic reads were both taxonomically annotated with the mOTUs profiler v 2.0[106] 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 normalized values as described in detail below.

Comparison between omics domains—taxonomic resolution. Taxonomic annotations from the Prophane protein group outputs were used for metaproteomics. This method uses only identified proteins and assesses annotations based on the LCA approach thus generating results for each protein at the best possible taxonomic resolution

The mOTU2 profiler used for the metagenomic taxonomic annotation takes advantage of marker genes for taxonomic annotation and thus annotates everything at the OTU level. Since this approach does not allow comparison at each taxonomic level, Kraken2[107] was used to compare taxonomic resolution across omics domains. Kraken2 was run on the sequencing reads with the maxikraken2_1903 database and a confidence threshold set to 0.7.

Comparison between omics domains - functional comparison. Each sequence database (SIHUMIx_REF, SIHUMIx_MO, and GUT_MO) was annotated with the Mantis[108] tool for consensus-driven protein annotation. For metaproteomics, abundance from Prophane outputs and annotation from Mantis were used to generate functional profiles. For metagenomics and metatranscriptomics, sequencing reads were mapped against the assembly contigs using bowtie2[109] and ORFs abundance was calculated using featureCounts[110] KEGG[111] annotations were retrieved from Mantis and used to compare functional profiles across omes.

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[112]. 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 4 and Supplementary Data 15 and 16). 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.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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. Assemblies and raw metagenomic and metatranscriptomic reads are available through the European Nucleotide Archive under the study accession number PRJEB42466. Source data are provided with this paper.

Code availability

All scripts and intermediary files are made available on github.com/metaproteomics/CAMPI[113].

Received: 2 April 2021; Accepted: 24 November 2021;
Published online: 15 December 2021

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**, 200115–19 (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. *Anal. Chem.* **85**, 4203–4214 (2013).
4. Rodriguez-Valera, F. Environmental genomics, the big picture? *FEMS Microbiol. Lett.* **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. *Environ. Microbiol.* **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. et al. Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis. *Expert Rev. Proteom.* **16**, 375–390 (2019).
9. Wang, D.-Z., Kong, L.-F., Li, Y.-Y. & Xie, Z.-X. Environmental microbial community proteomics: status. *Chall. Perspect. 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. Zhang, X. et al. Assessing the impact of protein extraction methods for human gut metaproteomics. *J. Proteom.* **180**, 120–127 (2018).
14. Wöhlbrand, L. et al. Impact of extraction methods on the detectable protein complement of metaproteomic analyses of marine sediments. *Proteomics* **17** 1700241 (2017).
15. Heyer, R. et al. Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **261**, 24–36 (2017).
16. Tanca, A. et al. The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* **4**, 227 (2016).
17. 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).
18. Muth, T. et al. Navigating through metaproteomics data: a logbook of database searching. *Proteomics* **15**, 3439–3453 (2015).
19. 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).
20. Colaert, N., Degroeve, S., Helsens, K. & Martens, L. Analysis of the resolution limitations of peptide identification algorithms. *J. Proteome Res.* **10**, 5555–5561 (2011).
21. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteom.* **4**, 1419–1440 (2005).
22. Heyer, R., Kohrs, F., Reichl, U. & Benndorf, D. Metaproteomics of complex microbial communities in biogas plants. *Microb. Biotechnol.* **8**, 749–763 (2015).
23. Verschaffelt, P. et al. Unipept CLI 2.0: adding support for visualisations and functional annotations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa553 (2020).
24. Gurdeep Singh, R. et al. Unipept 4.0: Functional Analysis of Metaproteome Data. *J. Proteome Res* **18**, 606–615 (2019).
25. Park, S. K. R. et al. *ComPIL 2.0: An Updated Comprehensive Metaproteomics Database* (2019).
26. Sajulga, R. et al. Survey of metaproteomics software tools for functional microbiome analysis. *PLoS ONE* **15**, e0241503 (2020).
27. Van Den Bossche, T. et al. Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for seamless end-to-end metaproteomics data analysis. *J. Proteome Res.* **19**, 3562–3566 (2020).
28. Muth, T. et al. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *J. Proteome Res.* **14**, 1557–1565 (2015).
29. 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. *Front. Microbiol.* **10**, 1883 (2019).
30. Liao, B. et al. iMetaLab 1.0: a web platform for metaproteomics data analysis. *Bioinformatics* **34**, 3954–3956 (2018).
31. Zhang, X. et al. Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease. *Nat. Commun.* **9**, 2873 (2018).
32. Heintz-Buschart, A. et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2**, 16180 (2017).
33. Erickson, A. R. et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE* **7**, e49138 (2012).
34. Juste, C. et al. Bacterial protein signals are associated with Crohn's disease. *Gut* **63**, 1566–1577 (2014).
35. 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. *J. Proteom.* **198**, 50–58 (2019).
36. Schneider, T. et al. Proteome analysis of fungal and bacterial involvement in leaf litter decomposition. *Proteomics* **10**, 1819–1830 (2010).
37. Teeling, H. et al. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* **336**, 608–611 (2012).
38. Morris, R. M. et al. Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J.* **4**, 673–685 (2010).
39. Petersen, J. M. et al. Chemosynthetic symbionts of marine invertebrate animals are capable of nitrogen fixation. *Nat. Microbiol.* **2**, 725 (2017).
40. 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. USA* **109**, E1173–82 (2012).
41. 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).
42. Heyer, R. et al. Metaproteome analysis reveals that syntrophy, competition, and phage-host interaction shape microbial communities in biogas plants. *Microbiome* **7**, 69 (2019).
43. 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).
44. Tanca, A. et al. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PLoS ONE* **8**, e82981 (2013).
45. Kleiner, M. et al. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat. Commun.* **8**, 6 (2017).
46. 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. *Front. Microbiol.* **10**, 238 (2019).
47. Mangul, S. et al. Systematic benchmarking of omics computational tools. *Nat. Commun.* **10**, 157 (2019).
48. Collins, B. C. et al. Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.* **8**, 291 (2017).
49. Bell, A. W. et al. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **6**, 423–430 (2009).

50. Sczyrba, A. et al. Critical assessment of metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
51. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466–1467 (2004).
52. Cox, J. & Mann, M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **80**, 273–299 (2011).
53. Meier, F. et al. Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteom.* **17**, 2534–2545 (2018).
54. 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).
55. 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).
56. 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).
57. Park, G. W. et al. Integrated proteomic pipeline using multiple search engines for a proteogenomic study with a controlled protein false discovery rate. *J. Proteome Res.* **15**, 4082–4090 (2016).
58. Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L. & Deutsch, E. W. Combining results of multiple search engines in proteomics. *Mol. Cell. Proteom.* **12**, 2383–2393 (2013).
59. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
60. 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* **20**, e1900351 (2020).
61. 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).
62. Martens, L. & Hermjakob, H. Proteomics data validation: why all must provide data. *Mol. Biosyst.* **3**, 518–522 (2007).
63. Pible, O. et al. Estimating relative biomasses of organisms in microbiota using 'phylopeptidomics'. *Microbiome* **8**, 30 (2020).
64. Verschaffelt, P. et al. MegaGO: a fast yet powerful approach to assess functional similarity across meta-omics data sets. *J. Proteome Res.* **20**, 2083–2088 (2021).
65. Serang, O. & Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Stat. Interface* **5**, 3–20 (2012).
66. Huang, T., Wang, J., Yu, W. & He, Z. Protein inference: a review. *Brief. Bioinformatics* **13**, 586–614 (2012).
67. Jagtap, P. et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13**, 1352–1357 (2013).
68. Kertesz-Farkas, A., Keich, U. & Noble, W. S. Tandem mass spectrum identification via cascaded search. *J. Proteome Res.* **14**, 3027–3038 (2015).
69. Potgieter, M. G. et al. MetaNovo: a probabilistic approach to peptide and polymorphism discovery in complex metaproteomic datasets. Preprint at *bioRxiv* https://doi.org/10.1101/605550 (2019).
70. Kumar, P. et al. A sectioning and database enrichment approach for improved peptide spectrum matching in large, genome-guided protein sequence databases. *J. Proteome Res.* **19**, 2772–2785 (2020).
71. Zhang, X. et al. In vitro metabolic labeling of intestinal microbiota for quantitative metaproteomics. *Anal. Chem.* **88**, 6120–6125 (2016).
72. von Bergen, M. et al. Insights from quantitative metaproteomics and protein-stable isotope probing into microbial ecology. *ISME J.* **7**, 1877–1885 (2013).
73. Davis, D. L., Palmblad, M. & Weintraub, S. T. iPRG 2019 metaproteomics study. *J. Biomol. Tech.* **30**, S53 (2019).
74. 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).
75. Narayanasamy, S. et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).
76. 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).
77. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
78. 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).
79. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
80. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
81. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).
82. Wolters, D. A., Washburn, M. P. & Yates, J. R. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).
83. 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).
84. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
85. Li, J. et al. An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
86. 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).
87. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
88. Jagtap, P. D. et al. Metaproteomic analysis using the Galaxy framework. *Proteomics* **15**, 3553–3565 (2015).
89. Blank, C. et al. Disseminating metaproteomic informatics capabilities and knowledge using the galaxy-P framework. *Proteomes* **6**, 7 (2018).
90. 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).
91. Barsnes, H. & Vaudel, M. SearchGUI: a highly adaptable common interface for proteomics search and de novo engines. *J. Proteome Res.* **17**, 2552–2555 (2018).
92. Geer, L. Y. et al. Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
93. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
94. Vaudel, M. et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **33**, 22–24 (2015).
95. Van Den Bossche, T. et al. Pout2Prot: an efficient tool to create protein (sub) groups from Percolator output files. Preprint at https://doi.org/10.1101/2021.08.11.455803 (2021).
96. 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.* **20**, 4 (2021).
97. Mesuere, B. et al. The Unipept metaproteomics analysis pipeline. *Proteomics* **15**, 1437–1442 (2015).
98. Schiebenhoefer, H. et al. A complete and flexible workflow for metaproteomics data analysis based on MetaProteomeAnalyzer and prophane. *Nat. Protoc.* **362**, 776 (2020).
99. Schneider, T. et al. Who is who in litter decomposition? Metaproteomics reveals major microbial players and their biogeochemical functions. *ISME J.* **6**, 1749–1762 (2012).
100. 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 Res.* **35**, D61–D65 (2007).
101. Jensen, L. J. et al. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36**, D250–D254 (2007).
102. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
103. 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).
104. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
105. Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).
106. Milanese, A. et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
107. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
108. Queirós, P., Delogu, F., Hickl, O., May, P. & Wilmes, P. Mantis: flexible and consensus-driven genome annotation. *Gigascience* **10**, 6 (2021).
109. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
110. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
111. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

112. 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).

113. Van Den Bossche, T., Kunath, B. J. & Schallert, K. Critical assessment of metaproteome investigation - a multi-lab comparison of established workflows. *Zenodo* https://doi.org/10.5281/zenodo.5588376 (2021).

Acknowledgements

This work has benefited from collaborations facilitated by the Metaproteomics Initiative (https://metaproteomics.org/) whose goals are to promote, improve and standardize metaproteomics. 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 were carried out using the high-performance computing facilities of the University of Luxembourg (https://hpc.uni.lu). This work was supported by the Research Foundation - Flanders (FWO) [grant no. 1S90918N (S.B.) to T.V.D.B.; 12I5217N to B.M.; G042518N to L.M.]; by a FEBS Summer Fellowship [to T.V.D.B.]; by the European Union's Horizon 2020 Program (H2020-INFRAIA-2018-1) [823839 to L.M.]; by the FEMS [R.T.G. to S.S.S.]; by the Norwegian Centennial Chair program [to T.J.G., P.D.J., and M.A.]; the Novo Nordisk Foundation grant NNF20OC0061313 to M.A.; the USDA National Institute of Food and Agriculture Hatch project [1014212 to M.K.]; the U.S. National Science Foundation [OIA 1934844 and IOS 2003107 to M.K.]; the Foundation for Food and Agriculture Research [Grant ID: 593607 to M.K.]; the Agence Nationale de la Recherche [ANR-17-CE18-0023-01 to G.M., O.P., J.A.]; Deutsche Forschungsgemeinschaft (DFG) [RE3474/5-1 and RE3474/2-2 to S.F., T.M., B.Y.R.]. Research by T.J.G., P.D.J, E.L was 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.; and the National Institutes of Health R01-DK70977 to RLH. 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 B.J.K., P.M. and P.W, and the European Research Council (ERC-CoG 863664) to P.W.

Author contributions

T.V.D.B., B.J.K., K.S. and S.S.S. contributed equally as the main authors of the manuscript; doing the data curation, data analysis and visualization, and writing and editing the manuscript. N.J. and D.B. started and supervised the initial laboratory intercomparison in the context of the 3rd International Metaproteomics Symposium. L.M. and T.M. started and jointly supervised the process of turning the intercomparison study into a coherent manuscript and supervised the writing. T.V.D.B., B.J.K., K.S., S.S.S., L.M., T.M., N.J. and D.B. contributed to the data selection, concept, and structure of the manuscript. B.J.K., S.S.S., J.-P.T, R.H., M.W., R.H., M.K., M.J., G.M., O.P., J.A., R.L.H., S.L.P., R.J.G., C.J., C.H., M.V.B., P.E.A. and T.L. did laboratory work for one of the eight independent laboratories, which included the sample provision, distribution and preparation, protein extraction and proteolytic digestion, mass spectrometry and nucleotide sequencing. T.V.D.B., K.S., P.J., R.H., M.W., M.Ø.A., L.H.H., H.S., S.F., M.J., O.P., S.L.P., R.J.G., A.B., K.T., E.L., A.S., P.T.Q., P.V., P.M., B.M., P.E.A. and O.L. did bioinformatics analyses that were used in the manuscript, which included the generation of protein databases from nucleotide sequences, mass spectrometry data processing and initial protein database search, further data analysis, and meta-analysis and visualization. J.A., R.L.H., P.J., C.L., M.K., O.P., R.H., P.T.Q., B.M., A.T., S.F. and P.M. reviewed and edited the manuscript and participated in discussions on the course of individual analysis. J.A., R.L.H., P.J., C.J., M.K., M.Ø.A., U.R., B.Y.R., S.U., M.V.B., P.W. and T.J.G. had a supervisory role and reviewed the manuscript. The four main authors T.V.D.B., B.J.K., K.S. and S.S.S. are listed at the beginning, while the two supervising authors L.M. and T.M. are listed at the end. All other co-authors are listed alphabetically.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-27542-8.

Correspondence and requests for materials should be addressed to Lennart Martens.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

[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, Esch-sur-Alzette, Luxembourg. [4]Bioprocess Engineering, Otto-von-Guericke University Magdeburg, Magdeburg, Germany. [5]Department of Molecular Systems Biology, Helmholtz-Centre for Environmental Research - UFZ GmbH, Leipzig, Germany. [6]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [7]Département Médicaments et Technologies pour la Santé (DMTS), Université Paris Saclay, CEA, INRAE, SPI, 30200 Bagnols-sur-Cèze, France. [8]Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences (NMBU), Ås, Norway. [9]INRAE, AgroParisTech, Micalis Institute, Université Paris-Saclay, 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, Minneapolis, 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]Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg. [22]Section eScience (S.3), Federal Institute for Materials Research and Testing, Berlin, Germany. [23]These authors contributed equally: Tim Van Den Bossche, Benoit J. Kunath, Kay Schallert, Stephanie S. Schäpe. [24]These authors jointly supervised this work: Lennart Martens, Thilo Muth. ✉email: lennart.martens@ugent.be