Abstract

Introduction: Metaproteomics is an established method to obtain a comprehensive taxonomic and functional view of microbial communities. After more than a decade, we are now able to describe the promise, reality, and perspectives of metaproteomics and provide useful information about the choice of method, applications, and potential improvement strategies.

Areas covered: In this perspective, we will discuss current challenges of species and proteome coverage, and also highlight functional aspects of metaproteomics analysis of microbial communities with different levels of complexity. To do this, we re-analyzed data from microbial communities with low to high complexity (8, 72, 200 and >300 species). High species diversity leads to a reduced number of protein group identifications in a complex community, and thus the number of species resolved is underestimated. Ultimately, low abundance species remain undiscovered in complex communities. However, we observed that the main functional categories were better represented within complex microbiomes when compared to species coverage.

Expert opinion: Our findings showed that even with low species coverage, metaproteomics has the potential to reveal habitat-specific functional features. Finally, we exploit this information to highlight future research avenues that are urgently needed to enhance our understanding of taxonomic composition and functions of complex microbiomes.

Article highlights

- We integrated four microbial community datasets to determine the effect of increasing community complexity on proteome, species and pathway coverage
- The taxonomic resolution is reduced in microbial communities with increasing complexity
- The identification of low abundance proteins present in complex microbial communities is challenging
- A unique strength of metaproteomics is the robust identification of habitat specific pathways, regardless of the underlying microbial community complexity

1. Introduction – A decade of metaproteomics

In nature, bacteria rarely occur axenically but are rather found in microbial communities that exhibit complex interactions and niche formations [1]. Microbial communities not only play a primary role in global biogeochemical cycling to make our planet habitable [2] but also form complex interactions with other organisms that are crucial for the development and maintenance of health in animals [3] and humans [4]. To characterize microbial communities and identify how they can potentially affect the host or the environment, it is common to profile the taxonomy and functionality of such communities [5].

The characterization of microbial community structures from assessing taxonomic marker gene profiles has been a widely performed and accepted practice over the past decade of microbiome research [6]. For example, the hypervariable regions of the 16S ribosomal RNA (16S rRNA) gene are often used as a targeted gene marker [7]. Profiling using 16S rRNA gene sequencing has shown that reproducibility can be strongly biased because of factors such as the genomic DNA extraction method, PCR primer selection, sequencing read length and the sequencing platform used [8]. Moreover, this approach is limited to the determination of taxonomic distribution and is not generally suitable for analysis of the actual functions of the community [9]. To alleviate this, bioinformatics toolsets such as PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states) [10] and Piphillin (improved prediction of metagenomics content by direct inference from human microbiomes) [11] have been developed to provide functional predictions based on 16S rRNA gene data of a microbial community. In these methods, a pseudo-metagenome is constructed using the 16S rRNA gene profiling data by picking genomes from a database containing known and sequenced bacterial genomes [10]. This approach has the drawback that the genomes of many taxa identified in the 16S rRNA gene sequencing data have not yet been whole genome sequenced or are not yet fully annotated in the available databases. Therefore, for these cases, the closest phylogenetically-related

specimen is selected instead. To achieve a more realistic picture of the functional potential of a community, next-generation sequencing can be used to analyze the whole metagenome of microbial communities [12]. The cost of metagenomics is still high per sample when compared to 16S rRNA gene profiling, but the associated costs are steadily decreasing and therefore metagenomics is being increasingly applied in microbiology studies [13].

Although the functional capacity of a given community can be investigated through analyzing the gene content, measuring the proteins as expression of genes is arguably more important for characterization of community functionality [14]. This has motivated subsequent studies to focus more on the proteome, since proteins are involved in metabolic processes and ultimately responsible for cellular functions wherefore proteomics has been established as an indispensable approach to study the complete protein inventory of a given species [15]. In 2004, this technique was applied to a microbial consortium for the first time and coined ‘metaproteomics’ by Wilmes and Bond [16, 17]. Currently, metaproteomics has developed into a widely practiced technique and offers the possibility of acquiring a comprehensive picture of the community structure and function. Moreover, it can be used to determine the microbial community interactions with external substrates or host metabolites [18]. However, it should be noted that analyzing a community via a combinatorial approach employing both metagenomics and metaproteomics can be even more successful for unraveling the composition of a given community [19]. This has been also shown in a recent study where a multi-omics approach was applied on a defined, evenly distributed mock microbial community [20]. Moreover, the increasing availability of metagenomes allows for the construction of smaller and more specific protein databases [21, 22] and therefore more accurate protein identifications.

Although the analysis of proteins present in a microbial community can provide information on the general functions performed by the consortium as a whole, it is also pivotal to determine which species are carrying out these functions and thus elucidate the active key players in the community [23]. To determine the cellular activity of a microbe in a community, specific activity

tests need to be applied. Targeted or untargeted metabolomics can assess the substrates and products of all metabolic enzymes, and therefore serve as a suitable approach to determine overall community activity [24]. Since metabolomics is very sensitive for present metabolites in a community, it can help to discover unknown metabolic functions as a supportive strategy for protein-based approaches [25]. Moreover, the inclusion of metabolic flux analysis can help to verify enriched functions identified in metaproteomics analysis [26]. However, this approach is limited by the rapid turnover of metabolites, which can lead to quenching of enzymatic reactions, and because metabolites are often not derived from the microbial community itself but rather from the host environment [27]. It is also challenging to determine the taxonomic origin of specific metabolites, unlike for proteins, of which the amino acid sequences can be traced back to the genomes of the microorganisms. Moreover, many of the cellular active proteins are structural proteins, regulatory factors, or proteins that interact with other proteins rather than catalyzing metabolic processes, which cannot be assessed by metabolomics [28]. In order to avoid the need for thousands of specific activity assays, the introduction of specific labelled substrates or nutrients into microbial communities and their subsequent incorporation into the biomass can determine species activity by methods such as protein-based stable isotope probing (protein-SIP) [29, 30] or protein-based stable isotope fingerprinting (protein-SIF) [31, 32].

The importance of metaproteomics in studying microbial communities has manifested during the last decade, as the number of publications utilizing the technique have increased about 8.9-fold (**Figure S1**). There has also been a consistent increase in the number of identified proteins (not peptide spectra matches, PSMs) each year. Particularly, there has been an exponential increase over the last four years, which is mainly due to the use of newer MS-instruments with higher resolution (**Figure S2**). Indeed, tandem mass spectrometry has been further developed to improve the accuracy and sensitivity of the instrument, and has become the principal high-throughput technology in metaproteomics studies [33-35]. The obtained metaproteome coverage, i.e., the limiting factor for optimally characterizing a microbial community, is, amongst other

factors, directly linked to the speed and accuracy of the MS-technology. It is worthwhile to consider these recent technological developments in order to assess the actual power of data acquisition for metaproteomics analyses, and such recent technological advances are discussed in detail in a former study [36]. Although the technological status of the MS-instruments is a notable influencing factor for the success of protein identifications, here we will be focusing on the database and bioinformatics issues encountered when analyzing microbial communities with increasing complexity, which is one of the main challenges of metaproteomics studies.

2. Metaproteomics for the characterization of complex communities

Since the analysis of proteomes complements other omics disciplines such as metagenomics, metatranscriptomics and metabolomics; metaproteomics has become a widely applied technique for building a comprehensive picture of the structure and functionality of microbial communities on a large scale [37]. Analyzing the metaproteome of a community has the unique strength, compared to other omics techniques, to characterize the covered metabolic pathways for the identification of habitat specific functions [38]. Further, in contrast to DNA based omics approaches, metaproteomics can also serve to characterize sequential variants of proteins resulting from splicing processes [39] or identify proteins altered in structure by additional post-translational modifications (PTMs) [40]. The aims that can addressed by metaproteomics are (i) obtaining information on the taxonomic distribution of a microbial community, (ii) identifying relevant functions covered by the community, (iii) matching the identified key players to their respective covered functions and (iv) analyzing interactions between species present in a community [32].

2.1 Challenge of metaproteomics - finding the suitable database

Although the use of metaproteomics is still a powerful approach to assign proteins on the taxonomic hierarchy and to understand the functional role of the present microbes, the methodology also involves some weaknesses. The common use of metaproteomics is to determine the structure of complex microbial communities in a wide range of environments. These natural environments harbor highly diverse microbes, of which many are unknown since they are uncultured and thus so far only rarely identified [41]. Moreover, assessing the entire diversity of a community is also challenging due to a high degree of nucleotide sequence diversities of the present microbes. As a result, sequence mutations and codon bias can lead to missing gene expressions and therefore numerous undetected proteins [42]. These factors consequently affect the completeness of the currently offered free and publicly available databases for metaproteomics.

On the other hand, the non-specificity of the available databases for metaproteomics creates another challenge. The use of such large protein databases presents difficulties in distinguishing homologous species from each other since they share many protein sequence similarities, with sometimes only one or two different amino acids [43]. This complicates the annotation of identified proteins to the present species. To circumvent this bottleneck, it is recommended to include smaller and environment specific protein databases to increase the probability of achieving a high number of protein identifications and taxonomic resolution. Indeed, it was recently shown that the selected protein search database for protein identification affects taxonomic and functional annotation [44].

To address this weakness, metagenomic sequencing of the entire community is indispensable for building up a small and specific reference database, which is increasingly being performed but still remains prohibitive for standard studies due to high costs [32]. Therefore, a direct consequence of including environmentally unspecific databases is protein inference, i.e., the

sequence of an identified peptide is shared by several distinct proteins often originating from different species [45]. Thus, it is a widely practiced strategy to group redundant proteins into so called metaproteins. These metaproteins or protein groups contain proteins with similar amino acid sequence or shared peptide identifications wherefore the protein groups represent the basic unit for downstream analysis, since most metaproteomics studies are based on a peptide-centric approach [46]. Further typical limitations of metaproteomics include obtaining a high amount of protein biomass from natural samples, e.g., soil and groundwater. Unlike gene oligonucleotides, proteins cannot be amplified and therefore the sensitivity of mass spectrometry depends on the net extracted proteins [47]. Further, protein based analyses require high time efforts for sample fractionation, separation and high-depth LC-MS/MS analysis [32, 48, 49].

2.2 Objectives of metaproteomics

Environments harbor microbial communities which are highly diverse and complex. Thus, proteome scientists have begun to focus more on the cultivation of simplified communities in *in-vitro* bioreactor systems with the aim to reduce the complexity of naturally diverse communities. It allows the identification of central functions with high coverage. Such a simplified system was recently established for the human intestinal microbiota, to overcome the challenge of proteome coverage in a complex community [50].

To demonstrate the effect of increasing complexity of microbial communities, in combination with a large and environmentally unspecific database, on protein group identifications, we focused on three objectives which can be addressed by metaproteomics: (i) species coverage (i.e., protein groups assigned to species), which is fundamental for a comprehensive high-resolution taxonomic characterization (ii) proteome coverage (i.e., protein groups identified from a certain species), which is crucial for the investigation of the taxonomy and function of a community, and (iii) pathway coverage (i.e., protein groups annotated to metabolic pathways), which offers a

promising strategy for metaproteomics to reveal deeper insights into the function of a microbial community (**Figure 1A**). In order to realize this approach, we integrated four datasets from communities with increasing community complexity and focused only on bacteria (not fungi or viruses). The first dataset was derived from a simple consortium of eight bacterial strains used as a model system, the extended simplified human intestinal microbiota (SIHUMIx), which comprise of functionally important species from the dominant phyla in the human gut. We cultivated the eight species consortium in a bioreactor under controlled conditions covering the main functions of the intestinal tract in order to create a representative stable sub-community of the human intestinal microbiome [51, 52]. The second dataset was derived from a 72 species community representative of the human intestinal microbiome, which is of intermediate complexity [53]. This consortium was isolated from a human fecal sample by the Allen-Vercoe group in order to create a more comprehensive model ecosystem, and was cultured in a bioreactor under similar conditions as SIHUMIx. The third dataset was derived from samples of the intestinal colonic microbiota of mice, which is complex and consists of approximately 200 species. The final dataset came from a highly complex microbial community derived from a subsurface aquifer with an estimated >300 species [54] (**Figure 1B**). The microbiome datasets of differing complexity were generated according to a standardized metaproteomics workflow which is explained in supplement II and figure S3. In principle, it consists of three main phases: (i) sample preparation, where the proteins are extracted from the cells and tryptic digested into smaller peptides during sample preparation, (ii) data acquisition, where the peptide species are first separated by nano-flow HPLC then individually analyzed by subsequent online ionization and MS/MS and (iii) bioinformatics data analysis, where the measured peptide spectra are matched against a protein database for identification and quantification (**Figure 1C**).

3. Taxonomic resolution of increasingly complex communities

We were interested in examining how the number of species in a microbial community impacted the observed species coverage by metaproteomics. Our aim was to determine the effect of community complexity on (i) the efficiency of protein group identifications, (ii) the rate of protein group annotation at different taxonomic levels and (iii) the observed diversity of the community. Foremost, we found that increasing the number of species leads to a reduced number of protein group identifications (**Figure 2A**). This restricts the potential for a comprehensive examination of the structure of complex microbial communities. Consequently, we wanted to find out how many of the identified protein groups could be classified to each rank of the taxonomic hierarchy (**Figure 2B**). This was accomplished for each protein group by determining the lowest common phylogenetic ancestor for the taxon of origin of all proteins in the protein group. We observed that the number of assignable protein groups decreased with (i) lowering levels of taxonomic hierarchy (kingdom to species) and (ii) increasing numbers of species. Thus, obtaining a high species coverage becomes increasingly challenging with growing complexity and hinders an exact reconstruction of the taxonomic composition of the community. This has led to the hypothesis that many microbes, especially from exotic habitats, remain undiscovered in complex ecosystems, which has been described as the “microbial dark matter” [55]. Because of this, we need better strategies to acquire valid information about the taxonomy of complex communities.

3.1 Low abundance species remain unexplored in complex communities

In the field of microbial ecology, it is quite common to perform diversity analyses, mostly described as alpha diversity, which serves as a proxy for the stability, productivity and migration of a community [56]. Alpha diversity consists of two basic parameters (i) species richness, a simple count of the microbial species present in a community and (ii) species evenness, the relative equality in the abundance of these microbial species. The species diversity of a community is

mostly represented by Shannon's diversity index, which is based on the species number and abundance [56-59]. Here, the Shannon diversity indexes of the integrated microbiome data were calculated as the effective number of species in order to investigate if the diversity of the communities with differing complexity remained consistent, even with more stringent filtering criteria for identified species.

We binned our data utilizing three separate criteria for considering species presence, where species were identified by at least 1, 2 or 5 protein groups, respectively. The objective was to show the change of species richness and diversity under increasingly stringent criteria. The species richness and evenness decreased severely by several orders of magnitude for the complex communities after filtering the species identified by presence of at least 2 or 5 protein groups (**Figure 2C, D**). In contrast, the simplified intestinal microbiome only showed a marginal decrease of identified species under all criteria. Therefore, the majority of species in complex microbial communities were only identified by 1 protein group, which is in agreement with the observed low species coverage. However, it is accepted that the low abundance species are challenging to identify in complex communities, but this fact has rarely been empirically shown. One commonly applied method to assess rarity in microbial communities is rarefaction curve analysis [60] (**Figure 2E, F**). This technique allows a standardized comparison of the identified species number between different communities [61] and should be implemented also as a standard quality control measure in metaproteomics studies. In complex communities, a small number of peptide spectrum matches (PSMs) were identified for a multitude of species. This demonstrates that complex communities predominantly consist of many low abundance species [62]. Nevertheless, this phenomenon results in both a loss of taxonomic and functional information, since low abundance species can have a disproportionate role in maintaining community functionality [62]. Therefore, this "rare biosphere" is receiving greater attention, since these microbes can be involved in central biogeochemical cycles that drive ecosystem functioning [63].

4. Functional profiling of microbial communities

Besides the taxonomic characterization of microbial communities, it is also of great importance to describe their functional traits [64]. Comprehensive functional analysis can provide information on biological processes, pathway regulations and descriptions of the active enzymes [38]. Additionally, functional profiling of the human or animal intestinal microbiome can reveal altered metabolisms in response to changed environmental factors and therefore support the identification of processes leading to clinical diseases [65]. In environmental studies, the functional characterization of microbiomes can elucidate the mechanism behind particular biogeochemical processes, nutrient cycling and decomposition of organic matter to describe ecosystem functioning [54, 66]. Moreover, the identified functions of a community can be traced back to the genomes from which the proteins were derived to determine which microbe within the community is responsible for which molecular function. The procedure for functional analysis is carried out by assigning the identified proteins to their respective functions through a functional identifier by matching the protein-coding sequence with public and hierarchical structured databases for functional annotations. Currently there are several of these databases available such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [67], evolutionary genealogy of genes: Non-supervised Orthologous Groups (EggNOG) database [68] and clusters of orthologous groups (COGs) database [69]. To search the protein sequences against the functional repository databases, we used the server based platform GhostKOALA, which is directly connected with the KEGG database to annotate KEGG orthology (KO) numbers to the proteins [70]. Therefore, we reanalyzed the integrated microbial data according to the exhibited functions, since the molecular function provides a biological relevance for the structure of microbial communities.

Here, we investigated the effect of increasing community complexity on the (i) rate of annotation of the identified protein groups to KEGG-functions and (ii) efficiency of pathway coverage. We

observed that community complexity had little impact on the percentage of identified protein groups which could be annotated to molecular functions (**Figure 3A**). This finding highlights that metaproteomics is still a useful tool for complex microbial community analysis to describe the ecosystem functioning performed by the entire community.

4.1 Pathway coverage and habitat specific functions

Cellular processes are conducted by interacting enzymes that can be grouped into biological pathways. These pathways are specific for distinct metabolisms and reveal the functions of a community [49]. To receive information on the depth of the underlying functionality, it is standard procedure to determine the pathway coverage and abundance of the functionally assigned protein groups. This is done by calculating the percentage of identified KEGG-annotated protein groups compared to the total number of proteins listed in the database for the KEGG pathway. In our analysis, we found that pathway coverages are reduced in communities with increasing complexity, although the number of proteins annotated to KEGG-functions was notably high, even in complex communities (**Figure 3B**). This is due to a decreased number of identified protein groups in complex communities, which leads to a decreased chance of successful pathway coverage.

The functional profiling of microbial communities is a unique feature of metaproteomics compared to other omics approaches Biochemical pathways are essential for mediating environmental stimuli, and thus it can be expected that microbiomes from different habitats generate distinctly abundant pathways according to their host environment [38]. The up or down-regulated pathways according to the host environment can be further analyzed on each taxonomic rank to determine which present species is performing which particular function. We found that a highly abundant pathway of subsurface environment microbiomes is involved in nitrogen metabolism. A further pathway was found with highly abundance for a gut related function of the simplified gut microbiome (**Figure 3C**). For instance, this can be used to find out which metabolic pathways are

responsible for the utilization of novel nutrients, stress response or amino acid biosynthesis [71]. However, this functional profiling of microbial communities is a unique feature of metaproteomics compared to other omics approaches. The trend in microbiome research is moving towards describing taxonomic distribution and elucidating functional networks, by determining the up or down-regulated pathways caused by stimuli and thereby constructing a comprehensive functional map of a community [72]. Therefore, we have provided an example of how metaproteomics yields insight into the overall functional reactions that describe the underlying environmental dynamics.

5. Reduced protein identifications of a single microbe present in complex communities

We were interested in determining how the number of identified protein groups for a single species changed with the increasing complexity of the microbial community. First, we selected the gram-negative bacterium *Escherichia coli*, which was present in all four datasets, and calculated its proteome coverage in each community. We clearly observed that the number of identified *E. coli* proteins decreased with an increasing number of species per sample (about 40% in an 8 species community compared to <5% in >150 species community) (**Figure 4A**). This observation could result from the typical challenges of metaproteomics analyses, which include uneven species distributions, broad ranges of protein expression levels between microorganisms, and the large genetic heterogeneity within microbial communities. Second, we constructed an *in-silico* model to retrieve protein abundance information for the identified *E. coli* proteins within the four datasets using the protein abundance database *PaxDB* [73]. Following the same trend, low abundance proteins (<1 part per million, ppm) are difficult to identify in intermediate to complex microbiomes relatively to the *PaxDB* (**Figure 4B**). This result highlights that high abundance proteins are predominantly identified and thus more prominent in functional analysis while the low abundance

proteins are underrepresented, although these can have an important impact on bacterial metabolism [74].

Conclusion

This perspective highlights the challenges of species and proteome coverage in metaproteomics for microbial communities of high complexity. We observed a severe reduction of assigned species to identified protein groups, from 45% for the complex intestinal community down to 19.4% for the highly complex environmental community. Furthermore, we identified a decrease of 85% for species richness and 96.5% for species evenness, by considering only species which are identified by at least two protein groups. In complex microbiomes, we observed that low abundance proteins are mostly undetected, and therefore potentially important cellular functions could be not identified. However, metaproteomics can analyze the functional traits of microbial communities as a whole. The functional assignment of protein groups was approximately 50% higher than the species coverage in complex microbiomes. Therefore, functional profiling of complex communities by metaproteomics is considered as a promising technique to investigate ecosystem functioning of environmental microbiomes.

Expert opinion

We have discussed the current limitations of taxonomic profiling, and also outlined functional perspectives of metaproteomics analysis of microbial communities with different levels of complexity. To achieve a more comprehensive characterization of the taxonomic composition and function of complex communities in the future, strategies by which to address the challenges occurring in metaproteomics analyses are needed. First, the metagenome of uncultivated

microbes of environmental communities are constructed thus far only rarely, and therefore the employment of metaproteomics has mainly required the use of large and unspecific protein databases. The lack of comprehensive, specific databases mainly results in the reduction of taxonomic information yielded from complex communities [75]. It has been shown that the use of sample specific databases revealed a comprehensive peptide and protein identification in the context of clinical studies [75].

It was recently shown in a gut microbiome study that the parallel search against publicly large and comprehensive metagenome based databases yielded more complete information regarding taxonomy and function [43]. Second, for handling the limitations of the current technological setups and standard metaproteomics workflow, we suggest performing the taxonomic analysis at a higher rank, e.g., phylum level, even though this only provides a rough overview of the microbes present in a community. Third, to increase the taxonomic resolution even on the species level, the implementation of other techniques that would complement the current metaproteomics approach should be considered. A prominent strategy is a multi-omics approach, which combines metaproteomics with other omics disciplines to build a holistic picture of the analyzed microbial communities. Mostly, metaproteomics is simply combined with a parallel metagenomics or metabolomics approach of the same community to allow for a deeper insight into the structure and function of microbial communities [76]. Metagenomics can help to improve the taxonomic characterization and metabolomics to understand metabolic processes. Therefore, a recently evolved area of focus is metaproteogenomics, a strategy at the interface of metaproteomics and metagenomics, where a protein sequence database is generated based on metagenomic and metatranscriptomic information to increase the annotation of peptides that are currently not present in a particular reference databases [77]. This approach was constructed for rather small communities. A relatively recent study refined this strategy by building a metaproteogenomics pipeline, and then applied it to diverse microbial communities, which improved protein detection, false-positive identifications and functional profiling [78]. However, to specify the active microbes

of a community, the stable isotope probing (SIP) approach has been established [29]. The principle of this method relies upon the incorporation of stable isotope atoms, e.g., ^{13}C or ^{15}N, into the proteins of active microbes within a community, which enables a direct link to the functions of a microbial community compared to other stable isotope probing approaches, e.g., DNA/RNA-SIP [79]. Since protein-SIP can be applied to communities of intermediate complexity, it is suitable for the analysis of the intestinal microbiota by delivering, e.g., ^{15}N containing chow [80]. The same is assumingly also true for SIF [31]. In combination with protein-SIP, the species coverage can be improved by specific enrichment strategies, where certain proteins that are present in all bacterial phyla can be isolated and enriched. Exemplarily, streptavidin coated beads allow for the isolation of biotinylated proteins to reduce the overall complexity of protein mixtures for a deeper metaproteome measurement, which might lead to more valid taxonomic information at the species level [81]. Another strategy to increase the species coverage of complex microbial communities is to focus on certain areas of the environment. For example, in the case of the scientifically important intestinal microbiota, a relevant sub-localization would be the mucus layer [82]. Finally, besides the descriptive characterization of microbial communities in their natural state, it is becoming more important to determine the response of a complex community to environmental stimuli or toxins. For such research questions, it is recommended to focus on simplified communities or even pure cultures to maximize the chance of identifying a high number of proteins per microbe, which is imperative for effect-mechanism studies. The increase in number of protein identifications during the last ten years (Figure S2) suggests that other improvements, such as the strategies highlighted here, will be increasingly employed in the future. We summarized our findings according to species and proteome coverage in **figure 5** to highlight the effect of community complexity on taxonomic analysis which is crucial for metaproteomics studies. Moreover, we hypothesized that applying these strategies, in particular including a suitable and environmental specific database, in combination with mass spectrometry improvements would result in a further increase of identifiable proteins and therefore might increase the species and

proteome coverage of several orders of magnitude even for highly complex communities. Our findings, especially if our recommended strategies for improving metaproteomics analyses are employed, reveal that metaproteomics is a highly useful research tool for improving our understanding of microbiomes.

Funding

This work was supported by the Collaborative Research Centre 1076 AquaDiva (CRC AquaDiva) which is founded by the German Research Foundation (DFG). The first author Patrick Lohmann was also supported by the Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE) and the integrated research training group (iRTG).

References

1. Wilpiszeski, R.L., et al., *Soil Aggregate Microbial Communities: Towards Understanding Microbiome Interactions at Biologically Relevant Scales.* Applied and Environmental Microbiology, 2019. **85**(14).
2. Falkowski, P.G., T. Fenchel, and E.F. Delong, *The microbial engines that drive Earth's biogeochemical cycles.* Science, 2008. **320**(5879): p. 1034-1039.
3. Harris, J.M., *The presence, nature, and role of gut microflora in aquatic invertebrates: A synthesis.* Microb Ecol, 1993. **25**(3): p. 195-231.
4. Macfarlane, G.T. and S. Macfarlane, *Human colonic microbiota: ecology, physiology and metabolic potential of intestinal bacteria.* Scand J Gastroenterol Suppl, 1997. **222**: p. 3-9.
5. Zhang, X. and D. Figeys, *Perspective and Guidelines for Metaproteomics in Microbiome Studies.* Journal of Proteome Research, 2019. **18**(6): p. 2370-2380.
6. Escobar-Zepeda, A., et al., *Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics.* Scientific Reports, 2018. **8**.
7. Singer, E., et al., *High-resolution phylogenetic microbial community profiling.* ISME J, 2016. **10**(8): p. 2020-32.
8. Raju, S.C., et al., *Reproducibility and repeatability of six high-throughput 16S rDNA sequencing protocols for microbiota profiling.* J Microbiol Methods, 2018. **147**: p. 76-86.
9. Yates, J.R., *Proteomics of Communities: Metaproteomics.* Journal of Proteome Research, 2019. **18**(6): p. 2359-2359.
10. Langille, M.G., et al., *Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.* Nat Biotechnol, 2013. **31**(9): p. 814-21.
11. Iwai, S., et al., *Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes.* Plos One, 2016. **11**(11).
12. Koboldt, D.C., et al., *The Next-Generation Sequencing Revolution and Its Impact on Genomics.* Cell, 2013. **155**(1): p. 27-38.
13. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.
14. Siggins, A., E. Gunnigle, and F. Abram, *Exploring mixed microbial community functioning: recent advances in metaproteomics.* Fems Microbiology Ecology, 2012. **80**(2): p. 265-280.
15. Aslam, B., et al., *Proteomics: Technologies and Their Applications.* Journal of Chromatographic Science, 2017. **55**(2): p. 182-196.
16. Wilmes, P. and P.L. Bond, *The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms.* Environ Microbiol, 2004. **6**(9): p. 911-20.
17. von Bergen, M., et al., *Insights from quantitative metaproteomics and protein-stable isotope probing into microbial ecology.* ISME J, 2013. **7**(10): p. 1877-85.
18. Braga, R.M., M.N. Dourado, and W.L. Araujo, *Microbial interactions: ecology in a molecular perspective.* Brazilian Journal of Microbiology, 2016. **47**: p. 86-98.
19. Xiao, J.Q., et al., *Metagenomic Taxonomy-Guided Database-Searching Strategy for Improving Metaproteomic Analysis.* Journal of Proteome Research, 2018. **17**(4): p. 1596-1605.
20. Kleiner, M., et al., *Assessing species biomass contributions in microbial communities via metaproteomics.* Nat Commun, 2017. **8**(1): p. 1558.

21. Hettich, R.L., et al., *Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities.* Anal Chem, 2013. **85**(9): p. 4203-14.
22. Heyer, R., et al., *Metaproteomics of complex microbial communities in biogas plants.* Microb Biotechnol, 2015. **8**(5): p. 749-63.
23. Jordan, F., et al., *Diversity of key players in the microbial ecosystems of the human body.* Scientific Reports, 2015. **5**.
24. Guijas, C., et al., *Metabolomics activity screening for identifying metabolites that modulate phenotype.* Nature Biotechnology, 2018. **36**(4): p. 316-320.
25. Prosser, G.A., G. Larrouy-Maumus, and L.P.S. de Carvalho, *Metabolomic strategies for the identification of new enzyme functions and metabolic pathways.* Embo Reports, 2014. **15**(6): p. 657-669.
26. Mumtaz, M.W., et al., *An overview of recent developments in metabolomics and proteomics - phytotherapic research perspectives.* Frontiers in Life Science, 2017. **10**(1): p. 1-37.
27. Lu, W.Y., et al., *Metabolite Measurement: Pitfalls to Avoid and Practices to Follow.* Annual Review of Biochemistry, Vol 86, 2017. **86**: p. 277-304.
28. Baumann, S., et al., *Requirements and Perspectives for Integrating Metabolomics with other Omics Data.* Current Metabolomics, 2013. **1**(1): p. 15-27.
29. Jehmlich, N., et al., *Protein-SIP in environmental studies.* Curr Opin Biotechnol, 2016. **41**: p. 26-33.
30. Seifert, J., et al., *Protein-based stable isotope probing (protein-SIP) in functional metaproteomics.* Mass Spectrom Rev, 2012. **31**(6): p. 683-97.
31. Kleiner, M., et al., *Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities.* Proc Natl Acad Sci U S A, 2018. **115**(24): p. E5576-E5584.
32. Kleiner, M., *Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities.* Msystems, 2019. **4**(3).
33. Herbst, F.A., et al., *Enhancing metaproteomics--The value of models and defined environmental microbial systems.* Proteomics, 2016. **16**(5): p. 783-98.
34. Muth, T., et al., *Searching for a needle in a stack of needles: challenges in metaproteomics data analysis.* Mol Biosyst, 2013. **9**(4): p. 578-85.
35. von Bergen, M., et al., *Insights from quantitative metaproteomics and protein-stable isotope probing into microbial ecology.* Isme Journal, 2013. **7**(10): p. 1877-1885.
36. Yates, J.R., 3rd, *Recent technical advances in proteomics.* F1000Res, 2019. **8**.
37. Zhang, X., et al., *Advancing functional and translational microbiome research using meta-omics approaches.* Microbiome, 2019. **7**(1).
38. Wilmes, P., A. Heintz-Buschart, and P.L. Bond, *A decade of metaproteomics: where we stand and what the future holds.* Proteomics, 2015. **15**(20): p. 3409-17.
39. Tress, M.L., F. Abascal, and A. Valencia, *Alternative Splicing May Not Be the Key to Proteome Complexity.* Trends Biochem Sci, 2017. **42**(2): p. 98-110.
40. Devabhaktuni, A., et al., *TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets.* Nat Biotechnol, 2019. **37**(4): p. 469-479.
41. Wang, Y., et al., *A Culture-Independent Approach to Unravel Uncultured Bacteria and Functional Genes in a Complex Microbial Community.* Plos One, 2012. **7**(10).
42. Caro-Quintero, A. and H. Ochman, *Assessing the Unseen Bacterial Diversity in Microbial Communities.* Genome Biology and Evolution, 2015. **7**(12): p. 3416-3425.
43. Tanca, A., et al., *The impact of sequence database choice on metaproteomic results in gut microbiota studies.* Microbiome, 2016. **4**(1): p. 51.

44. Geron, A., et al., *Deciphering the Functioning of Microbial Communities: Shedding Light on the Critical Steps in Metaproteomics.* Frontiers in Microbiology, 2019. **10**.
45. 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 Proteomics, 2019. **16**(5): p. 375-390.
46. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.
47. Cox, J. and M. Mann, *Is proteomics the new genomics?* Cell, 2007. **130**(3): p. 395-398.
48. Feist, P. and A.B. Hummon, *Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples.* International Journal of Molecular Sciences, 2015. **16**(2): p. 3537-3563.
49. Heyer, R., et al., *Challenges and perspectives of metaproteomic data analysis.* J Biotechnol, 2017. **261**: p. 24-36.
50. Schape, 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, 2019. **7**(12).
51. Becker, N., et al., *Human intestinal microbiota: characterization of a simplified and stable gnotobiotic rat model.* Gut Microbes, 2011. **2**(1): p. 25-33.
52. Krause, J.L., et al., *Changes in pH can modify the ability of a simplified human intestinal microbiota to stimulate MAIT cells.* European Journal of Immunology, 2019. **49**: p. 39-39.
53. McDonald, J.A.K., et al., *Simulating distal gut mucosal and luminal communities using packed-column biofilm reactors and an in vitro chemostat model.* Journal of Microbiological Methods, 2015. **108**: p. 36-44.
54. Starke, R., et al., *Candidate Brocadiales dominates C, N and S cycling in anoxic groundwater of a pristine limestone-fracture aquifer.* J Proteomics, 2017. **152**: p. 153-160.
55. Rinke, C., et al., *Insights into the phylogeny and coding potential of microbial dark matter.* Nature, 2013. **499**(7459): p. 431-7.
56. Stirling, G. and B. Wilsey, *Empirical Relationships between Species Richness, Evenness, and Proportional Diversity.* Am Nat, 2001. **158**(3): p. 286-99.
57. Davis, S.C., et al., *Gut microbiome diversity influenced more by the Westernized dietary regime than the body mass index as assessed using effect size statistic.* Microbiologyopen, 2017. **6**(4).
58. Macarthur, R.H., *Patterns of Species Diversity.* Biological Reviews, 1965. **40**(4): p. 510-+.
59. Shannon, C.E., *A Mathematical Theory of Communication.* Bell System Technical Journal, 1948. **27**(3): p. 379-423.
60. Gotelli, N.J. and R.K. Colwell, *Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness.* Ecology Letters, 2001. **4**(4): p. 379-391.
61. Hughes, J.B. and J.J. Hellmann, *The application of rarefaction techniques to molecular inventories of microbial diversity.* Methods Enzymol, 2005. **397**: p. 292-308.
62. Jousset, A., et al., *Where less may be more: how the rare biosphere pulls ecosystems strings.* Isme Journal, 2017. **11**(4): p. 853-862.
63. Karpinets, T.V., et al., *Linking Associations of Rare Low-Abundance Species to Their Environments by Association Networks.* Frontiers in Microbiology, 2018. **9**.
64. Escalas, A., et al., *Microbial functional diversity: From concepts to applications.* Ecology and Evolution, 2019.
65. Haange, S.B. and N. Jehmlich, *Proteomic interrogation of the gut microbiota: potential clinical impact.* Expert Rev Proteomics, 2016. **13**(6): p. 535-7.
66. Keiblinger, K.M., et al., *Soil and leaf litter metaproteomics-a brief guideline from sampling to understanding.* Fems Microbiology Ecology, 2016. **92**(11).

67. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**(1): p. 27-30.
68. Huerta-Cepas, J., et al., *eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences.* Nucleic Acids Research, 2016. **44**(D1): p. D286-D293.
69. Galperin, M.Y., et al., *Expanded microbial genome coverage and improved protein family annotation in the COG database.* Nucleic Acids Research, 2015. **43**(D1): p. D261-D269.
70. Kanehisa, M., Y. Sato, and K. Morishima, *BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences.* Journal of Molecular Biology, 2016. **428**(4): p. 726-731.
71. Partridge, J.D., et al., *Escherichia coli transcriptome dynamics during the transition from anaerobic to aerobic conditions.* Journal of Biological Chemistry, 2006. **281**(38): p. 27806-27815.
72. Xiao, Y., et al., *Mapping the ecological networks of microbial communities.* Nat Commun, 2017. **8**(1): p. 2042.
73. Wang, M., et al., *Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines.* Proteomics, 2015. **15**(18): p. 3163-8.
74. Xiong, W.L., et al., *Microbial metaproteomics for characterizing the range of metabolic functions and activities of human gut microbiota.* Proteomics, 2015. **15**(20): p. 3424-3438.
75. Rechenberger, J., et al., *Challenges in Clinical Metaproteomics Highlighted by the Analysis of Acute Leukemia Patients with Gut Colonization by Multidrug-Resistant Enterobacteriaceae.* Proteomes, 2019. **7**(1).
76. Gutleben, J., et al., *The multi-omics promise in context: from sequence to microbial isolate.* Critical Reviews in Microbiology, 2018. **44**(2): p. 212-229.
77. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nature Methods, 2014. **11**(11): p. 1114-1125.
78. Schiebenhoefer, H., et al., *Challenges and promise at the interface of metaproteomics and genomics: an overview of recent progress in metaproteogenomic data analysis.* Expert Review of Proteomics, 2019. **16**(5): p. 375-390.
79. Taubert, M., et al., *Protein-SIP enables time-resolved analysis of the carbon flux in a sulfate-reducing, benzene-degrading microbial consortium.* ISME J, 2012. **6**(12): p. 2291-301.
80. Oberbach, A., et al., *Metabolic in Vivo Labeling Highlights Differences of Metabolically Active Microbes from the Mucosal Gastrointestinal Microbiome between High-Fat and Normal Chow Diet.* J Proteome Res, 2017. **16**(4): p. 1593-1604.
81. Blumert, C., et al., *Analysis of the STAT3 interactome using in-situ biotinylation and SILAC.* Journal of Proteomics, 2013. **94**: p. 370-386.
82. Haange, S.B., et al., *Metaproteome analysis and molecular genetics of rat intestinal microbiota reveals section and localization resolved species distribution and enzymatic functionalities.* J Proteome Res, 2012. **11**(11): p. 5406-17.

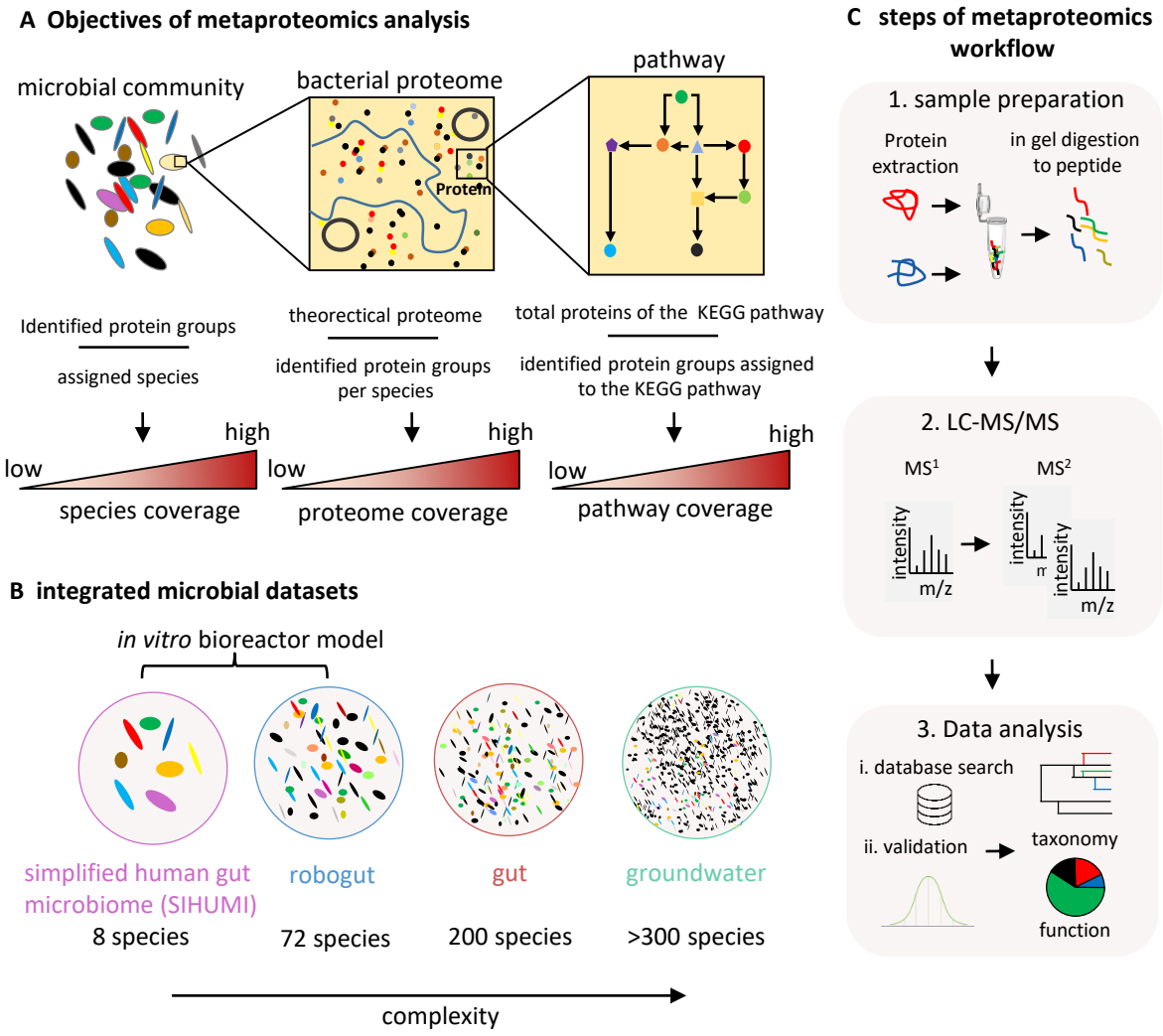

Figure 1: Objectives of metaproteomics. (A) Objectives of metaproteomics analysis. (B) Integrated datasets of the increasing complex microbiomes. (C) steps of a standard metaproteomics workflow. MS^1=mass spectrometry full scan, MS^2=tandem mass spectrometry scan, m/z=mass-to-charge ratio

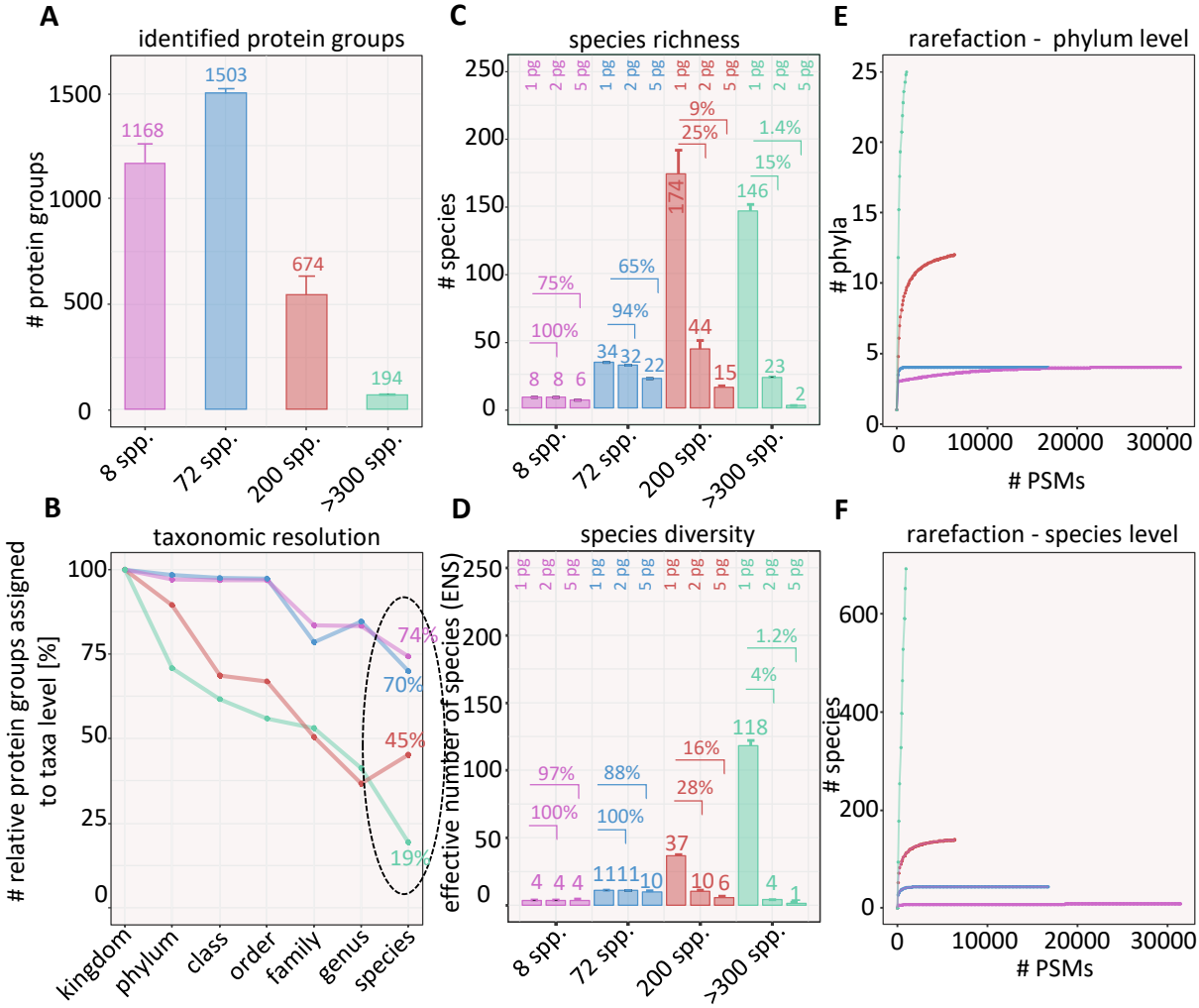

Figure 2: Taxonomic resolution of microbial communities. (A) Total number of identified protein groups on species level. (B) Relative number of protein groups assigned to each taxonomical level. (C) Species richness calculated by the count of different species for each microbiome. Visualization of species identified with at least 1, 2 or 5 identified protein groups (pg). (D) Species evenness is shown by the effective number of species (ENS) calculated by the exponential shannon-index (SIHUMI; 1.43, Robogut; 2.41, GUT; 3.6, GW; 4.77). Depiction of ENS for at least 1, 2 or 5 identified protein groups (pg). (E, F) Rarefaction curves for phylum and species level of the integrated datasets.

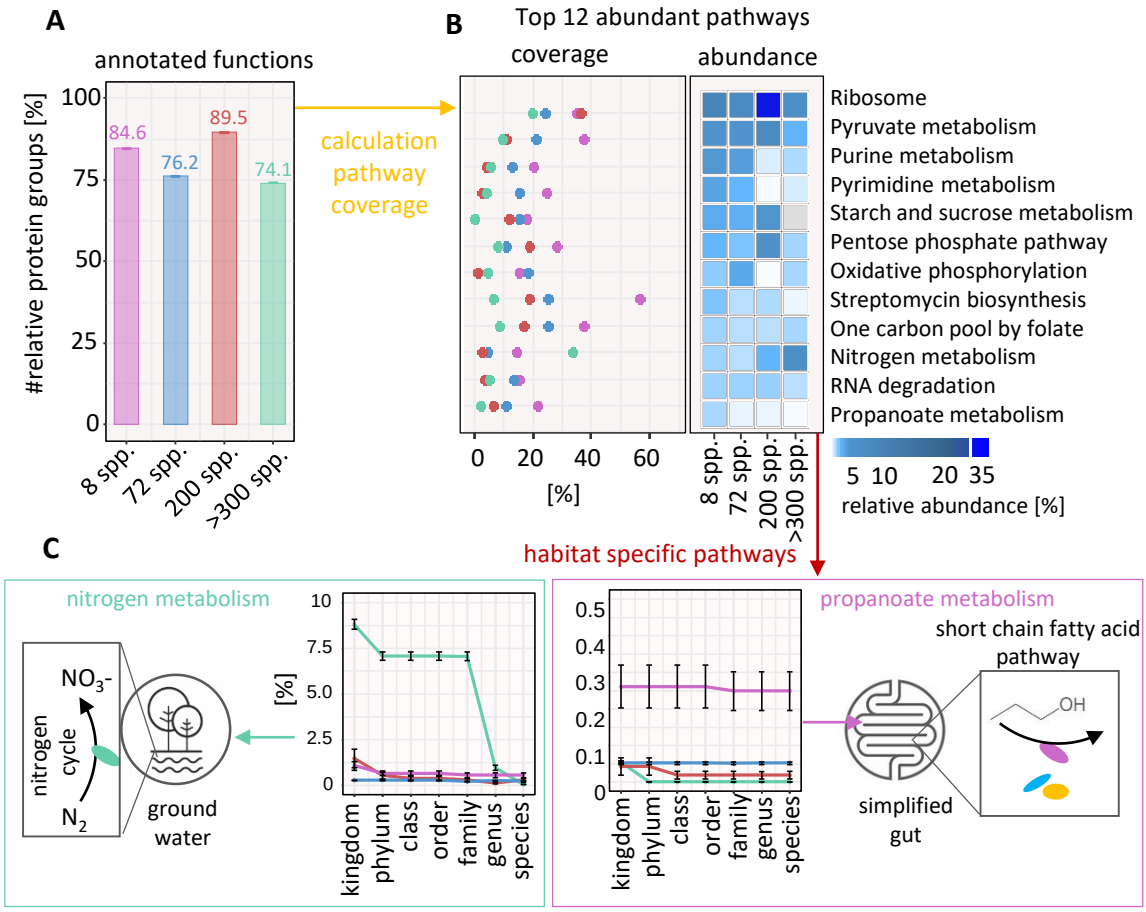

Figure 3: Functional profiling of microbial communities. (A) Relative number of protein groups assigned to a KEGG-function. (B) Heatmap of top 12 relative abundant pathways and the pathway coverage calculated by unique identified protein groups for a pathway. (C) Selected pathways for gut and groundwater related microbiomes to show relative number of identified protein groups (pathway abundances) of each taxonomical level.

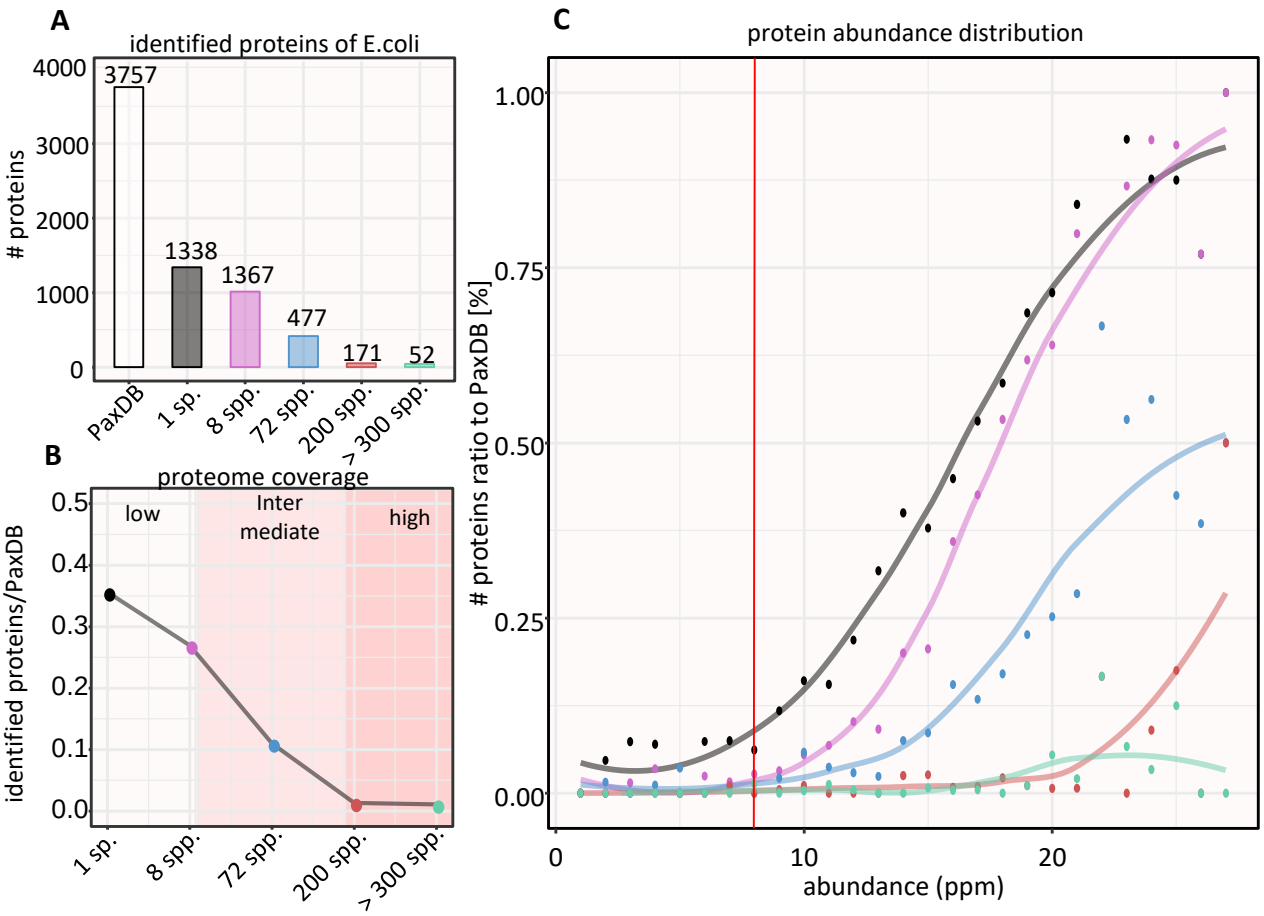

Figure 4: Proteome coverage of a single microbe present in complex communities. (A) absolute number of identified proteins of *E.coli* in the different complex microbial communities. (B) relative number of identified proteins (proteome coverage) of *E.coli* in the different complex microbial communities (C) Abundance distribution of relative number of proteins compared to PaxDB of *E.coli* in communities with increasing complexity. The protein abundance is calculated in parts per million (ppm). The red line represents the edge of the low abundance range.

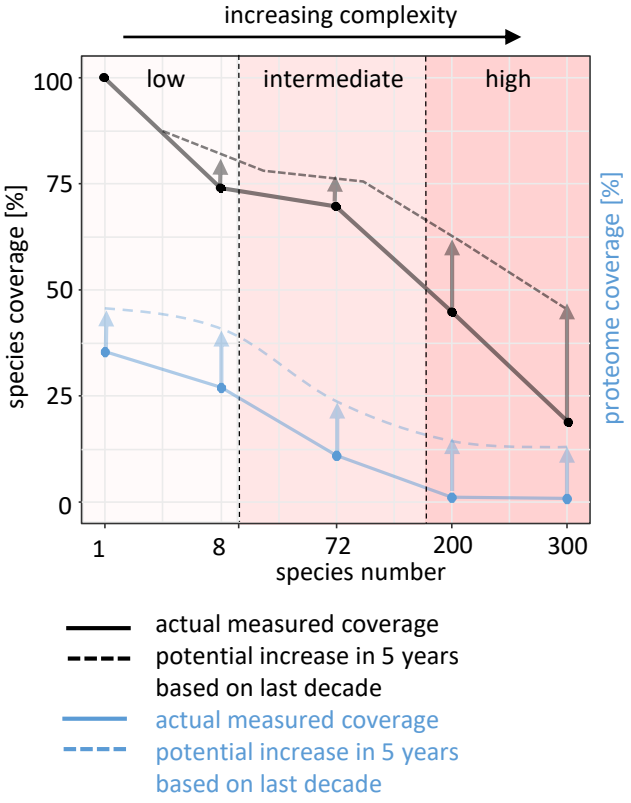

Figure 5: Summary of challenges of species coverage (black) and proteome coverage (blue) with increasing number of species. Solid line represents the actual known range of species or proteome identifications. Dashed line represents the hypothetical increase of new identified species as a result of protein enrichment strategies. This hypothetical increase for the next five years is based on the technical improvements in mass spectrometry since the last decade.

Supplement I

Protein identifications during the last decade

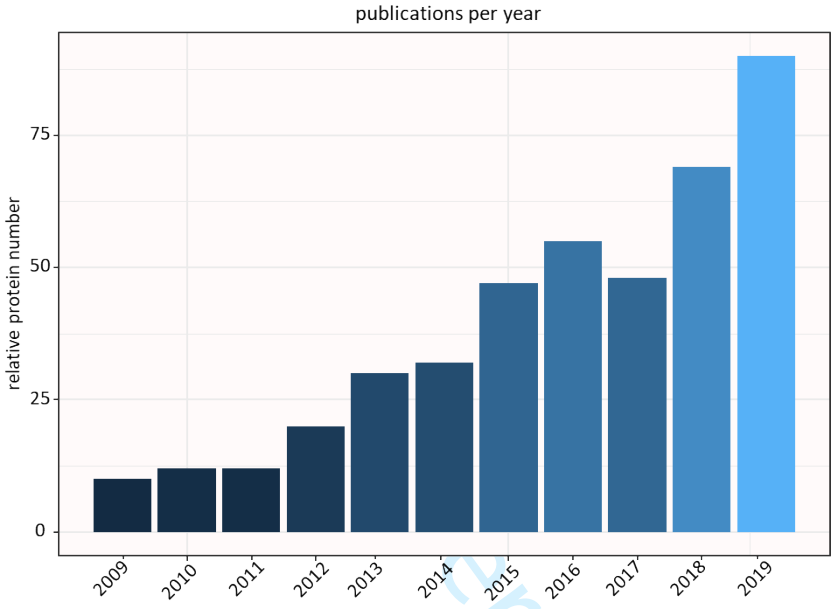

Fig.S1: Number of publications during last decade of metaproteomics studies. Publications per year were found by PubMed with the keyword: “metaproteomics”.

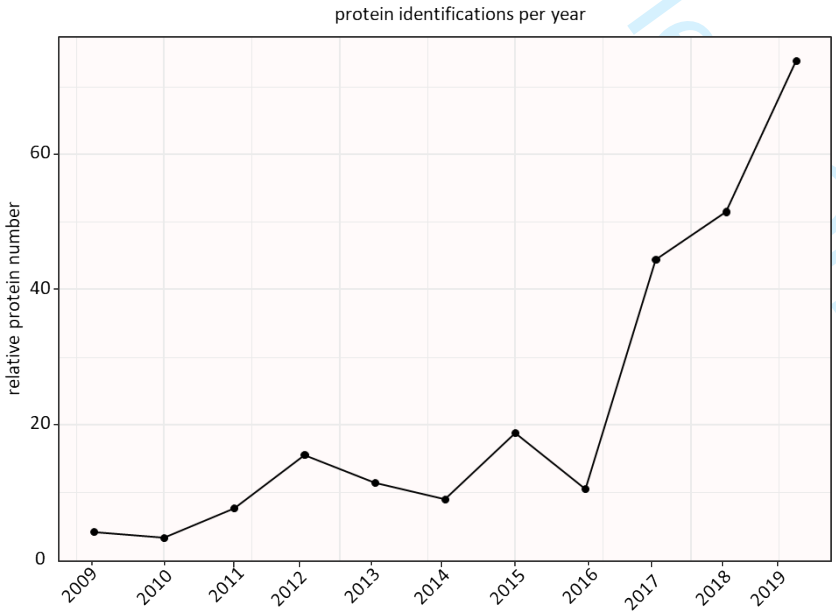

Fig.S2: The mean value of the number of proteins per minute of gradient length of several publications (table S1) published in the given year found in PubMed with the keyword:”metaproteomics”. The publications were selected by providing following parameter: Respective environment, number of proteins, instrument, LC-gradient, search engine

Table S1: Identified proteins and selected parameter of metaproteomics studies during last decade

published [year]	environment	# proteins	instrument	FDR (%)	LC-gradient (min.)	Search engine	proteins/min	reference
2009	human gut	634	LTQ orbitrap MS	5	120	SEQUEST	**4.2**	Verberkmoes et al., 2009
2010	soil	333	LTQ XL MS	-	120	SEQUEST	**4.6**	Chourey et al., 2010
2010	human salivary	139	LTQ linear ion trap MS	1	65	SEQUEST	**2.1**	Rudney et al., 2010
2011	lung lichen	463	LTQ orbitrap MS	3.9	60	MASCOT	**7.7**	Schneider et al., 2011
2012	Human feces	839	LTQ orbitrap velos	5	77	MASCOT	**10.9**	Haange et al., 2012
2012	forest soil	881	LTQ orbitrap MS	4.9	60	MASCOT	**14.7**	Keiblinger et al., 2012
2012	human gut	1790	LTQ orbitrap XL MS	5	85	OMSSA	**21.1**	Kolmeder et al., 2012
2013	forest soil	494	LTQ orbitrap XL MS	-	105	MASCOT/SEQUEST	**4.7**	Becher et al., 2013
2013	coastal east antarctica	548	LTQ FT ultra MS	5	30	MASCOT	**18.3**	Williams et al., 2013
2014	enrichment culture	576	LTQ orbitrap MS	1	82	MASCOT	**7.0**	Bozionovskia et al., 2014
2014	seawater	664	LTQ orbitrap MS	1.2	60	SEQUEST	**11.1**	Dong et al., 2014
2015	sludge	1395	ESI-amaZon iontrap MS	1	120	MASCOT	**11.6**	Püttker et al., 2015
2015	soil	1351	Q Exective MS	1.3	120	MASCOT	**11.3**	Bastida et al., 2015
2015	human gut/feces	4031	LTQ orbitrap velos MS	2.4	120	SEQUEST	**33.6**	Young et al., 2015
2016	chicken gut	1719	Q exective plus MS	1	240	MASCOT	**7.2**	Tilocca et al., 2016
2016	human salivary	2090	Q exective HC MS	1	150	Andromeda	**13.9**	Belstrøm et al., 2016
2017	human gut	20558	orbitrap fusion MS	-	260	Maxquant	**79.1**	Zhang et al., 2017
2017	rhizosphere	294	triple TOF MS	-	30	MASCOT	**9.8**	Mattarozzi et al., 2018
2018	mice gut	5610	LTQ orbitrap XL MS	1	70	Blazmass	**80.1**	Moon et al., 2018
2018	human oral surface	3671	Qexective HF orbitrap MS	1	160	maxquant	**22.9**	Jersie-Christensen et al., 2018
2019	human gut	9171	LTQ orbitrap velos pro MS	1	80	MASCOT/Sequest	**114.6**	Hickl et al., 2019
2019	human salvia/tongue	3969	Q exective plus MS	6	120	Comet	**33.1**	Rabe et al., 2019

Supplement II

Material and Methods

MM1. Metaproteomics workflow and data analysis

A complete metaproteomics workflow of sample preparation, data acquisition and data analysis is shown in figure S3.

1.1 Sample preparation

Cell lysis and protein extraction

Cells were harvested and resuspended in 1-5 ml Lysis-buffer (0.29% NaCl, 0,01M Tris-HCl, 5mM EDTA, 0.4% SDS) with 1 µl PMSF solution. The suspended cells were further lysed by bead-beating with 3 cycles of FastPrep for 1 min. The lysate was then heated and mixed for 15 min. at 60°C in a Thermomixer. The cell debris were removed by centrifugation at 10 000 g for 10 min. at 4°C. The proteins were precipitated in 5 volumes of acetone with overnight incubation at -20°C (for the communities: SIHUMI, 8 spp.; Robogut, 72 spp.; Gut, 200 spp.). The protein extraction for the community Groundwater, >300 spp. was performed according to Starke et al., 2017. The precipitated proteins were centrifuged at 15 000 g for 10 min. at 4°C. The pellet was evaporated using a SpeedVac for 5 min. The dry protein pellet was stored at -20°C.

SDS-PAGE, proteolytic digestion, and peptide extraction

For SDS-PAGE we used 25 µg protein per sample, added 20 µl SDS loading buffer to each sample and incubated them for 5 min in a ThermoMixer at 95°C and 1400 rpm. After SDS-PAGE and staining with colloidal Coomassie brilliant blue (Merck, Darmstadt, Germany) overnight, the coloured gel bands containing all proteins was cut out and sliced into smaller gel pieces to increase accessibility to the protease and destained. In order to reduce the cysteine residuals, proteins in each band were modified with 10 mM Dithioerythritol (DTT) and 100 mM 2-iodacetamide (IAA) and incubated for 30 min. at room temperature. The alkylated proteins were proteolytically digested using 0.5 µg trypsin (Sigma-Aldrich, St. Louis, USA) at 37°C, overnight. Digestion was stopped by adding 10 mM ammonium bicarbonate in 0.1% formic acid (FA). After peptide extraction using extraction buffer (50% acetonitrile and 5% formic acid) the samples were evaporated using the SpeedVac for 2h and stored at -20°C. The extracted peptides were desalted using ZipTip filter (Thermo Fischer Scientific, Waltham, USA) following the manufacturer's instructions. Peptides were dissolved in 0.1% FA and injected into the liquid chromatography-mass spectrometer.

1.2. Data acquisition

Liquid chromatography-tandem mass spectrometry (LC-MS/MS)

Samples were analyzed using liquid chromatography (HPLC, Ultimate 3000 RSLCnano, Dionex/Thermo Fisher Scientific, Idstein, Germany) coupled via a TriVersa NanoMate (Advion, Ltd., Harlow, UK) source in LC chip coupling mode with an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific, Waltham, USA). Samples (5 µl) were first loaded for 5 min on the precolumn (µ-pre-column, Acclaim PepMap C18, 2 cm, Thermo Scientific) at 4% mobile phase B (80% acetonitrile in nanopure water with 0.08% formic acid) and 96% mobile phase A (nanopure water with 0.1% formic acid) at a flow rate of 300 nl/min and at 35°C. Then they were eluted from the analytical column (Acclaim PepMap C18 LC column, 25 cm, Thermo Scientific) over a 100-min linear gradient of mobile phase B (4%–50%). The MS was set on Top Speed for 3 s using the Orbitrap analyzer for MS and MS/MS scans with higher energy collision dissociation (HCD) fragmentation at normalized collision energy of 30%. MS scans were measured at a resolution of 120,000 in the scan range of 400–1,600 *m/z*. The MS ion count target was set to $4x10^5$ at an injection time of 60 ms. Most intense peaks (charge state 2-7) were isolated for MS/MS scans by a quadrupole with an isolation window of 2 Da and were measured with a resolution of 15,000. The dynamic exclusion was set to 30 s with a +/-10 ppm tolerance. The automatic gain control target was set to $5x10^4$ with an injection time of 150 ms

1.3 Data analysis

The acquired raw data were searched against the database: bacterial all DB (6.5 GB, $>10^6$ sequences), downloaded 2017 from Uniprot. The experimental acquired sequenced were matched against the *in-silico* sequences of the database. We considered only proteins with a false-discovery rate of 1%. The identified proteins were filtered according the following criteria: (i) at least 2/3 replicates show an abundance value, (ii) proteins contain at least one unique peptide, (iii) non-bacterial proteins were removed, (iv) proteins assigned to only one protein group ID were considered. The proteins were than grouped into protein groups according to the lowest common ancestor (lca) for the different taxonomic ranks. Protein groups containing proteins which were not assigned to the same taxon were annotated to heterogeneous. The number of protein groups with a unique taxon were counted (without heterogeneous). The panels were created by R version 3.6.1 with the installed packages ggplot2, export, extrafont and readr.

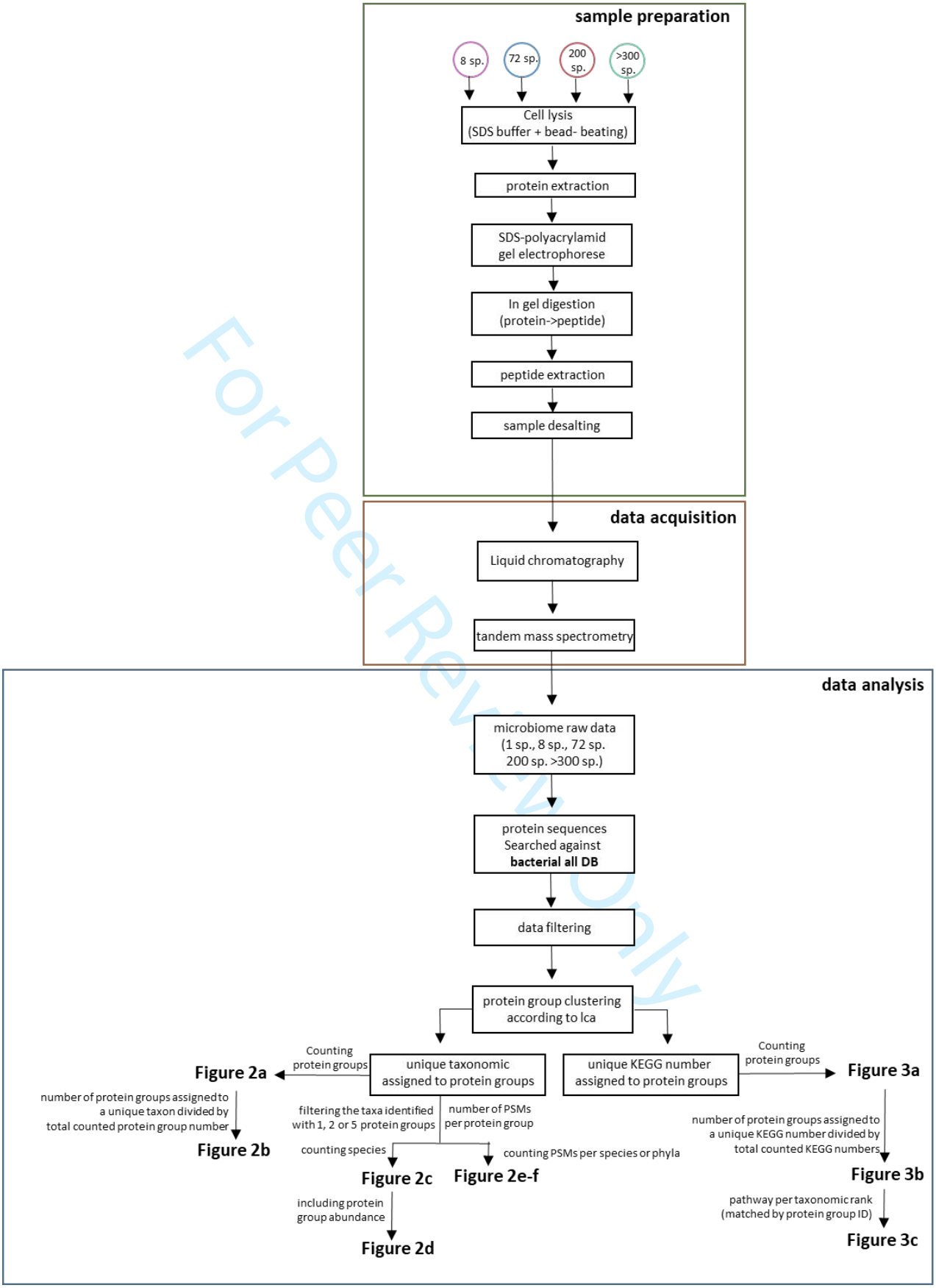

Fig.S3: Schematic of the metaproteomics workflow. Shown are the three main steps: sample preparation, data acquisition and data analysis.

MM2. Proteome coverage of *E.coli* present in complex communities

In order to determine if the increasing complexity of the involved microbiomes influences the identifiable number of proteins, proteome coverage and abundance distribution of a single microbe, we integrated the datasets of the four microbiomes. The measured peptide sequences of each dataset was searched against an *E.coli* database (1.75 MB; 4306 Sequences). This leads to the identification of *E.coli* proteins present in the dataset. After filtering of the identified proteins (removing human proteins, proteins without abundance, proteins without at least one unique peptide), the protein accession numbers were searched against the PaxDB to (1) confirm the identification of *E.coli* proteins and (2) normalize the abundance according to the proteins of the PaxDB.

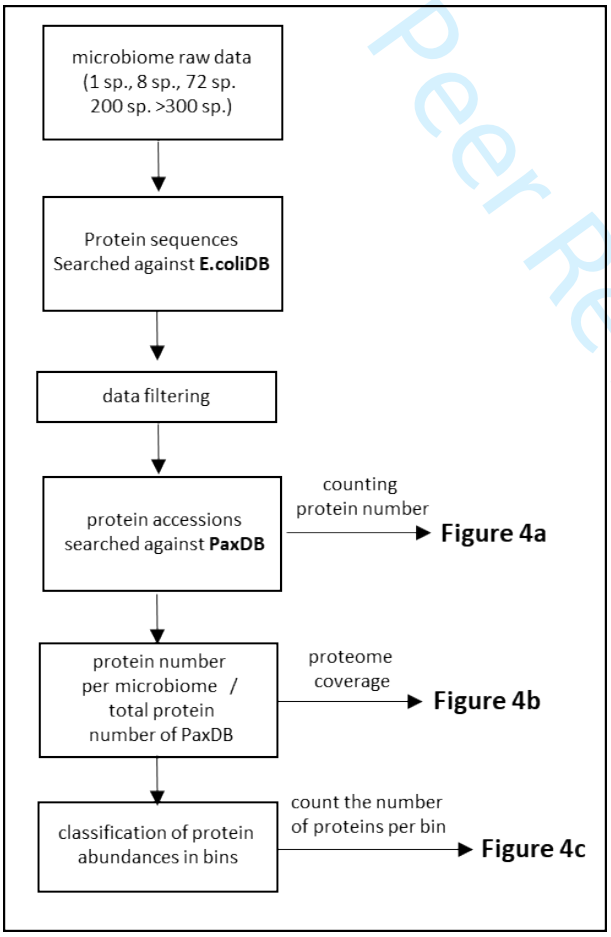

Fig.S4: Schematic of the data analysis workflow to calculate the protein identification of a single microbe (*E.coli*) present in increasingly complex microbial communities (shown in figure 4).

Supplement III

Terms and definitions

Term	Definition
16S rRNA gene	small subunit of the ribosomal ribonucleic acid, a target gene for sequencing in genetics
PICRUSt	phylogenetic investigation of communities by reconstruction of unobserved states, a tool for functional predictions based on 16S rRNA gene (Langille et al., 2013)
protein-SIP/-SIF	protein-stable isotope probing/-fingerprinting is a technique for identifying active species in a community (Jehmlich et al., 2016)
LC-MS/MS	liquid chromatography-tandem mass spectrometry, analytical technique that connects the separation capabilities of liquid chromatography (HPLC) with the mass analysis capabilities of mass spectrometry (MS) (Dass et al, 2007)
microbial dark matter	microbial clades which are not identified and not able to culture in the lab and remain unknown (Jeff Bowman, 2018)
alpha diversity	analysis of species diversity by calculation the number and abundance of present species in a environment (Prehn-Kristensen et al., 2018)
shannon diversity index	a standard quantitative diversity index to reflect the number and distribution of species in a community (Morris et al., 2014)
species richness	the number of different species present in a community (Stirling et al., 2001)
species evenness	how equal the species of a community are (Stirling et al., 2001)
effective number of species	equally-common species in a community assessing species evenness by the exponential of shannon index (Chiu and Chao, 2016)
rare biosphere	low abundant and rarely identified species in a ecosystem which contributes to ecosystem functioning (Jousset et al., 2017)
KEGG orthology	KEGG orthology database containing molecular functions represented in terms of functional orthologs numbers (Kanehisa et al., 2016)
rarefaction curves	calculation of species richness for a given number of individual samples (Gotelli et al., 2001)
SIHUMI	simplified human gut microbiome model system, microbial community mimicking the human gut consisting of 8 species (Schaepe et al., 2019)
Robogut	simplified microbial community model system mimicking the human gut consisting of 72 species

Box S1: Definitions of often used terms in the manuscript