Database-independent de novo metaproteomics of complex microbial communities

Kleikamp, Hugo B.C.; Pronk, Mario; Tugui, Claudia; Guedes da Silva, Leonor; Abbas, Ben; Lin, Yue Mei; van Loosdrecht, Mark C.M.; Pabst, Martin

DOI
10.1016/j.cels.2021.04.003

Publication date
2021

Document Version
Final published version

Published in
Cell Systems

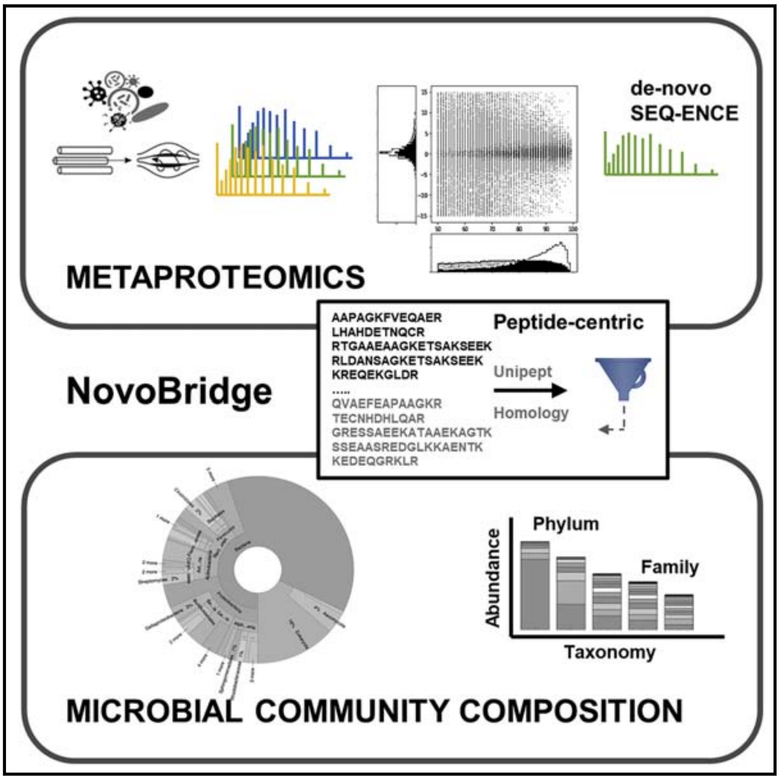
Citation (APA)
Kleikamp, H. B. C., Pronk, M., Tugui, C., Guedes da Silva, L., Abbas, B., Lin, Y. M., van Loosdrecht, M. C. M., & Pabst, M. (2021). Database-independent de novo metaproteomics of complex microbial communities. *Cell Systems*, *12*(5), 375-383.e5. https://doi.org/10.1016/j.cels.2021.04.003

Cell Systems

Methods in Brief

Database-independent *de novo* metaproteomics of complex microbial communities

Graphical abstract



Highlights

- NovoBridge—a novel *de novo* metaproteomics pipeline
- Database-matching independent metaproteomic analysis of microbes
- Rapid quantitative profiling of microbial communities
- Novel validation strategy for taxonomic annotations of *de novo* sequences

Authors

Hugo B.C. Kleikamp, Mario Pronk, Claudia Tugui, ..., Yue Mei Lin, Mark C.M. van Loosdrecht, Martin Pabst

Correspondence

m.pabst@tudelft.nl

In brief

Metaproteomics has emerged as one of the most promising post-genomics approaches. Kleikamp et al., describe a novel *de novo* metaproteomics pipeline (NovoBridge) that enables rapid community profiling without the need for constructing protein sequence databases. The performance was systematically evaluated using pure reference strains, synthetic communities, and natural communities.

Kleikamp et al., 2021, Cell Systems *12*, 375–383
May 19, 2021 © 2021 Elsevier Inc.
https://doi.org/10.1016/j.cels.2021.04.003

CellPress

Methods in Brief

Database-independent *de novo* metaproteomics of complex microbial communities

Hugo B.C. Kleikamp,[1] Mario Pronk,[1] Claudia Tugui,[1] Leonor Guedes da Silva,[1] Ben Abbas,[1] Yue Mei Lin,[1] Mark C.M. van Loosdrecht,[1] and Martin Pabst[1,2,*]

[1]Delft University of Technology, Department of Biotechnology, van der Maasweg 9, 2629 HZ Delft, the Netherlands
[2]Lead contact
*Correspondence: m.pabst@tudelft.nl

https://doi.org/10.1016/j.cels.2021.04.003

SUMMARY

Metaproteomics has emerged as one of the most promising approaches for determining the composition and metabolic functions of complete microbial communities. Conventional metaproteomics approaches rely on the construction of protein sequence databases and efficient peptide-spectrum-matching algorithms, an approach that is intrinsically biased towards the content of the constructed sequence database. Here, we introduce a highly efficient, database-independent *de novo* metaproteomics approach and systematically evaluate its quantitative performance using synthetic and natural microbial communities comprising dozens of taxonomic families. Our work demonstrates that the *de novo* sequencing approach can vastly expand many metaproteomics applications by enabling rapid quantitative profiling and by capturing unsequenced community members that otherwise remain inaccessible for further interpretation.

Kleikamp et al., describe a novel *de novo* metaproteomics pipeline (NovoBridge) that enables rapid community profiling without the need for constructing protein sequence databases.

INTRODUCTION

State-of-the-art approaches for analyzing the composition of microbial communities are based on *in situ* staining, 16S ribosomal RNA sequencing, or whole-genome shotgun-based approaches. Moreover, metatranscriptomics provides additional gene activity information, but unfortunately, mRNA levels often only poorly correlate with actual protein abundances (Maier et al., 2009). Therefore, those approaches do not directly assess the actual phenotype of a community, and the actively expressed pathways responsible for metabolic conversions remain elusive (Martin and Uroz, 2016).

On the other hand, metaproteomics targets the functional parts—the proteins—of a community directly, and therefore provides insights into the community phenotype. Furthermore, because proteins make up the bulk mass of a cell, metaproteomics also estimates the contribution of individual community members to the community biomass (Kleiner et al., 2017).

In recent years, metaproteomics has gained substantial momentum with the development of high-resolution proteomics workstations and the establishment of next-generation sequencing (NGS) technologies, which provide affordable high-quality (protein) sequence databases from complete communities (Wilmes and Bond, 2006).

Classical metaproteomics approaches employ peptide-spectrum-matching algorithms used for subsequent protein and species identification. The quality and completeness of the employed databases are therefore of utmost importance (Timmins-Schiffman et al., 2017; Xiao et al., 2018). A complete database covers the genetic potential of all community members and may contain hundreds of thousands of sequences. Alternatively, comprehensive (and even larger) public sequence databases such as NCBI, UniProtKB/Swiss-Prot, or GenBank may be accessed (in addition) (Xiao et al., 2018), which, however, require advanced focusing/filtering strategies to manage computational efforts (Heyer et al., 2017; Muth et al., 2015, 2016; Potgieter et al., 2019). Very large protein sequence databases challenge the common ‘‘peptide-spectrum-matching’’ algorithms and associated statistical parameters, which have been historically developed for single-species proteomics. Consequently, conventional metaproteomics experiments can be compromised in regard to sensitivity, accuracy, and throughput (Heyer et al., 2017; Muth et al., 2015; Timmins-Schiffman et al., 2017). Moreover, database-matching inherently biases the outcome of a metaproteomics measurement toward the (constructed) protein sequence database.

A database-independent approach, such as *de novo* peptide sequencing, that directly annotates mass spectrometric fragmentation spectra with amino acid sequences overcomes the above mentioned database-related limitations. Ultimately, the generation of the peptide sequence lists from the mass spectrometric raw data can be regarded as inherently unbiased (Muth et al., 2016). Following a successful *de novo* sequencing, the sequence lists only require retrieving taxonomic and functional annotations from comprehensive taxonomic databases using efficient ‘‘text-search’’ tools. Thereby, *de novo* sequencing also avoids the loss of taxonomic and functional information from community members not covered by the database. Those signals (not covered by the target database) can be further matched to related species through sequence homology searching approaches (Ma and Johnson, 2012). Homology searching further increases proteome coverage, by annotating also "partially correct" sequences (sequence tags), which are common "by-products" of the *de novo* sequencing process (Ma and Johnson, 2012).

Moreover, *de novo* sequencing may serve as a direct measure of the proportion of unsequenced members in a community. In a similar manner, the usefulness of *de novo* sequencing for evaluating the target sequence database completeness, or "suitability," has been demonstrated only recently (Johnson et al., 2020).

On the other hand, *de novo* peptide sequencing strongly depends on high-quality mass spectrometric data and efficient sequence annotation tools. Therefore, *de novo* sequencing commonly provides fewer spectral identifications when compared with database search approaches (Medzihradszky and Chalkley, 2015). Nevertheless, whether *de novo* sequencing provides sufficient qualitative and quantitative information for (quantitative) metaproteomic applications has not been effectively established to date.

Over the past years, several high-performance *de novo* sequencing algorithms have been introduced (Tran et al., 2019; Ma et al., 2003; Behsaz et al., 2020), and some have also been proposed for taxonomic profiling applications (Lee et al., 2018; Mooradian et al., 2019). In addition, a number of advanced web-based services that support taxonomic and functional analyses of metaproteomic protein and peptide sequences have been introduced only recently (Mesuere et al., 2015; Boekel et al., 2015; Zhang et al., 2016; Singh et al., 2019; Riffle et al., 2017).

In this study, we introduce and evaluate a newly established *de novo* metaproteomics workflow for its quantitative performance and taxonomic resolution using synthetic and natural environmental community data. Furthermore, we introduce a new validation strategy and demonstrate how to establish the actual content of individual community members within community proteomics data. The new pipeline (NovoBridge) efficiently bridges individual components from *de novo* sequencing, automated annotation of sequences with taxonomies, a new validation procedure, and the provision of an output summary.

RESULTS

The presented metaproteomics pipeline employs conventional high-resolution shotgun proteomics data in which fragmentation spectra are subsequently translated into peptide sequence lists by *de novo* sequencing. The lists are then submitted by programmed access to the (public) Unipept database to retrieve taxonomic and metabolic information (Singh et al., 2019). Annotations are then processed by the established pipeline, which includes grouping into taxonomic branches and translation of enzyme commission numbers into KEGG pathways.

We investigated fundamental aspects and evaluated the performance of the established workflow using synthetic and natural microbial communities.

Taxonomic resolution

The first question concerns the taxonomic resolution that can be achieved when matching *de novo* peptide sequences against particularly large taxonomy databases to retrieve taxonomic and functional annotations. A large number of peptide sequences is common to several taxa and can therefore only be unique to a certain taxonomic ranking. Hence, the number of unique peptide sequences decreases from higher to lower taxonomic rankings. For example, because of the relatedness between taxa, there will be many more peptide sequences unique only to the phylum level compared with the more distinguished genus or species levels.

For our study, we aimed to retrieve taxonomic information from the Unipept database, which contains processed peptide sequences pre-allocated with taxonomic and functional annotations derived from the Uniprot database, using NCBI taxonomy (Mesuere et al., 2012, 2016). The Unipept ranking uses the hierarchical structure of the NCBI taxonomy for which consensus taxa have been determined using the lowest common ancestor approach (Mesuere et al., 2012). To test the Unipept database for the achievable taxonomic resolution, we generated *in silico* peptide sequences from >1,000 species retrieved from the NCBI reference sequences database (www.ncbi.nlm.nih.gov/refseq/). This provided for approximately 90% of all peptide sequences taxonomic annotations, but as expected, showed a steady decrease in the number of assigned peptides from higher to lower taxonomic rankings (= "drop-off rate"), with a particularly large drop between genus and species levels (Figure 1C). It is worth noting that deviations from this "drop-off rate" can be observed for species from highly sampled taxa and species with inconsistent taxonomic classifications. This impacts not only the quantitative performance but may also limit the taxonomic resolution, because a certain number of peptides is required for the identification of a respective taxon.

Furthermore, because there is no complete taxonomy database available, there is always a high likelihood of "unsequenced" community members—those that are not in the taxonomy database—being present in the community. Those retrieve annotations through related species mostly at higher taxonomic rankings and will therefore provide only a comparatively low taxonomic resolution.

A quantitative analysis should therefore aim to investigate the "drop-off rates" for individual taxonomic branches, in order to flag poorly quantitative traits. For this, *in silico* peptidomes may serve as highly useful comparators to establish the actual content of a member within the community proteomics data.

A validation procedure

De novo sequencing commonly generates a fraction of only partially correct peptide sequences. This raises the question of whether those incomplete sequences lead to false-positive assignments, which bias the taxonomic representation of the community.

As a measure of confidence for *de novo*-established peptide sequences, the software platform PEAKS provides the average local confidence (ALC) score, and DeepNovo, the p score (Ma et al., 2003; Tran et al., 2019, 2017). Although these parameters

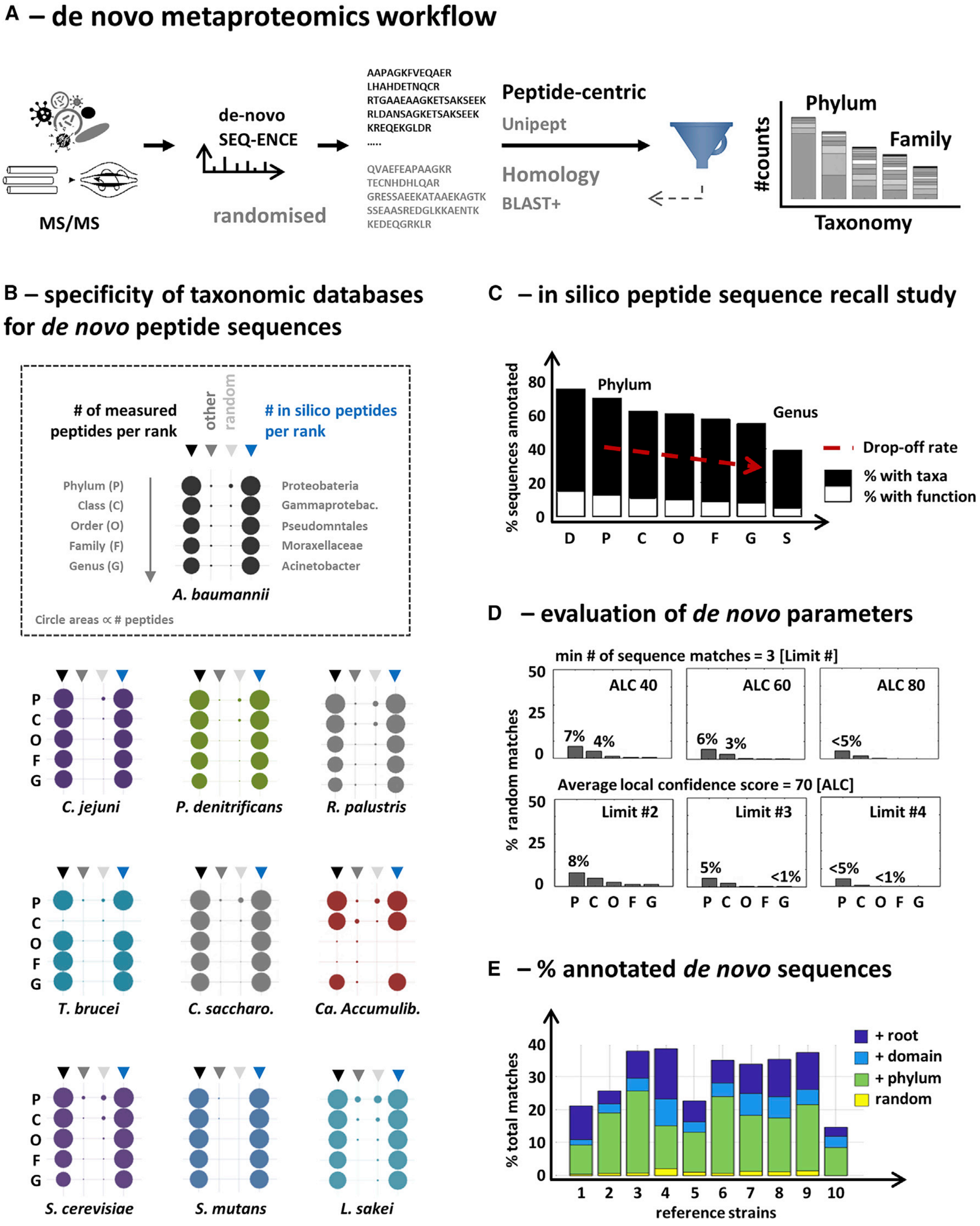


Figure 1. Overview of the *de novo* metaproteomics workflow and an evaluation of fundamental characteristics

(A) Shotgun metaproteomics workflow. Shotgun metaproteomic raw data from microbial communities are *de novo* sequenced and processed through the established pipeline as “correct” and randomized sequences. The peptide-centric approach accesses Unipept (Mesuere et al., 2016) to obtain taxonomic and functional annotations. Further processing includes grouping into taxonomic branches and translation of functional annotations into KEGG pathways. High-quality unmatched sequences are further made accessible for homology search approaches such as BLAST+.

(B) Specificity of taxonomy databases for *de novo* peptide sequence lists. Shotgun proteomic data from pure reference strains were *de novo* sequenced and processed through the established *de novo* metaproteomics pipeline to retrieve taxonomic annotations. The annotated sequences were then grouped into taxonomic lineages (phylum, class, order, family, and genus) and represented as circle graphs. The circle areas correlate to the normalized sequence counts of the respective taxonomic rank. Every reference strain is represented by four circle lanes: black triangle arrow, “# of measured peptides per rank,” which counts the number of peptide sequences annotated to the lineage of the target strain, e.g., *A. baumannii*; gray triangle arrow, “other,” which counts the number of

(legend continued on next page)

are useful for ranking *de novo* sequences based on their quality, an estimate on the actual number of incorrect sequences is not provided.

Consequently, additional measures are required to give confidence in the taxonomic representation achieved by *de novo*-generated sequences. A recently proposed solution employs a taxonomic database containing sequences not only in correct but also in reverse order. This strategy enables to make use of the widely employed target/decoy approach (Mooradian et al., 2019). However, database volumes are thereby duplicated, and considering only single taxonomic points does not allow performance of a quantitative investigation of the taxonomic profiles.

Therefore, we aimed not to randomize the target database sequences but to randomize the peptide query sequences instead. To qualify this approach, we processed proteomics data from pure reference species, once in correct order, and once after peptide sequence randomization. The randomized sequences retrieved a surprisingly large number of taxonomic annotations at the root (>20%) and super kingdom levels (>10%) but were consistently low for the lower taxonomic rankings (Figures 1B and 1D). Only small proportions of other taxa were observed, mostly related to culturing and sample preparation conditions, or the samples themselves (such as *virus L-A related proteins* for the yeast *S*. *cerevisiae*). Several of those unexpected matches were only identified at certain taxonomic levels, which underlines the importance of measuring the taxonomic profiles across several taxonomic rankings (e.g., from phylum, family, or genus level) rather than single taxonomic points (e.g., only genus level) (Figure 1B; Table S5).

Next, we constructed the theoretical drop-off rates using the reference proteomes of the test strains to investigate for "hidden" side populations, not covered by the taxonomic database. This, however, showed that the theoretical and the observed drop-off rates were very comparable, which confirmed the purity of the selected reference strains.

In summary, using the pure reference strain samples and the sequence randomization strategy, we could demonstrate that *de novo* sequence lists provide only small numbers of erroneous assignments at lower taxonomic rankings (phylum and genus).

Quantitative community profiling

Finally, we investigated the quantitative aspect when measuring more complex communities. Kleiner et al. only recently demonstrated the usefulness of metaproteomics for estimating species biomass contributions (Kleiner et al. 2017). Thereby, the authors generated highly useful metaproteomic reference data from synthetic communities consisting of species with "equal protein" and "equal cell" content. We *de novo* sequenced the publicly available raw data from both synthetic communities and subjected the obtained sequence lists to our data-processing pipeline. By employing the abovementioned multi-point taxonomic evaluation, we achieved a particularly good quantitative representation of the community as shown for the "equal protein" community (phylum and family) in Figure 2A. The 17 genus-level identifiers provided a comparably good correlation, although 3 strains did not provide sufficient unique peptides at this lower level. The same good species abundance correlation was achieved when analyzing another dataset of the same "equal cell" community, thereby also comparing 2 different *de novo* sequencing platforms, PEAKS and DeepNovo (Figure 3). Verification of parameters such as ALC scores and mass error, including species abundance correlations, obtained for the "equal cell" synthetic community are shown in Figures S1–S4.

Furthermore, we aimed to apply the *de novo* pipeline to datasets from two natural communities. Thereby, we first processed a publicly available metaproteomic dataset published by Mikan et al., representing microbiomes sampled from the Bering Sea (Mikan et al., 2020). We generated peptide sequences once using *de novo* sequencing and once using peptide-spectrum-matching employing the metagenomics constructed database published by the authors. Thereby, the taxonomic profiles

peptide sequences annotated to other taxonomic lineages than the target strain; light gray triangle arrow, "random," which counts the number of randomized peptide sequences which received a taxonomic annotation; blue triangle arrow, "# of *in silico* peptides per rank," which counts the number of *in silico* target strain sequences for every rank. The experiment confirms that erroneous or only partially correct *de novo* sequences only insignificantly interfere with the taxonomic representation of the metaproteomic sample. Furthermore, the low number of "other" strain assignments confirmed the purity of the selected reference strain samples. Except for the *in silico* experiments, the averages of duplicate analyses are shown.

(C) *In silico* proteome recall study. The bar graph shows the average number of *in silico* peptide sequences which retrieved taxonomic or "enzyme commission number" annotations. The *in silico* peptide sequences were generated from a large number of proteomes (>1,000, retrieved from the NCBI reference proteome database). The individual taxonomic rankings domain (D), phylum (P), class (C), order (O), family (F), genus (G), and species (S) are shown as separate bars. Approximately 90% of the peptides obtained taxonomic annotations (black bars), and 10%–20% retrieved additional functional annotations (enzyme commission numbers, white bars). The number of sequence annotations per taxon showed a steady decrease from the phylum to the genus level ("drop-off" rate, red arrow).

(D) Evaluation of *de novo* sequence quality parameters. The bar graph shows the average number of random sequences which obtained a taxonomic annotation, when considering different quality parameter thresholds. The randomized sequences were generated from the "correct" reference strains *de novo* sequence lists (excluding *T. brucei* and *Ca*. Accumulibacter). The quality parameter thresholds evaluated were the average local confidence score (ALC, PEAKS platform) and frequency limits (# of peptide sequences observed for an individual taxonomic identifier). ALCs below 60 and frequency limits <3 increased the percentage of random sequence annotations to >5%. Therefore, an ALC of 70 and a minimum of 3 sequence annotations per taxon were set as default thresholds for the experiments in this study.

(E) The percentage of annotated *de novo* sequences. The bar graph outlines the percentage of *de novo* sequences submitted to Unipept, which retrieved taxonomic annotations. The bars (1–10) represent the strains shown in Figure 1B (*A. baumannii*, top of the image; *L. sakei*, bottom right of the image). The blue bars represent all annotations including "root" level, which are sequences common to all domains of life; the light blue bars represent annotations assigned to domain level and lower; and the green bars show annotations assigned to phylum level and lower. The yellow bars indicate the average number of random annotations at the lower taxonomic rankings. The observed differences in the degree of sequence annotations are supposedly a consequence of differences in employed sample preparation protocols and instrumental setups. Therefore, although the percentage of assigned sequences are difficult to compare between different laboratories, those parameters are likely to provide a useful quality parameter when operations are standardized within one laboratory.

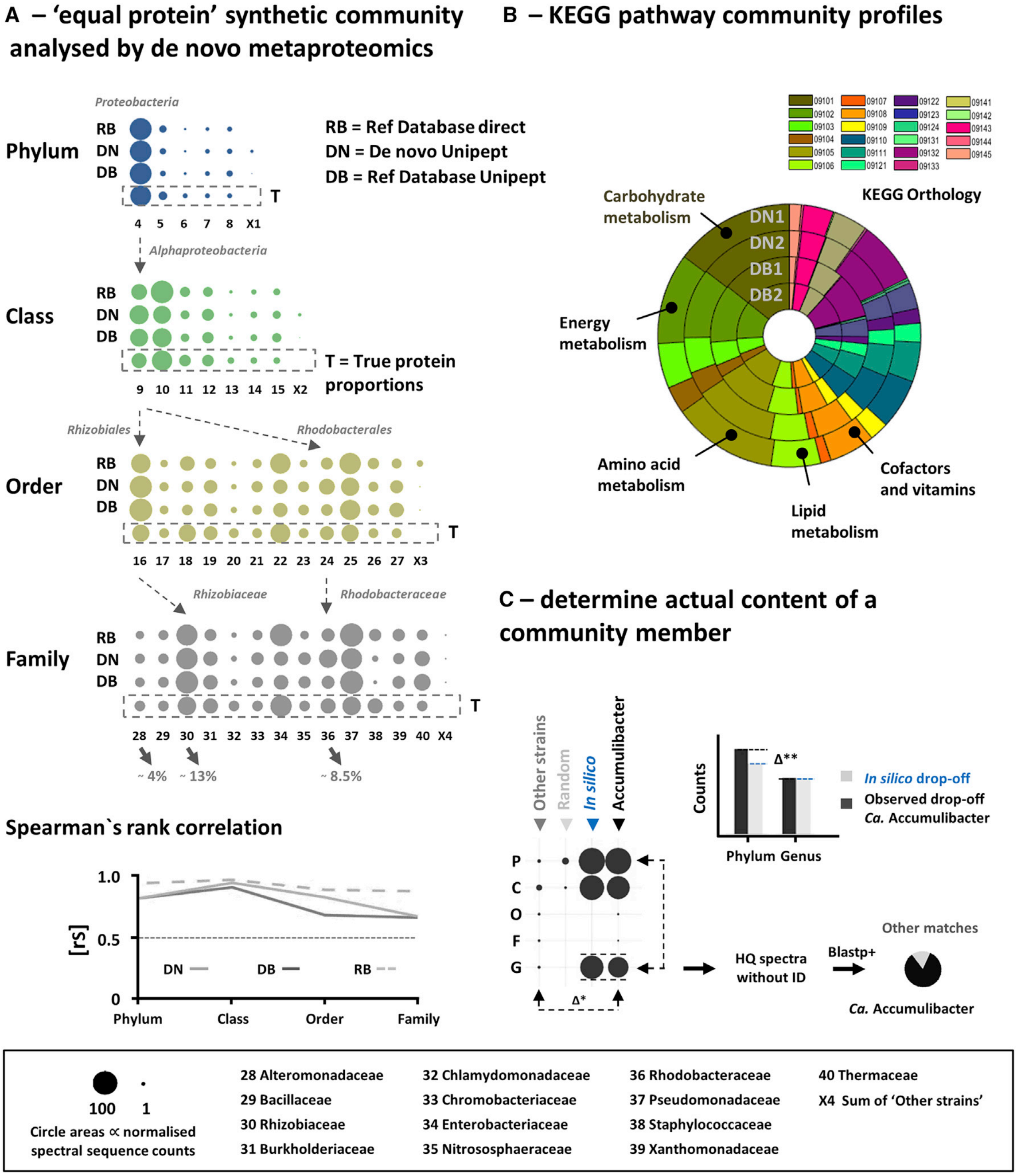


Figure 2. Quantitative taxonomic profiling of microbial communities

(A) Analyzing the community composition by *de novo* metaproteomics. Proteomics data from a synthetic community, as established by Kleiner et al. (Kleiner et al., 2017), were used to evaluate the quantitativeness of the established *de novo* metaproteomics workflow. For this, the raw data were once *de novo* sequenced and once analyzed using the constructed target database published by the authors. The taxonomic rankings from phylum and family are represented as circle graphs. Thereby, rows annotated with: "DN" show the protein abundances of each taxon using the *de novo* sequences; "DB" show the protein abundances obtained for each taxon using the sequences established by database matching; "RB" show the protein abundances obtained after grouping the taxon annotated database matched peptides directly; "T" shows the theoretical (true) protein abundances for each taxon. The circle areas correlate to the normalized spectral sequence counts of the respective taxon. All community members show abundance profiles, which strongly correlate to the expected/true (T) species protein abundances. The taxonomic lineages of *Rhizobiaceae* and *Rhodobacteriaceae* are outlined with arrows for exemplification purposes. Those account for approximately 13% and 8.5% of the total community protein content, respectively. Shown is the average of duplicate analyses. The taxonomic identifiers with the numbers 1–27 represent: (1) Bacteria, (2) Eukaryotes, (3) Archaea, (4) Proteobacteria, (5) Firmicutes, (6) Chlorophyta, (7) Thaumarchaeota, (8) Deinococcus-Thermus, (9) Alphaproteobacteria, (10) Gammaproteobacteria, (11) Bacilli, (12) Betaproteobacteria, (13) Chlorophyceae, (14) Nitrososphaeria, (15) Deinococci, (16) Rhizobiales, (17) Alteromonadales, (18) Bacillales, (19) Burkholderiales, (20) Chlamydomonadales, (21) Neisseriales, (22) Enterobacterales, (23) Nitrososphaerales, (24) Rhodobacterales, (25) Pseudomonadales, (26) Xanthomonadales, and (27) Thermales. The lower graph shows the Spearman's rank correlation between the

(legend continued on next page)

between both approaches were highly comparable (Figure S5A), where only some of the very-low-abundance members were not resolved by the *de novo* approach. However, the metaproteomics approaches indicate a stronger contribution of the Alphaproteobacteria, compared with the 16S rRNA sequencing data published by the authors (Figure S5A).

The second dataset was derived from the metaproteomic analysis of a wastewater treatment plant community, published by Hansen et al. (Hansen et al., 2014). The authors investigated different protein extraction procedures to maximize extraction reducibility and community coverage. We therefore analyzed the mass spectrometric raw data obtained from the most efficient protocol through the established metaproteomics pipeline. Furthermore, the observed community profiles appeared very comparable between the *de novo*-generated peptide sequences and the (metagenomics) database search peptide sequence matches. Again, differences were only observed in the very-low-abundance community members. Moreover, the *de novo* phylum-level profile of the *de novo* dataset was found highly comparable to qFISH data established from the same community at an earlier time point (Albertsen et al., 2012) (Figure S6).

Database incompleteness and spectral volume dependency

To evaluate the impact of incomplete databases, we simulated scenarios where the taxonomies present in the microbiomes are not covered by the taxonomic database (e.g., Unipept). As a consequence, measured peptide sequences from those taxonomic identifiers would only match to related taxa (potentially) present in the same database.

Interestingly, when all species or genera (present in the synthetic "Kleiner community" or Bering Sea microbiome) were removed, the obtained community profiles at the higher taxonomic rankings (e.g., family/phylum) changed only marginally compared with the unfiltered database output (Figures S2B and S5B). However, an incomplete taxonomic database unavoidably limits the achievable taxonomic resolution. Nevertheless, this is expected to become an increasingly less impactful parameter over time. Proteome/genome databases have been rapidly expanding over the past years, and this is likely to further accelerate due to the continuous advancements in sequencing technologies. For example, the RefSeq database expanded by approximately 25,000 entries for bacteria alone over 5 years (November 2015–November 2020), which corresponds to an increase of >100 million protein sequences in that period of time (www.ncbi.nlm.nih.gov/refseq/statistics/).

Furthermore, to evaluate the dependency of achieving a comprehensive taxonomic coverage on the volume of peptide sequences, we performed a random downsampling of the peptide sequences from the synthetic "Kleiner community" and the Bering sea metaproteomics datasets. To evaluate the impact of the downsampling procedure, we plotted the number of the remaining sequences against the obtained number of taxonomic identifiers. This showed a plateau for the number of obtained taxonomic identifiers at a certain percentage of the original number of peptide sequences (approximately 40%–60% for the "Kleiner community" and approximately 80%–90% for the Bering Sea microbiome) for both metaproteomic experiments. This means that (nearly) no new taxonomic identifiers were obtained after this fraction of peptide sequences and that the acquired datasets therefore indeed comprehensively cover the microbiome biomass.

Establishing the actual content of a community member

Finally, we aimed to investigate the usefulness of *in silico* drop-off curves (the decrease in the number of peptides, assigned to different taxonomic ranks from the higher to the lower taxonomic ranks using the (lowest common ancestor [LCA] approach) and BLAST+ homology search, for investigating the actual content of an enrichment culture. Evaluating the drop-off rates of a lineage enables one to evaluate whether the observed numbers of peptides at the higher taxonomic levels (e.g., phylum level = Proteobacteria) are aligned with the number of peptides observed at lower taxonomic levels (e.g., *Ca.*

peptide sequence lists (obtained by DB matching, "DB," and DN sequencing, "DN") and the expected protein abundance ratios ("T"). Overall, the correlation to the expected protein abundances was strong for both sequence list approaches (e.g., >0.82 for the DN sequence lists from phylum to order, and 0.67 only at the family level, considering all taxonomic identifiers, including "x"). The very comparable correlation between the *de novo* and the database-matching generated sequence lists confirms the high quality of the *de novo* established peptide sequences. The profiles obtained after directly grouping the database spectrum-matched peptides show, as expected, a slightly better correlation. Therefore, the difference between these profiles and the profiles obtained by the sequence lists shows the impact of the database, such as sequence coverage and volumes. The database used for the database-matching experiments consisted of the reference proteomes of the strains present in the synthetic community, and therefore represented a comparatively focused, complete, and non-redundant database. Moreover, the very large and generic Unipept peptide sequence database, used to annotate the peptide sequence lists, contained only closely related taxa for some strains (e.g., for *Roseobacter sp*. AK199).

(B) KEGG pathway community profiles. The graphs compare profiles for the major KEGG categories "metabolism" and "genetic information processing," obtained by sequence lists from *de novo* (outer circles) or peptide-spectrum matching approaches (inner circles) of the "equal protein" community. Both *de novo* (DN) and the database (DB) sequences provide very comparable profiles. Nevertheless, since peptide sequence lists are compared against a large genomic space, sequences can be matched to several enzymes or different pathways, which may inflate functional annotations. See also Figure S6.

(C) Establishing the actual contribution of community members. The *de novo* metaproteomic analysis of a *Ca*. Accumulibacter enrichment culture suggests a very high enrichment (>95%, "other" versus "Accumulibacter," Δ*). Furthermore, comparing the experimental with the *in silico* "drop-off" rates, shows only a discrepancy of approximately 17% (small bar graphs, Δ**). To investigate for potential "hidden" members not covered by the taxonomic database, the high-quality (HQ) unmatched sequences (top 20% fraction based on ALC scores) were analyzed using BLAST+ for homolog sequences. Thereby, more than 80% of the newly retrieved annotations were again assigned to *Ca*. Accumulibacter (small pie chart), confirming the content estimated after drop-off correction. The individual circle graph columns represent: black triangle arrow, "# of measured peptides per rank," which counts the peptide sequences annotated to the lineage of *Ca*. Accumulibacter; blue triangle arrow, "# of *in silico* peptides per rank," which represents the number of *Ca*. Accumulibacter in silico sequences per taxon; light gray triangle arrow, "random," which counts the number of randomized peptide sequences which received a taxonomic annotation; gray triangle arrow, "other," which counts the number of measured peptide sequences annotated to other taxonomic lineages than *Ca*. Accumulibacter. The circle areas correspond to spectral sequence (peptide) counts for the respective taxonomic ranking.

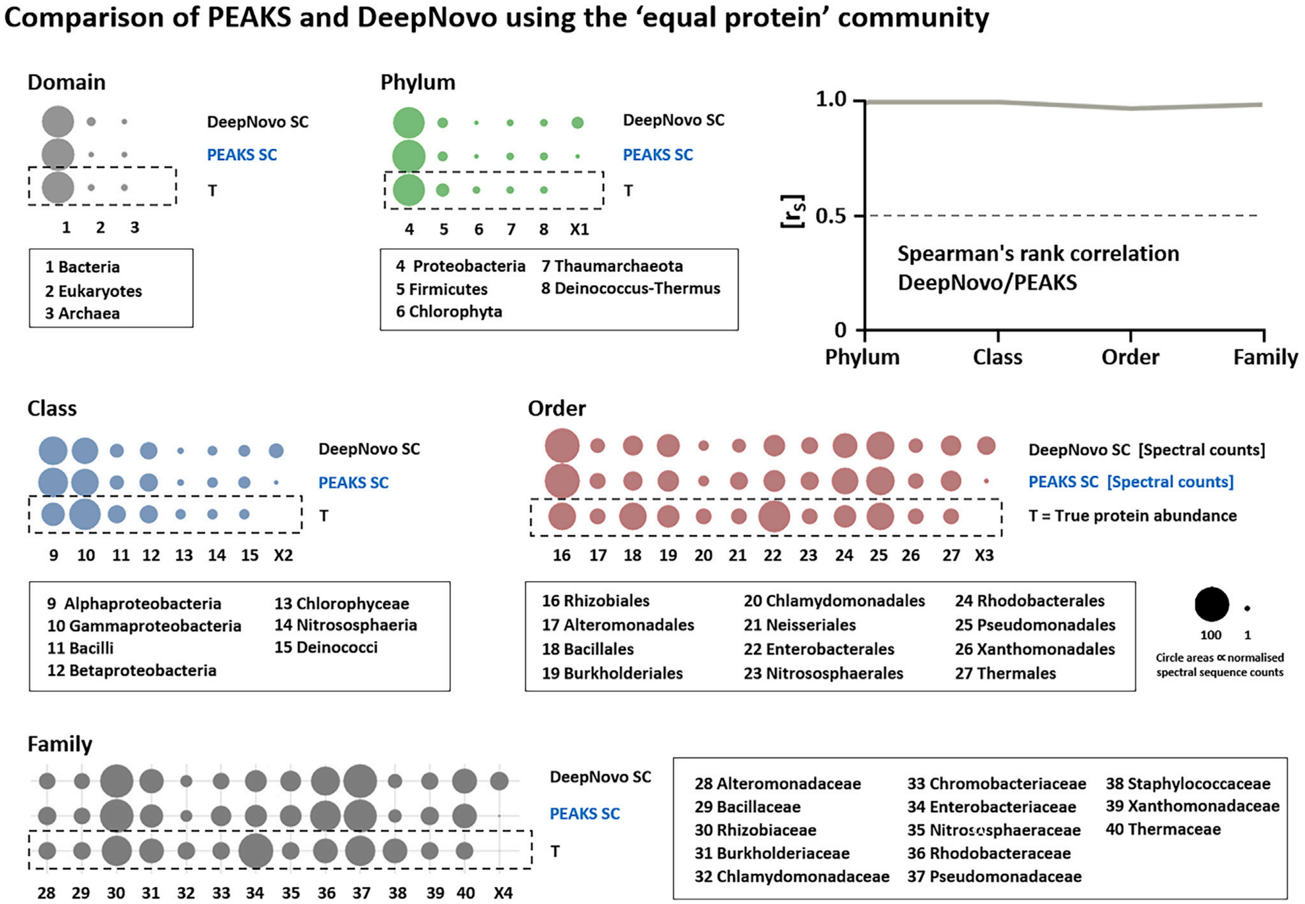


Figure 3. Comparison of microbiome profiles established by PEAKS and DeepNovo

(A) Community profiles of the "equal protein" community established by PEAKS and DeepNovo. The circle graphs show the taxonomic profiles obtained from the "equal protein" community (Kleiner et al., 2017) established by PEAKS or DeepNovo. *De novo* sequence lists from both platforms were processed by the established *de novo* metaproteomics pipeline using the same parameters. "T" represents the true abundance of the respective community members (dashed box). "PEAKS SC" represents the established profiles obtained from the PEAKS *de novo* sequences using spectral sequence counting. "DeepNovo SC" represents profiles obtained from the DeepNovo *de novo* sequences using spectral sequence counting. The unexpected, "other" taxonomic annotations were summed and are shown as circles labeled with "X." The experiment demonstrates that both tools provide very comparable taxonomic profiles and only differ in the proportions of the unexpected "other" matches. The circles represent the average of 2 analyses, where the circle areas correlate to the normalized spectral sequence counts. The left upper graph shows the Spearman's rank correlation of the taxonomic profiles between PEAKS and DeepNovo. The very strong correlation (r_S between 1.0–0.97, from phylum–family, considering the expected taxonomic identifiers, 1–40) confirms that both tools provide highly comparable peptide sequence lists.

Accumulibacter, genus level). This approach allows one to evaluate whether the proportion of proteobacteria is likely derived from *Ca*. Accumulibacter or whether there are other lineages present that are not covered by the database. *Ca*. Accumulibacter has been described frequently as showing strong discrepancies in the proposed community contribution when comparing between FISH- and 16S RNA sequencing-based techniques (Stokholm-Bjerregaard et al., 2017). Therefore, we analyzed an Accumulibacter enrichment culture metaproteomic dataset through the described pipeline, which indicated a particularly high enrichment (Figure 2C, approximately 98% at the genus level [Δ*], in contrast to 16S RNA data for the same reactor at an earlier time point of approximately 34% [Da Silva et al., 2018]). When comparing the experimental drop-off rate for the lineage of *Ca*. Accumulibacter with the *in silico* constructed drop-off curve, we observed a discrepancy of only approximately 17% (Δ**), meaning that nearly all sequences assigned to proteobacteria translate to the *Ca*. Accumulibacter genus-level annotations. Nevertheless, to fully exclude significant quantities of potential other populations—e.g., from other phyla, not captured by the (Unipept) database—the high-quality unmatched sequences (top 20% based on ALC scores) were analyzed using BLAST+ against the non-redundant NCBI protein sequence database (for the sake of speed using a local installation). Thereby, approximately 83% of newly retrieved (genus level) sequences could be attributed again to *Ca*. Accumulibacter (Table S7; Figure 2D), reflecting the estimated content obtained after drop-off correction. Moreover, the high degree of enrichment indicated by our metaproteomics experiments is in good agreement with the observed phosphate accumulation activity, observed for this culture during lab experiments (data not shown).

Determining the fraction of unmatched (high-quality) spectra has already been proposed as an indicator for the presence of community members not captured by the database (Kleiner et al., 2017; Johnson et al., 2020). The fraction of unmatched high-quality spectra, however, may considerably depend on the applied analytical procedures. The same was observed for the reference strains used in this study, in which raw data were acquired from different laboratories and thus showed large variations in their fraction of peptides that obtained taxonomic annotations (Figure 1E). Although this approach appears very promising, it may provide misleading conclusions if not corrected for individual analytical procedures.

DISCUSSION

Metaproteomics has emerged as one of the most promising post-genomics approaches to study microbial dynamics in nature or in the context of human health, such as the microbial dynamics of the gut microbiome (Behsaz et al., 2020; Timmins-Schiffman et al., 2017). However, common metaproteomics workflows require the laborious construction of high-quality protein sequence databases. Thus, spectrum-matching algorithms are challenged by very large databases or unsequenced community members not covered by the database. Furthermore, the quantitative aspect is often only poorly supported, despite being utmost important when investigating community dynamics.

Here, we introduce a newly established *de novo* metaproteomics workflow, which enables quantitative profiling of microbial communities within a very short analysis time. We provide a systematic evaluation of the taxonomic resolution and quantitative performance using reference strains and natural communities. Thereby, we introduce a validation procedure and demonstrate how to establish the actual content of community members within community proteomics data. The established pipeline automates data filtering, taxonomic annotation, additional validation procedures, grouping, and reporting of taxonomic and functional outputs with only minutes of processing time for a typical shotgun metaproteomics dataset. In comparison, metagenomics including database construction, or the analysis of the mass spectrometric data against very large generic databases, typically requires (several) days of processing time.

Notably, because our approach is database independent, it generates peptide sequences also from “not-in-the-database” community members, making them accessible for further interpretation. The achievable resolution in *de novo* metaproteomics, however, depends not only on the taxonomic database but also on the abundance of the individual community members. Moreover, a completely metagenomics-independent evaluation of a community, containing only unsequenced community members, will likely provide only a comparatively low taxonomic resolution or provide assignments only to the closest taxa present in the database.

The evaluation we performed demonstrates that the highest accuracy is achieved up to the family level, which could therefore be suggested as the default level of operation. However, an improved resolution and quantification (number of peptide matches) for the lower taxonomic rankings—such as genus or even species level—could currently be achieved by performing a *de novo*/database-matching hybrid approach. *De novo*-established taxonomies thereby guide the construction of a focused database from large generic databases, which subsequently can be used for comparatively efficient peptide-spectrum-matching experiments.

Nevertheless, the current vast technical advancements in the field of mass spectrometry and sequencing algorithms are likely to continue improving the quality of the sequencing spectra and thus the number of correct *de novo* sequence annotations in the near future. Ultimately, this will strengthen and expand the scope of *de novo* metaproteomics as either a hybrid, orthogonal, or stand-alone approach.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Application of publicly available data
  - ○ Whole cell lysate proteolytic digestion
  - ○ Shotgun metaproteomic analysis
  - ○ PEAKS and DeepNovo raw data processing
  - ○ NovoBridge data processing pipeline
  - ○ Taxonomic annotation of metagenomic sequence database
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ In silico evaluation of ‘drop-off curves’
  - ○ Simulation of peptide sequence database lacking specific taxonomies
  - ○ Simulation of metaproteomics data containing different volumes of peptide sequences
  - ○ Spearman rank correlation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2021.04.003.

ACKNOWLEDGMENTS

The authors are grateful for valuable discussions with colleagues from the Department of Biotechnology, TU Delft, and acknowledge Carol de Ram for support with sample processing and the SIAM (Soehngen Institute of Anaerobic Microbiology) consortium for funding. The authors further acknowledge all who made their mass spectrometric raw data available through public repositories, in particular the M. Kleiner and M. Strous labs, for establishing the synthetic community data.

AUTHOR CONTRIBUTIONS

H.B.C.K., Y.M.L., MC.M.v.L., and M. Pabst designed and evaluated experiments; H.B.C.K., B.A., and M. Pabst performed in-house proteomics experiments on microbial samples. H.B.C.K., L.G.d.S., M. Pronk, and M. Pabst established and analyzed datasets for true content determination. H.B.C.K., C.T., and M. Pabst performed coding and evaluated codes. H.B.C.K. and M. Pabst performed processing of the proteomics and metaproteomics raw data. H.B.C.K. performed annotation of metagenomics DB. H.B.C.K. and M. Pabst wrote the manuscript. All authors discussed the results and revised the manuscript text.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 16, 2020
Revised: January 16, 2021
Accepted: April 7, 2021
Published: April 30, 2021

REFERENCES

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.

Behsaz, B., Mohimani, H., Gurevich, A., Prjibelski, A., Fisher, M., Vargas, F., Smarr, L., Dorrestein, P.C., Mylne, J.S., and Pevzner, P.A. (2020). De novo peptide sequencing reveals many cyclopeptides in the human gut and other environments. Cell Syst. *10*, 99–108.e5.

Boekel, J., Chilton, J.M., Cooke, I.R., Horvatovich, P.L., Jagtap, P.D., Käll, L., Lehtiö, J., Lukasse, P., Moerland, P.D., and Griffin, T.J. (2015). Multi-omic data analysis using Galaxy. Nat. Biotechnol. *33*, 137–139.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Camacho, C., Madden, T., Tao, T., Agarwala, R., and Morgulis, A. (2008). BLAST® Command Line Applications User Manual (National Center for Biotechnology Information).

Da Silva, L.G., Gamez, K.O., Gomes, J.C., Akkermans, K., Welles, L., Abbas, B., Van Loosdrecht, M.C., and Wahl, S.A. (2018). Revealing metabolic flexibility of Candidatus Accumulibacter phosphatis through redox cofactor analysis and metabolic network modeling. bioRxiv https://www.biorxiv.org/content/10.1101/458331v2.full#:∼:text=This%20metabolic%20flexibility%20is%20enabled,to%20NADPH%20dependent%20PHA%20synthesis.

Hansen, S.H., Stensballe, A., Nielsen, P.H., and Herbst, F.A. (2014). Metaproteomics: evaluation of protein extraction from activated sludge. Proteomics *14*, 2535–2539.

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.

Johnson, R.S., Searle, B.C., Nunn, B.L., Gilmore, J.M., Phillips, M., Amemiya, C.T., Heck, M., and MacCoss, M.J. (2020). Assessing protein sequence database suitability using de novo sequencing. Mol. Cell. Proteomics *19*, 198–208.

Junqueira, M., Spirin, V., Balbuena, T.S., Thomas, H., Adzhubei, I., Sunyaev, S., and Shevchenko, A. (2008). Protein identification pipeline for the homology-driven proteomics. J. Proteomics *71*, 346–356.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kleiner, M., Thorson, E., Sharp, C.E., Dong, X., Liu, D., Li, C., and Strous, M. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. Nat. Commun. *8*, 1558.

Lee, J.-Y., Mitchell, H.D., Burnet, M.C., Jenson, S.C., Merkley, E.D., Shukla, A.K., Nakayasu, E.S., and Payne, S. (2018). Proteomics of natural bacterial isolates powered by deep learning-based de novo identification. bioRxiv https://www.biorxiv.org/content/10.1101/428334v1.

Ma, B., and Johnson, R. (2012). De novo sequencing and homology searching. Mol. Cell. Proteomics *11*, O111.014902.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. *17*, 2337–2342.

Madden, T. (2013). Chapter 16: The BLAST sequence analysis tool. The NCBI handbook, Second Edition (National Center for Biotechnology Information)), pp. 1–15.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Lett. *583*, 3966–3973.

Martin, F., and Uroz, S. (2016). Microbial environmental genomics (MEG) (Springer).

Medzihradszky, K.F., and Chalkley, R.J. (2015). Lessons in de novo peptide sequencing by tandem mass spectrometry. Mass Spectrom. Rev. *34*, 43–63.

Mesuere, B., Debyser, G., Aerts, M., Devreese, B., Vandamme, P., and Dawyndt, P. (2015). The Unipept metaproteomics analysis pipeline. Proteomics *15*, 1437–1442.

Mesuere, B., Devreese, B., Debyser, G., Aerts, M., Vandamme, P., and Dawyndt, P. (2012). Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. J. Proteome Res. *11*, 5773–5780.

Mesuere, B., Willems, T., Van Der Jeugt, F., Devreese, B., Vandamme, P., and Dawyndt, P. (2016). Unipept web services for metaproteomics analysis. Bioinformatics *32*, 1746–1748.

Mikan, M.P., Harvey, H.R., Timmins-Schiffman, E., Riffle, M., May, D.H., Salter, I., Noble, W.S., and Nunn, B.L. (2020). Metaproteomics reveal that rapid perturbations in organic matter prioritize functional restructuring over taxonomy in western Arctic Ocean microbiomes. ISME J. *14*, 39–52.

Mooradian, A.D., Van Der Post, S., Naegle, K.M., and Held, J.M. (2019). ProteoClade: a taxonomic toolkit for multi-species and metaproteomic analysis. bioRxiv https://www.biorxiv.org/content/10.1101/793455v1.

Muth, T., Kolmeder, C.A., Salojärvi, J., Keskitalo, S., Varjosalo, M., Verdam, F.J., Rensen, S.S., Reichl, U., De Vos, W.M., Rapp, E., and Martens, L. (2015). Navigating through metaproteomics data: a logbook of database searching. Proteomics *15*, 3439–3453.

Muth, T., Renard, B.Y., and Martens, L. (2016). Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. Expert Rev. Proteomics *13*, 757–769.

Potgieter, M.G., Nel, A.J., Tabb, D.L., Fortuin, S., Garnett, S., Blackburn, J., and Mulder, N.J. (2019). MetaNovo: a probabilistic approach to peptide and polymorphism discovery in complex mass spectrometry datasets. bioRxiv https://www.biorxiv.org/content/10.1101/605550v6.

Riffle, M., May, D.H., Timmins-Schiffman, E., Mikan, M.P., Jaschob, D., Noble, W.S., and Nunn, B.L. (2017). MetaGOmics: a web-based tool for peptide-centric functional and taxonomic analysis of metaproteomics data. Proteomes *6*, 2.

Sayers, E. (2009). The E-utilities in-depth: parameters, syntax and more. In Entrez Programming Utilities Help (National Center for Biotechnology Information), pp. 1–156.

Singh, R.G., 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.

Stokholm-Bjerregaard, M., Mcilroy, S.J., Nierychlo, M., Karst, S.M., Albertsen, M., and Nielsen, P.H. (2017). A critical assessment of the microorganisms proposed to be important to enhanced biological phosphorus removal in full-scale wastewater treatment systems. Front. Microbiol. *8*, 718.

Timmins-Schiffman, E., May, D.H., Mikan, M., Riffle, M., Frazar, C., Harvey, H.R., Noble, W.S., and Nunn, B.L. (2017). Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. ISME J. *11*, 309–314.

Tran, N.H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., and Li, M. (2019). Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nat. Methods *16*, 63–66.

Tran, N.H., Zhang, X., Xin, L., Shan, B., and Li, M. (2017). De novo peptide sequencing by deep learning. Proc. Natl. Acad. Sci. USA *114*, 8247–8252.

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

Xiao, J., Tanca, A., Jia, B., Yang, R., Wang, B., Zhang, Y., and Li, J. (2018). Metagenomic taxonomy-guided database-searching strategy for improving metaproteomic analysis. J. Proteome Res. *17*, 1596–1605.

Zhang, X., Ning, Z., Mayne, J., Moore, J.I., Li, J., Butcher, J., Deeke, S.A., Chen, R., Chiang, C.-K., Wen, M., et al. (2016). MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiome *4*, 31.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Reagents/Equipment		
Sequencing Grade Modified Trypsin	Promega	V5111
Oasis HLB 96-well Plate	Waters	WAT058951
B-PER™ Bacterial Protein Extraction Reagent	Thermo Scientific	78243
Y-PER™ Yeast Protein Extraction Reagent	Thermo Scientific	78991
QE plus Orbitrap mass spectrometer	Thermo Scientific	
EASY nano LC 1200	Thermo Scientific	
Acclaim PepMap RSLC RP C18 reverse phase, (75μm x 150mm, 2μm)	Thermo Scientific	164568
Deposited Data		
Candidatus Accumulibacter phosphatis	PXD016992	MP_Ser_01122018_Accum_2hr_DDA01.raw MP_Ser_01122018_Accum_2hr_DDA02.raw
Acinetobacter baumannii	PXD011302	Nsco_20170712_LFQ_P_negative_ABCA1_WC_B1.raw Nsco_20170712_LFQ_P_negative_ABCA1_WC_B2.raw
Campylobacter jejuni	PXD005306	Cj_media_DOC_R1_23Feb15_Arwen_14-12-03.raw Cj_media_DOC_R2_23Feb15_Arwen_14-12-03.raw
clostridium saccharolyticum	PXD016992	MP_AM27072018_S1SC_No1_DDA01.raw MP_AM27072018_S1SC_No2_DDA01.raw
Lactobacillus sakei	PXD011417	BBM_079_P064_01_FRH_12_R1.raw BBM_079_P064_01_FRH_13_R1.raw
Paracoccus denitrificans	PXD013274	PDL-2-5.raw PDL-2-6.raw
Rhodopseudomonas palustris	PXD013729	Biodiversity_R_palustris_long_WT_Ar_20d_01_QEP_20Aug18_Wally_18-07-04 Biodiversity_R_palustris_long_WT_Ar_20d_02_QEP_20Aug18_Wally_18-07-04
Streptococcus mutans	PXD006735	WT1.raw WT2.raw
Saccharomyces cerevisiae	PXD016992	MP18112018_yeast_Y1_1uL_DDA01.raw MP18112018_yeast_Y2_1uL_DDA01.raw
Trypanosoma brucei	PXD009073	mb161104_01.raw mb161104_02.raw
Synthetic community equal protein	PXD006118	Run2_P1_2000ng.raw Run2_P2_2000ng.raw Run4_P1_2000ng.raw Run4_P2_2000ng.raw
Synthetic community equal cell	PXD006118	Run2_C2_2000ng.raw Run2_C4_2000ng.raw
Bering Sea T1 (#14 BS.T1.Control)	PXD008780	2014_Sept_08_BeringSea32.raw 2014_Sept_08_BeringSea33.raw
Bering Sea T6 (#49 BS.T6.Control)	PXD008780	2014_Sept_08_BeringSea36.raw 2014_Sept_08_BeringSea37.raw
Bering Sea T10 (#60 BS.T10.Control)	PXD008780	2014_Sept_08_BeringSea38.raw 2014_Sept_08_BeringSea39.raw

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Wastewater treatment plant community	PXD000862	FAH_SludgeExtr_B_BB_1.raw FAH_SludgeExtr_B_BB_2.raw
Software, Databases and Algorithms		
Matlab	The MathWorks, Inc.	https://www.mathworks.com
Python	Python Software Foundation	https://www.python.org/
NovoBridge	This paper	https://github.com/hbckleikamp/NovoBridge
PEAKS Studio X	Bioinformatics solutions Inc.	https://www.bioinfor.com/
BLAST	NCBI blast-2.9.0+	https://blast.ncbi.nlm.nih.gov ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+
DeepNovo	DeepNovo_2017	https://github.com/nh2tran/DeepNovo
Taxonomy NCBI server	https://www.ncbi.nlm.nih.gov/books/NBK21100/ (Handbook)	https://ftp.ncbi.nih.gov/pub/taxonomy/
KEGG BRITE Database	https://www.kegg.jp/	https://www.genome.jp/kegg-bin/get_htext?ko00001
Unipept	Gent University	https://unipept.ugent.be/apidocs
Diamond	v2.0.6	https://github.com/bbuchfink/diamond

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Martin Pabst (m.pabst@tudelft.nl).

Materials availability

This study did not generate new materials.

Data and code availability

Generated mass spectrometric raw data have been deposited at ProteomeXchange server and are publicly available under the project code PXD016992. Moreover, this paper analyzes existing, publicly available data. These datasets' accession numbers are summarized in the key resource table and are outlined in more detail in the method details part.

A conversion (including description) of the original NovoBridge code into Python code is available via github.com: https://github.com/hbckleikamp/NovoBridge. The Matlab codes are freely available upon request from the Lead Contact.

The functions used to generate the figures reported in this paper are available via Matlab 2017b or later versions plus Bioinformatics Toolbox (https://nl.mathworks.com/) and their use is described in the STAR Methods.

Any additional information required to reproduce this work is available from the Lead Contact.

METHOD DETAILS

Application of publicly available data

The synthetic community proteomic raw data were downloaded from ProteomXchange server project PXD006118, established by M. Kleiner and M. Strous labs (Kleiner et al., 2017). Protein content and taxonomic lineages of the synthetic community samples used have been further outlined in the Tables S2 and S3. Due to incomplete coverage of viral strains in the Unipept database, viruses were not further considered in the quantitative analysis. Shotgun proteomic raw data from *Rhodopseudomonas palustris* were retrieved from the project PXD013729 generated by E. Nakayasu, Pacific Northwest National Laboratory and C.S. Harwood, University of Washington, *Campylobacter jejuni* raw data were retrieved from PXD005306 generated by M. Monroe, and J. Adkins, Pacific Northwest National Laboratory, Paracoccus denitrificans raw data were downloaded from project PXD013274 generated by T. J. Erb and M. Glatter, MPI Marburg, respectively. *Lactobacillus sakei* PXD011417 from C. Ludwig, Bavarian Center for Biomolecular Mass Spectrometry (BayBioMS), Technical University Munich. Acinetobacter baumannii PXD011302 from M. Feldmann Washington University School of Medicine and J. Scott, University of Melbourne; Streptococcus mutans PXD006735 from J. Koh and K.C. Rice, University of Florida; Trypanosoma brucei PXD009073 from J.W. Dupuy form Centre de Génomique Fonctionnelle Bordeaux, France and M. Boshart, from Ludwig-Maximilians-University Munich Martinsried, Germany. Additional environmental community reference dataset shown in Figure S5, was obtained from PXD008780, as published by B. L. Nunn and E. Timmins-Schiffman of the University of Washington (Mikan et al., 2020). The waste water treatment plant community data shown in Figure S6, were obtained from processing PXD000862, which were published by S. A. Hansen and F. A. Herbst, from Aalborg University.(Hansen et al., 2014) Comparative database-search peptide sequences were retrieved from published supplemental information, which were filtered for sequences with PEP<0.01 before processing through the pipeline. qFISH abundances were obtained from the paper published by Albertsen et al. (2012),(Albertsen et al., 2012) using the 'GetData Graph Digitizer' tool.

Whole cell lysate proteolytic digestion

Approximately 25-50mg biomass (wet weight) of each cell pellet/material were homogenised by beads beating in TEAB/B-PER reagent (Thermo Scientific™, for bacterial cells such as *Ca*. Accumulibactor phosphatis enrichment and Clostridium sacch.) or Y-PER reagent (Thermo Scientific™, for yeast cells), respectively. The supernatant was collected by centrifugation at 14.000xg. The protein content was precipitated using TCA (1 vol TCA 100 w/v % to 4 vol sample) followed by washing with ice cold acetone. The protein pellet was resuspended in 200 mM ammonium bicarbonate containing 6M Urea, reduced in a 10 mM DTT solution at 40C for 1 hour, and alkylated using 20 mM IAA in the dark, at room temperature, for 30 minutes. The solution was diluted to below 1 M Urea and digested using sequencing grade Trypsin at a protease to protein ratio of approximately 1:50. Peptides were desalted using Oasis HLB solid phase extraction cartridges (Waters Corporation) according to the protocol provided by the manufacturer, speed-vac dried and resuspended in 3% acetonitrile in H_2O, containing 0.1% formic acid.

Shotgun metaproteomic analysis

An aliquot of each sample was analysed using a nano-liquid-chromatography system consisting of an EASY nano LC 1200 equipped with an Acclaim PepMap RSLC RP C18 reverse phase column (75μm x 150mm, 2μm) coupled to a QE plus Orbitrap mass spectrometer (Thermo, Germany). Solvent A was H_2O containing 0.1% formic acid, and solvent B consisted of 80% acetonitrile in H_2O, containing 0.1% formic acid. The flow rate was maintained at 300 nL/min. The Orbitrap was operated in top 10 data dependent acquisition (DDA) mode, acquiring peptide signals form 350-1400 m/z, at 70K resolution in MS1 with an AGC target of 3e6 and max IT of 100ms. For yeast, approx. 250ng protein digest were analysed using a short linear gradient from 4 to 30% B over 32.5 minutes, and further to 70% B over 12.5 minutes. MS2 acquisition was performed at 17.5K resolution, with an AGC target of 2e5, and a max IT of 54ms, using a NCE of 28. Unassigned, singly charged as well as 7, 8 and >8 charged mass peaks were excluded. For bacterial samples, approx. 100ng protein digest were analysed using a linear gradient from 5-30% B over 85 minutes and further to 75% B over 25 minutes. MS2 acquisition was performed at 17.5K resolution, with an AGC target of 1e5, and a max IT of 54ms, at a NCE of 30. Unassigned, singly charged, 8 and >8 times charged mass peaks were excluded.

PEAKS and DeepNovo raw data processing

Comparative database-search peptide sequences were retrieved from published supplemental information, which were filtered for sequences with PEP<0.01 before processing through the pipeline. qFISH abundances were obtained from the paper published by Albertsen et al. (2012),(Albertsen et al., 2012) using the 'GetData Graph Digitizer' tool. Peptide sequencing procedures: Mass spectrometric raw data were processed using PEAKS Studio X (Bioinformatics Solutions Inc., Canada)(Ma et al., 2003) for database search and *de novo* sequencing, or DeepNovo(Tran et al., 2019) for comparative *de novo* sequencing studies. Both, *de novo* sequencing and database search was performed allowing 15ppm parent ion and 0.015Da fragment mass error (depending on the acquisition, slightly more tolerant parameters such as 20ppm/0.02Da were applied). Carbamidomethylation was set as fixed and methionine oxidation as variable modifications. Database search allowed in addition N/Q deamidation as variable modifications. The same settings were applied to DeepNovo where applicable, otherwise software default settings were used. Database search further used decoy fusion for estimation of false discovery rates (FDR) and subsequent filtering of peptide spectrum matches for 1% FDR. Only the top ranked *de novo* sequence annotations were considered for processing. Both, sequence lists were further processed through the same metaproteomics pipeline. Except for the comparative study, shown in Figure 3, PEAKS was used to generate sequence lists.

NovoBridge data processing pipeline

The NovoBridge Matlab pipeline is freely available upon request from the lead contact. A conversion (including description) of the original pipeline into Python code is available via github.com.https://github.com/hbckleikamp/NovoBridge

A Matlab 'main script' was constructed that links together functions for pre-filtering, sequence randomisation, automated submissions to Unipept to obtain taxonomic and functional information, threshold filtering, taxonomic grouping and visualisation of output data. The pipeline was established and tested with peptide sequence lists generated by *de novo* sequencing using PEAKS or DeepNovo, from high-resolution QE Orbitrap shotgun proteomics raw data. The script was constructed using Matlab 2017b and 2019 respectively.

Function 1, pre-filtering, sequence randomisation and Unipept submission

The first part of the script involves importing peptide sequence lists (obtained from PEAKS/DeepNovo) into the Matlab environment and to perform pre-filtering based on the sequence annotation quality parameters. The default pre-filtering thresholds were set to ALC scores >40, less than 20ppm mass error and a minimum peptide length of 7 amino acids. Sequence lists were 'cleaned' from peptide modification annotations and mass errors were corrected for mass drifts. The Matlab 'rand' function was further used to generate additional randomised sequences from imported *de novo* lists. Thereby, the order of amino acids in front of the cleavage site (R or K) of every sequence was randomised, keeping original sequence parameters attached. Automated sequence submission to Unipept was done using Unipept's inbuilt API (https://unipept.ugent.be/apidocs) option.(Mesuere et al., 2015) For retrieving taxonomic information, '*pep2lca*' including the options *'&equate_il=true'*, to equate leucine and isoleucine, were used. Further, '*&extra=true &names=true*' are specified to get the complete taxonomic lineage and the names of every taxonomic rank. The script automatically filters for the main categories super kingdom, phylum, class, order family, genus and species. The '*pept2-funct*' combined with the option *'&equate_il=true'* was used to retrieve additional EC number information.(Mesuere et al., 2015) Thereby, a single peptide sequence can generate multiple EC numbers or pathways which cause functional inference and inflation, particularly when searching against a large sequence database space. For this study, only the top scoring peptide sequence per scan was considered.

Function 2, compositional analysis

The compositional analysis considered the major taxonomic categories super kingdom, phylum, class, order, family, genus and species. Depending on data quality/abundance, lower ranks (such as species or genus) were excluded from quantitative analysis/representation due to low numbers or insufficient annotations. In a first step, tables were filtered for sequences with ALCs >70 (or less than -0.1 for DeepNovo), and a mass error of less than 15 ppm. To exclude random matches from erroneous *de novo* sequences or low-abundance signals, a taxonomic identifier of a branch was only considered when occurring at least 3 times. Frequency and ALC cut-offs/thresholds were established using randomised sequences of the pure reference strains. Remaining taxonomic branches are further grouped and visualised using the 'bar(x..,stacked)' function in Matlab for both, absolute and normalized peptide sequence counts (or areas/intensities, respectively). Visualising the relative abundances of the individual community members were performed using circle graphs using the 'surf' function in Matlab. Circle areas represent thereby the number of normalised spectral sequence counts and show the average of 2 separate analyses (except stated otherwise). True/expected abundances of individual community members of the synthetic communities were retrieved from the supplemental information materials, as published by Kleiner et al. (2017).(Kleiner et al., 2017)

Function 3, functional analysis

KEGG pathways, from global classifications to individual conversions within a pathway, correspond to the KEGG orthology (KO) codes.(Kanehisa and Goto, 2000) Therefore, we established a script, which translates the retrieved enzyme commission numbers (EC) into KO codes. This was done by integrating the KEGG annotation database, downloaded from https://www.genome.jp/kegg-bin/get_htext?ko00001 (10/19), into the Matlab environment. The analysis of the global community metabolic functions, considered thereby only branches which were also used for compositional analysis. Sequences assigned to root and super kingdom levels were excluded. EC assignments matched more than twice (based on unique spectral sequence counts) were further translated into KO codes, normalised to the total number of spectral sequence counts and grouped into pathways. Obtained functional community profiles were visualised using heat maps or circle graphs based on KEGG pathways/category levels 2 (global) and 3 (carbohydrate and energy metabolism). Further information regarding 'KEGG pathway categories' are outlined below.(Kanehisa and Goto, 2000)* Heat maps were generated using the 'heatmap' function, and circle graphs were created using Matlab's 'donut.m' function as available through www.mathworks.com 'file exchange' website.

- * Second category codes: 09101 Carbohydrate metabolism, 09102 Energy metabolism, 09103 Lipid metabolism, 09104 Nucleotide metabolism, 09105 Amino acid metabolism, 09106 Metabolism of other amino acids, 09107 Glycan biosynthesis and metabolism, 09108 Metabolism of cofactors and vitamins, 09109 Metabolism of terpenoids and polyketides, 09110 Biosynthesis of other secondary metabolites, 09111 Xenobiotics biodegradation and metabolism, 09121 Transcription 09122 Translation, 09123 Folding, sorting and degradation, 09124 Replication and repair, 09131 Membrane transport, 09132 Signal transduction, 09133 Signalling molecules and interaction, 09141 Transport and catabolism, 09143 Cell growth and death, 09144 Cellular community – eukaryotes, 09145 Cellular community – prokaryotes, 09142 Cell motility.
- * Third category codes: 00010 Glycolysis/Gluconeogenesis, 00020 Citrate cycle (TCA cycle), 00030 Pentose phosphate pathway, 00040 Pentose and glucuronate interconversions, 00051 Fructose and mannose metabolism, 00052 Galactose metabolism, 00053 Ascorbate and aldarate metabolism, 00500 Starch and sucrose metabolism, 00520 Amino sugar and nucleotide sugar metabolism, 00620 Pyruvate metabolism, 00630 Glyoxylate and dicarboxylate metabolism, 00640 Propanoate metabolism, 00650 Butanoate metabolism, 00660 C5-Branched dibasic acid metabolism, 00562 Inositol phosphate metabolism, 00190 Oxidative phosphorylation, 00195 Photosynthesis, 00196 Photosynthesis - antenna proteins, 00710 Carbon fixation in photosynthetic organisms, 00720 Carbon fixation pathways in prokaryotes, 00680 Methane metabolism, 00910 Nitrogen metabolism, 00920 Sulfur metabolism. *www.genome.jp/kegg/pathway.html

Function 4. Peptide sequence outputs

To interface with other tools, a peptide sequence table output is provided in form of '.xls' or '.mat' files. Thereby either all sequences, only identified or non-identified sequences can be selected. The later can be filtered for high quality spectra, such as selecting for the top 20% (based on ALC score), which was exemplified using the BLAST+ homology search module, to investigate for potential un-sequenced community members.

Alternative BLAST+ search of unidentified spectra

Alternatively, high quality unidentified *de novo* sequences were subjected to BLAST+ homology search(Madden, 2013; Camacho et al., 2008). Even though there are homology search web services available(Junqueira et al., 2008), we used a local installation to maintain sufficient throughput and integrity with the established *de novo* metaproteomics pipeline. For this ncbi-blast-2.9.0+ and the non-redundant protein sequence database 'nr.gz' (segmented for more efficient use, due to size) were downloaded from the

NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/blast, updated 12/19) and installed on a local windows 10 workstation. BLAST searches were operated using the Matlab 'system' command function. All BLAST searches used the PAM30 scoring matrix. The top 5 assignments per query sequence (based on bit-scores) were combined and filtered for best e values and scores, respectively. Taxon ID and name databases were downloaded from the NCBI server. Full taxonomic lineages were retrieved form NCBI using E-utilities calls 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=taxonomy&id=' and

'taxurl_right='&retmode=xml'.(Sayers, 2009)

Taxonomic annotation of metagenomic sequence database

The metagenomics protein (assembly) sequence database from Mikan et al.(Mikan et al., 2020) was annotated with taxonomies using DIAMOND v2.0.6 and the non-redundant bacterial NCBI RefSeq database (Reference Sequence, release 203) and default parameters.(Buchfink et al., 2015) Furthermore, for the top 20% of sequence alignments (based on bit score), complete lineages were determined using the Unipept taxonomy API. The LCA was established using the LCA approach, and the taxonomy ID was retrieved from the prot.accession2taxid database downloaded from the NCBI repository (ftp.ncbi.nih.gov).

QUANTIFICATION AND STATISTICAL ANALYSIS

In silico evaluation of 'drop-off curves'

Large-scale reference proteomes in silico study: A large number of reference proteomes (>1500) covering all 3 domains of life were retrieved from the NCBI reference database (www.ncbi.nlm.nih.gov/refseq/). *In silico* trypsin cleavage, random selection of 1K sequences (each) and programmed submission to Unipept was done and determination of drop-off curves were performed using Matlab2017b (The MathWorks, Inc., US). Reference proteome *in silico* drop off analysis: A random selection of 3.5K unique trypsin cleaved *in silico* peptides (7-15 amino acids length, to approximate real samples) for the pure strains analysed in this study, as listed in the Table S4, was performed using MATLAB's bioinformatics toolbox. The *in silico* peptidomes were processed through the same NovoBridge pipeline, as described above.

Simulation of peptide sequence database lacking specific taxonomies

Peptide sequences were submitted (as usual), using the above-mentioned NovoBridge pipeline, to Unipept to retrieve taxonomic lineages based on the lowest common ancestor (LCA) approach using 'pept2lca'. Unlike in the default processing pipeline, peptide sequences that obtained a class or lower taxonomic annotation were further annotated with taxIDs using the 'pept2prot' and 'taxonomy' API to obtain all underlying taxonomic lineages. This moreover enabled to selectively remove lineages for peptides from taxonomic entries at defined taxonomic rankings, e.g. species, genus, or family. Following the selective filtering, the remaining taxonomic lineages (of the respective peptide sequences) were regrouped using the LCA approach and further processed through the conventional NovoBridge pipeline with default parameters. The evaluation of the obtained taxonomic profiles was compared to the true (synthetic Kleiner community) or the initially determined taxonomic profiles (Bering Sea community) by determining the Spearman rank-order correlation coefficient (r_s) using the MATLAB 'corr' function and the 'Spearman' option.

Simulation of metaproteomics data containing different volumes of peptide sequences

To evaluate the taxonomic profiles obtained from metaproteomics containing different amounts of spectral information, we performed a (random) down-sampling of the peptide sequences. For this, the metaproteomics data from the synthetic 'Kleiner community' or the natural Bering Sea microbiome were down-sampled stepwise to finally contain only 90, 80, 70, 60, 50, 40, 30, 20, 10, 8, 6, 4, 2, or 1% of the original number of peptide sequences. The remaining sequences were further processed through the NovoBridge pipeline using default parameters. The change in the obtained number of taxonomic identifiers (at different taxonomic ranks) was compared using line plots created with the MATLAB 'plot' function and hill equation curve fitting.

Spearman rank correlation

Generally, the evaluation of the obtained taxonomic profiles were compared to the true or otherwise comparatively determined taxonomic profiles by determining the Spearman rank-order correlation coefficient (r_s) using the MATLAB 'corr' function and the 'Spearman' option.