Journal of Biotechnology 261 (2017) 24–36

Contents lists available at ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Review

Challenges and perspectives of metaproteomic data analysis

Robert Heyer[a,*], Kay Schallert[a], Roman Zoun[b], Beatrice Becher[a], Gunter Saake[b], Dirk Benndorf[a,c,**]

[a] Otto von Guericke University, Bioprocess Engineering, Universitätsplatz 2, 39106 Magdeburg, Germany
[b] Otto von Guericke University, Institute for Technical and Business Information Systems, Universitätsplatz 2, 39106 Magdeburg, Germany
[c] Max Planck Institute for Dynamics of Complex Technical Systems, Bioprocess Engineering, Sandtorstraße 1, 39106, Magdeburg, Germany

ARTICLE INFO

Keywords:
Bioinformatics
Software
Big data
Environmental proteomics
Microbial communities
Mass spectrometry

ABSTRACT

In nature microorganisms live in complex microbial communities. Comprehensive taxonomic and functional knowledge about microbial communities supports medical and technical application such as fecal diagnostics as well as operation of biogas plants or waste water treatment plants. Furthermore, microbial communities are crucial for the global carbon and nitrogen cycle in soil and in the ocean. Among the methods available for investigation of microbial communities, metaproteomics can approximate the activity of microorganisms by investigating the protein content of a sample. Although metaproteomics is a very powerful method, issues within the bioinformatic evaluation impede its success. In particular, construction of databases for protein identification, grouping of redundant proteins as well as taxonomic and functional annotation pose big challenges. Furthermore, growing amounts of data within a metaproteomics study require dedicated algorithms and software. This review summarizes recent metaproteomics software and addresses the introduced issues in detail.

1. Introduction

Microorganisms represent 50–78% of Earth‘s total biomass (Kallmeyer et al., 2012) and occur in all environments. Some microorganisms produce biomass by photosynthesis whereas others act as composers and degrade dead biomass. Microbial species live in complex microbial communities in which they have to compete or cooperate with each other. Understanding the functioning of the microbial communities is important, because microbial communities in the human gut effect health (Erickson et al., 2012; Heintz-Buschart et al., 2016; Kolmeder et al., 2016) and several technical applications such as waste water treatment plants (Püttker et al., 2015; Wilmes et al., 2008) and biogas plants (Abram et al., 2011; Hanreich et al., 2012) rely on the metabolic activity of microbial communities.

Methods for the investigation of microbial communities target the microbial cells, their genes, their transcripts, their proteins and their metabolites (Heyer et al., 2015). Since proteins carry out most functions in cells, including catalysis of biochemical reactions, transport and cell structure, protein amounts correlate quite well with microbial activity (Wilmes and Bond, 2006). The investigation of all proteins from one species is called proteomics. In contrast metaproteomics is the study of proteins from multiple organisms. It was introduced by Wilmes and Bond (2006, 2004) and Rodriguez-Valera (2004). The typical metaproteomics workflow comprises protein extraction and purification, tryptic digestion into peptides, protein or peptide separation and tandem mass spectrometry (MS/MS) analysis. Proteins are identified by comparing experimental mass spectra and theoretical mass spectra predicted from comprehensive protein databases. For a detailed discussion about the metaproteomics workflow please refer to Hettich et al. (2013), Becher et al. (2013), Heyer et al. (2015), Wohlbrand et al., (2013). Up to now most metaproteomics studies characterize the taxonomic and functional composition of complex microbial communities in their specific environment (Abram et al., 2011; Kan et al., 2005; Ram et al., 2005; Wilmes and Bond, 2006). A few recent studies additionally correlated the taxonomic and functional composition with certain environmental/process parameters or diseases (Erickson et al., 2012; Heyer et al., 2016; Kolmeder et al., 2016). However, three issues within bioinformatic data evaluation hampered previous

Abbreviations: CPU, central processing unit; COG, clusters of orthologous groups; DBMS, database management system (DBMS); de.NBI, German Network for Bioinformatics Infrastructure; EC, enzyme commission number; eggNOG, evolutionary genealogy of genes non-supervised orthologous; FDR, false discovery rate; GPU, graphical processing unit; GO, gene ontologies; iPath, interactive pathways explorer; LC, liquid chromatography; LCA, lowest common ancestor; KO, KEGG ontologies; MPA, MetaProteomeAnalyzer; MS, mass spectrometer; MS/MS, tandem mass spectrometer; *m*/*z*-ratio, mass-to-charge ratio; NoSQL, not only SQL; SQL, structured query language

[*] Corresponding author.

[**] Corresponding author at: Otto von Guericke University, Bioprocess Engineering, Universitätsplatz 2, 39106 Magdeburg, Germany.

E-mail addresses: heyer@mpi-magdeburg.mpg.de (R. Heyer), kay.schallert@ovgu.de (K. Schallert), roman.zoun@ovgu.de (R. Zoun), beatrice.becher@st.ovgu.de (B. Becher), saake@iti.cs.uni-magdeburg.de (G. Saake), benndorf@mpi-magdeburg.mpg.de (D. Benndorf).

http://dx.doi.org/10.1016/j.jbiotec.2017.06.1201
Received 15 February 2017; Received in revised form 20 June 2017; Accepted 23 June 2017
Available online 27 June 2017

metaproteomics studies (Muth et al., 2013).

First, metaproteomes consist of up to 1000 different species (Schlüter et al., 2008). Due to high complexity metaproteomics data analysis requires a greater computational effort, necessitating bigger hard drives, more memory, more processors and more efficient algorithms. A main issue is the database search against comprehensive protein databases. Whereas handling of small protein databases below 1 GB is not critical, usage of the entire NCBI reference database requires extended computational time and may fail due to software or hardware limitations.

Second, identical peptides belonging to homologous proteins cause redundant protein identification (Herbst et al., 2016). As a result taxonomic and functional interpretation of results becomes ambiguous. A peptide may belong to the lactate dehydrogenase (EC. 1.1.1.27) (1.1.1.27) of different members of the genus *Lactobacillus*, which ferment sugars to lactate. But it may also belong to some representatives of the order *Clostridiales fermenting* lactate to acetate (Kohrs et al., 2014).

Third, protein identification is difficult if the taxonomic composition is unknown or protein entries are missing from protein databases. For example the UniProt/TrEMBL database contains only proteins from 698,745 species (http://www.ebi.ac.uk/uniprot/TrEMBLstats, status 16.12.2016), but the number of microbial species on Earth is estimated to be up to one trillion (Locey and Lennon, 2016). Thereby, already small changes in the protein sequence between related microorganisms have a big impact on protein identification. One mutation in every tenth amino acid leads to completely different tryptic peptides which hinder the identification of any peptide for the investigated protein. Thus, researchers started to sequence metagenomes alongside metaproteomics studies (Ram et al., 2005; Tyson et al., 2004). Alternatively, they use metagenomes from similar samples for protein identification.

As a consequence of these issues, standard proteomics software is often insufficient for metaproteomics studies missing the identification of unsequenced species or the comprehensive taxonomic and functional description of microbial communities. Thus, researchers favor special tools. Therefore, this review provides an overview about dedicated metaproteomics software and bioinformatic strategies.

In addition to two previous reviews on bioinformatics in metaproteomics (Muth et al., 2013, 2016) we present the impact of combining metagenomes on protein identification and address future hardware requirements and handling of big data.

After a brief introduction current metaproteomics software tools are discussed. Subsequently, this review illuminates the creation of protein databases for protein identification investigating several biogas plant samples in a use case. Then the grouping of redundant protein identifications, the evaluation of taxonomic and functional results as well as quantification in metaproteomics studies are discussed. Finally, data storage and deployment solutions for big data as well as future challenges, perspectives and demand for metaproteomics software are considered.

2. Status of proteomics software and latest trends

For the comprehensive bioinformatic processing of MS data different software tools exist. These include software for peak picking in MS-spectra, software for protein identification via database search algorithms and tools for comparison of protein expression patterns. A comprehensive summary of all these software tools can be found in the OMIC tools database (http://omictools.com/, retrieved: 09-02-2017, (Henry et al., 2014)) and in several reviews (Cappadona et al., 2012; Gonzalez-Galarza et al., 2012).

Latest trends in proteomics software are the development of proteomics tool libraries such as OpenMS (Sturm et al., 2008), Compomics (Barsnes et al., 2011) or Trans-Proteomic Pipeline (Keller and Shteynberg, 2011). These libraries comprise software tools for each step of the processing workflow, ranging from data management to data analysis. Noteworthy are also webservices, such as Expasy (Gasteiger et al., 2003), which provide a collection of small bioinformatic tools for biochemical analyses of proteins.

Repositories for MS-data such as PRIDE are used to enable long-term storage and to make published MS-data available to other researchers (Vizcaino et al., 2016). In this context general formats for exchange of MS results are necessary. Current standard in the proteomics community are the mzIdentML format (Jones et al., 2012), mzTab format (Griss et al., 2014) and mzML format (Martens et al., 2011).

Recent proteomics software combines several database search algorithms. For example, the SeachGUI tool (Vaudel et al., 2011) enables the parallel protein database search with eight different database search algorithms. Further developments are software tools for improved MS-operation and quantification. Search items for these developments are "data independent acquisition" (Doerr, 2015), "multiple and single reaction monitoring" (Colangelo et al., 2013) as well as "absolute quantification" (Cappadona et al., 2012).

Within the last years many powerful software tools were developed but their use was often restricted to a few scientific groups. Reasons were missing maintenance or availability after funding periods ended. Furthermore, many biological research groups lack bioinformatic skills to set up comprehensive software workflows or client-server architectures. In some cases even the conversion of data into the required input formats fail. In order to tackle these problems governments started to fund the collection, maintenance and support of research software tools. Examples are the Galaxy project (https://usegalaxy.org/, retrieved: 09-02-2017), (Afgan et al., 2016), ELIXIR (https://www.elixir-europe.org/, retrieved: 09-02-2017, (Crosswell and Thornton, 2012)) or de.NBI (https://www.denbi.de/, retrieved: 09-02-2017).

3. Software dedicated for metaproteomics

To address the three issues specific to metaproteomics bioinformatic data evaluation, researchers started to develop special software tools and workflows [Table 1, Fig. 1]. These tools apply different concepts, which will be discussed later. Graph 2Pep/Graph2Pro (Tang et al., 2016) and Compile (Chatterjee et al., 2016) focus on tailoring protein databases for optimal protein identification. UniPept (Mesuere et al., 2015), Prophane (Schneider et al., 2011), Megan CE (Huson et al., 2016) and Pipasic (Penzlin et al., 2014) enable taxonomic analysis, functional data evaluation and/or protein grouping. Additionally, several groups assembled comprehensive software workflows for metaproteomics, e.g. Galaxy-P (Jagtap et al., 2015), MetaPro-IQ (Zhang et al., 2016), MetaProteomeAnalyzer (Muth et al., 2015a) and others (Heintz-Buschart et al., 2016; May et al., 2016; Tanca et al., 2013). Among these workflows, the MPA is particularly user-friendly. It allows the user to control the entire bioinformatic workflow via an intuitive graphical user interface. Another noteworthy metaproteomics software tool is MetaProSIP (Sachsenberg et al., 2015). It supports the detection and quantification of isotope ratios for Protein-SIP experiments.

To ensure comparability of results between all these tools, standards for data exchange are crucial (Timmins-Schiffman et al., 2017). Consequentially, the Human Proteomics Standard Initiative is planning to extend the proteomics mzIdentML format in order to support metaproteomics data. Version 1.2.0 of the mzIdentML format (Jones et al., 2012) will support the representation of redundant protein groups (http://www.psidev.info/mzidentml, retrieved: 09-02-2017).

Another often neglected aspect is the reproducibility of results using different metaproteomics software tools. So far, only Tanca et al. (2013) tested their complete metaproteomics workflow for a defined mixed culture of nine different microorganisms. A comparison where multiple research groups evaluate an identical sample would also be desirable.

4. Construction of user databases for protein identification

Protein database selection affects the number of identified proteins as well as the identified taxonomies and identification increases. In

Table 1
Overview about metaproteomic specific issues and appropriated software resp. bioinformatic strategies.

Issue	Solution/bioinformatic strategie	Reference
Grouping of redundant homologous proteins	1. Flexible grouping to metaproteins based on protein, peptide and taxonomy similarity 2. Grouping by shared peptide	MetaProteomeAnalyzer (Muth et al., 2015a) Prophane (Schneider et al., 2011)
Database tailoring	1. Two step database search 2. Metapeptide database 3. A "Graph-Centric Approach"	Jagtap et al. (2013) May et al. (2016) Graph2Pep/Graph2Prot (Zhang et al., 2016)
Taxonomic and functional evaluation	1. Calculate taxonomic value for each identified peptide (LCA) and visualize results 2. Calculate taxonomic value for peptides using peptide similarity estimation and expression level weighting 3. Taxonomic evaluation (LCA) and functional prediction using RPSBLAST or HMMER3 4. Taxonomic (LCA) and functional evaluation using ECs, KEGG Ontologies and KEGG Pathways. Unknown sequences can be annotated using Diamond. 5. Taxonomic (LCA) and functional evaluation using UniProt Keywords, ECs, KEGG Ontologies, KEGG Pathways. Unknown sequences can be annotated using BLAST.	UniPept (Mesuere et al., 2015) Pipasic (Penzlin et al., 2014) Prophane (Schneider et al., 2011) Megan CE (Huson et al., 2016) MPA (Muth et al., 2015a)
Storing and deployment of big data	1. Scalable set of sequence databases and specific database search algorithm	Compile and Blazmass (Chatterjee et al., 2016)
Quantitation	1. Detection and quantification of isotope ratios for Protein-SIP	MetaProSip (Sachsenberg et al., 2015)

consequence, the estimated FDR and thus, the threshold for accepting protein identifications are higher and may lead to the rejection of true protein identifications.

Optimal databases would only include proteins and posttranslational modifications present in the sample and detectable by MS. However, taxonomic composition and protein abundance are usually unknown for environmental samples. Furthermore, protein content between analyzed samples may differ significantly. Therefore, database selection is a challenging task (Muth et al., 2015b; Tanca et al., 2016). This issue is further complicated by the adherence of the research community to the FDR concept (Muth et al., 2015b).

Originally Elias and Gygi, (2007) established the FDR concept for comparable protein identification in pure culture proteomics. In particular, the FDR enables comparability between different mass spectrometers and database search algorithms. Subsequently, the proteomics community accepted the FDR calculation as the standard to control the quality of protein identifications. An FDR of 1% was defined as threshold (Barnouin, 2011). However, a condition for the successful estimation of the FDR is that the database fits well to the sample. This is not guaranteed for metaproteomics studies, resulting in inaccurate approximations of the FDR. Therefore, it would be desirable that the metaproteomics community revises the FDR concept questioning the decoy based approach. Instead protein identifications could be classified using machine learning approaches.

Principally researches have two options to construct their database for metaproteomics studies. The first strategy is to sequence the whole metagenome or metatranscriptome [Fig. 2A] (Ram et al., 2005; Tyson et al., 2004) and to translate the genes to proteins by tools such as Transeq or Sixpack (http://www.ebi.ac.uk/Tools/st/, retrieved 07.06.2017). The second is to use comprehensive sequence databases [Fig. 2_1] and apply reasonable constraints. Recently, sequencing of metagenomes became affordable, due to high-throughput sequencing technologies such as Illumina sequencing (Bentley et al., 2008; Junemann et al., 2013; Junemann et al., 2014). However, several different processing states of metagenomes could be used as protein databases [Fig. 2A]. After Illumina sequencing and quality control, metagenome data are present as reads. Reads are short fragments of about 150 base pairs, which can be translated into about 50 amino acids [Fig. 2B]. Subsequently, the translated reads are assembled to contigs and redundant reads are removed [Fig. 2C]. In some high resolution metagenome studies, it is even possible to assemble the entire genome of single microorganisms (Campanaro et al., 2016). The disadvantage of reads and contigs is that all six reading frames are considered during the translation of DNA sequences into protein sequences. This multiplies the amount of data by six. Contigs may also contain several genes, which complicates the taxonomic and functional interpretation. Hence, genes are predicted from the contigs and non-coding DNA fragments are removed [Fig. 2D]. Therefore, assembled metagenomes with gene predictions are the preferable databases for protein identification. Sometimes it is even possible to reconstruct the whole genome of single microorganisms within the microbial community, which is called binning.

Since these assembled metagenome protein databases match the actual sample, FDR estimation should be valid. However, the bioinformatic workflow to assemble metagenomes can also influence the protein identification (Tanca et al., 2016). For example, during metagenome assembly redundant reads where only one amino acid differs are sometimes condensed into a single read. This ignores protein isoforms and can lead to the loss of protein identifications. In contrast, a high number of translated reads in a database decrease protein identifications due to an increase in the FDR. In line with these problems, some authors experienced a higher number of protein identifications with read databases instead of contig databases (Timmins-Schiffman et al., 2017). Better protein identification was also observed by Tang et al. (2016) applying a graph-centric usage of reads as database.

The sequencing of metatranscriptomes is similar to metagenome sequencing [Fig. 2A]. In principle only translation of RNA to DNA is required. Identification of proteins against metatranscriptomes is beneficial, since organisms only transcript genes that are currently used (Wilmes et al., 2015).

Sequencing a metagenome or metatranscriptome for each sample is not always possible due to the high cost and effort for the sequencing and the data processing. Thus, researchers use metagenomes from similar samples or comprehensive databases such as UniProtKB/SwissProt, UniProtKB/TrEMBL (UniProt, 2015), UniRef (Suzek et al., 2007), NCBI (Coordinators, 2017) or Ensemble (Yates et al., 2016) [Fig. 2_1]. Database searches against complete comprehensive databases require long computation times and decrease the number of identified proteins due to the overestimation of the FDR. Reasonable constraints on these comprehensive databases are therefore necessary. For example Jagtap et al. (2013) proposed to search in two steps. Taxonomies or proteins identified in the first error-tolerant search are used to restrict the protein database for the second search [Fig. 2_2]. This obviously increases computation times, but reduces the FDR and the threshold for protein identifications. In the end more proteins are identified, but how well this approximates the real FDR remains unclear. Another option for reduction of the FDR is to perform several searches against smaller sub databases and to merge their results afterwards (Muth et al., 2016; Tanca et al., 2016) [Fig. 2_3]. A more reasonable approach to constrain the protein database is taxonomic foreknowledge, because in some cases taxonomic composition of the sample is known (Tanca et al., 2016) [Fig. 2_4]. For example,

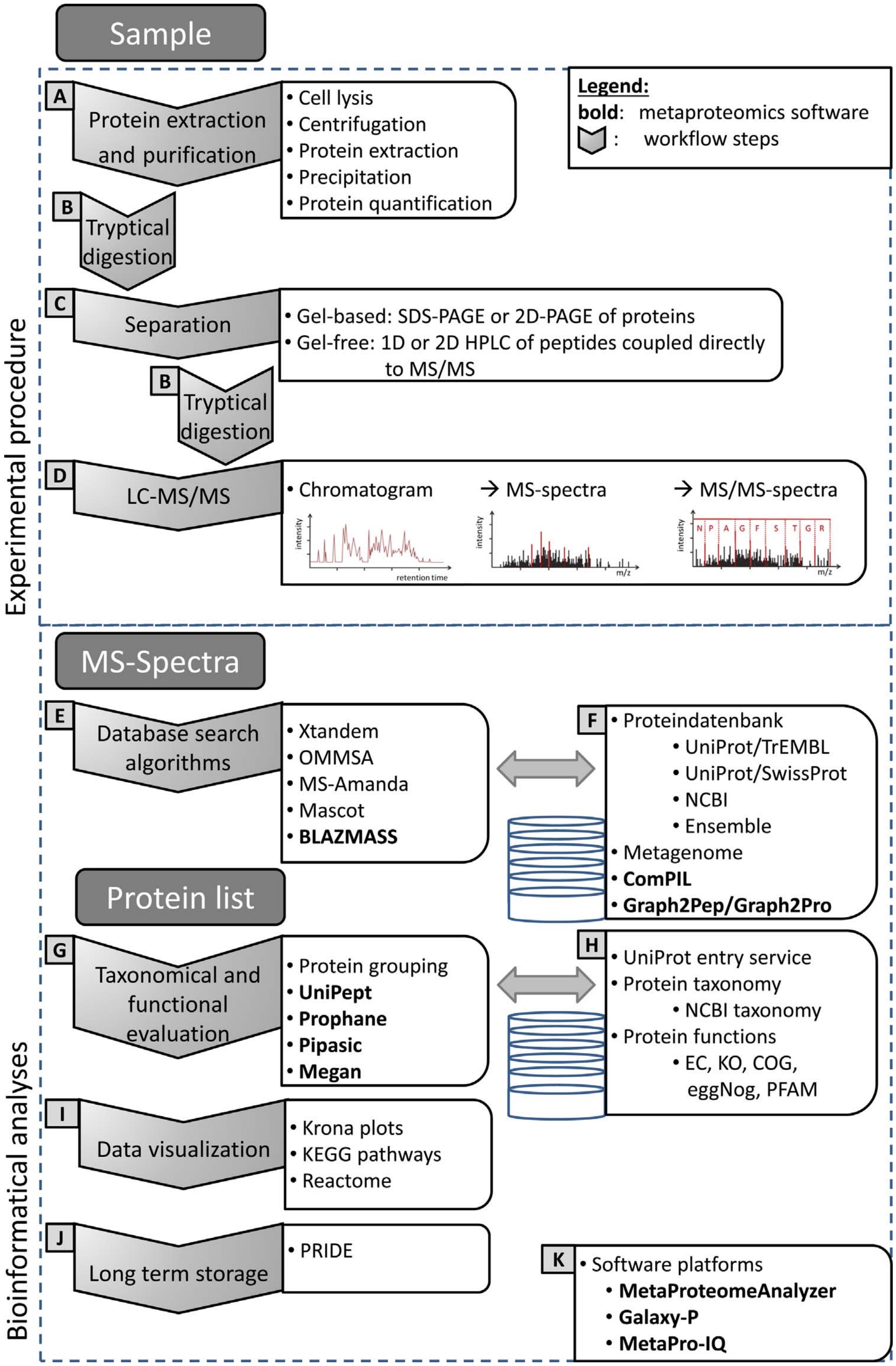

Fig. 1. Workflow for metaproteome analyses. Software tools specific for metaproteomics are highlighted in bold. Beside tools for single steps of the bioinformatic analysis also comprehensive software platforms are available (K).

sequencing of the 16S-rRNA gene provides a taxonomic profile. Nevertheless, performing pre-searches against all taxonomies can help to avoid excessive constraints on protein taxonomy during the actual searches.

A smart idea to decrease computational time for protein database searches was recently proposed by May et al. (2016). They searched against peptide databases instead of protein databases [Fig. 2E]. This reduces the size of the search space due to the grouping of identical peptides from homologous proteins.

To summarize, all strategies to constrain protein databases carry some pitfalls and we would recommend researchers to try different approaches. Despite all these strategies for protein database construction, inaccurate FDR estimation hampers metaproteomics studies. Solutions other than the target-decoy approach are required to validate protein identifications across different MS and database search algorithms. A promising step towards this direction represent semi-supervised machine learning algorithms such as the software tools Percolator (Kall et al., 2007) or Nokoi (Gonnelli et al., 2015). They distinguish correct and incorrect peptide-to-spectrum matches using a classificator based on learning algorithms from real data.

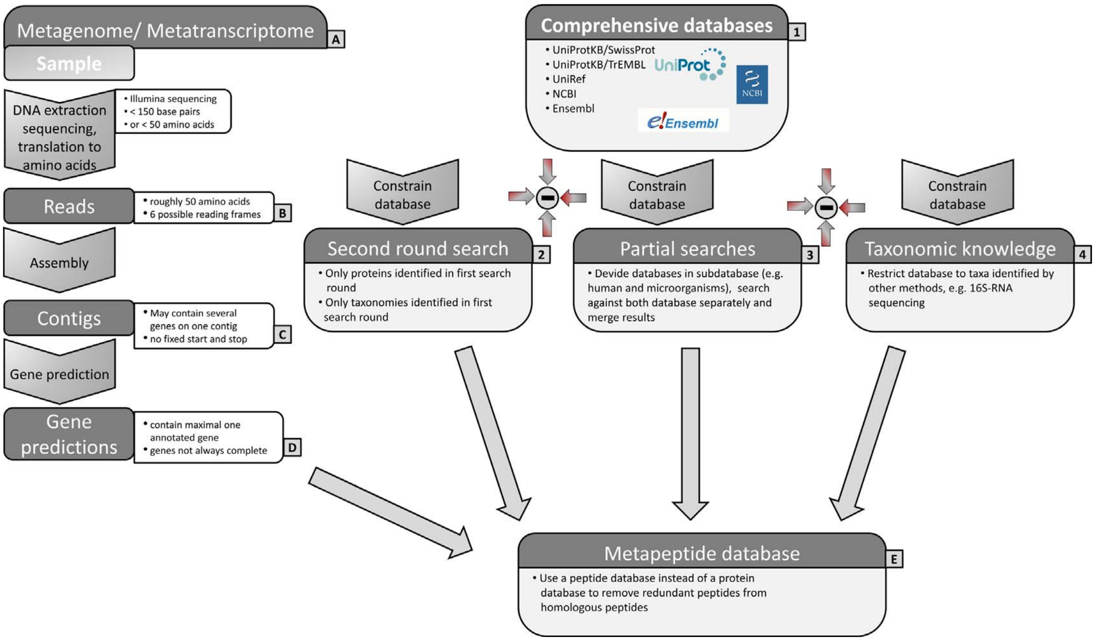
Metagenome/ Metatranscriptome
A
Sample
DNA extraction sequencing, translation to amino acids
• Illumina sequencing
• < 150 base pairs
• or < 50 amino acids
Reads
• roughly 50 amino acids
• 6 possible reading frames
B
Assembly
Contigs
• May contain several genes on one contig
• no fixed start and stop
C
Gene prediction
Gene predictions
• contain maximal one annotated gene
• genes not always complete
D
Comprehensive databases
1
• UniProtKB/SwissProt
• UniProtKB/TrEMBL
• UniRef
• NCBI
• Ensembl
UniProt
NCBI
Ensembl
Constrain database
Second round search
2
• Only proteins identified in first search round
• Only taxonomies identified in first search round
Constrain database
Partial searches
3
• Devide databases in subdatabase (e.g. human and microorganisms), search against both database separately and merge results
Constrain database
Taxonomic knowledge
4
• Restrict database to taxa identified by other methods, e.g. 16S-RNA sequencing
Metapeptide database
E
• Use a peptide database instead of a protein database to remove redundant peptides from homologous peptides

Fig. 2. Database construction for protein identification.

5. Construction of user databases for protein identification: a use case

In order to visualize the impact of user databases a case study was conducted for a metaproteome analysis of three different biogas plant samples (BGP01, BGP02, BGP03). After phenol extraction, SDS-PAGE separation into ten fractions (Heyer et al., 2013) and LC–MS/MS measurement using an Orbitrap Elite (Heyer et al., 2016) different protein databases were tested [Fig. 3]. First the samples were searched against the UniProtKB/SwissProt database. Second several metagenomes from biogas plants were tested (metagenome 1, metagenome 2, metagenome 4, metagenome 5 (Stolze et al., 2016), metagenome 6 (Schlüter et al., 2008)). Of these metagenomes number 1 and 2 were prepared for BGP01 resp. BGP02. A metagenome from a waste water treatment plant (WWTP) (Püttker et al., 2015) was used as a negative control. Furthermore, the impacts of combining databases as well as of combining the results were evaluated.

The smallest numbers of identified metaproteins could be identified by the protein database search against the WWTP metagenome followed by the search against the UniProtKB/SwissProt database. Better results were obtained with the biogas plant metagenomes. Instead of 900 metaproteins for the protein database search against UniProtKB/SwissProt database about 2.000 metaproteins were identified using the biogas plant metagenomes. In some cases metagenomes appeared to be interchangeable, because metagenomes from other biogas plant samples showed equal or even better numbers of identified metaproteins as matching metagenomes, e.g. BGP02 and metagenome 2. This result questions whether the generation of a corresponding metagenome for each sample is always necessary. The combination of different metagenomes additionally increased the number of identified metaproteins to about 4.000 (combination metagenome 1 + 2 + 4 + 5 + 6). However, the number of additional metaprotein identifications decreased for each additional metagenome included in the search. In contrast the combination of metagenome 5 and the poorly matching metagenome from a waste water treatment plant (WWTP) decreased the number of identified metaproteins showing that an increased size of the database led to an increased chance of false positive hits and an increased FDR. The highest number of identified metaproteins was obtained with the separate search against all metagenomes (metagenome 1;2;4–6) and subsequent combination of the results. Focusing

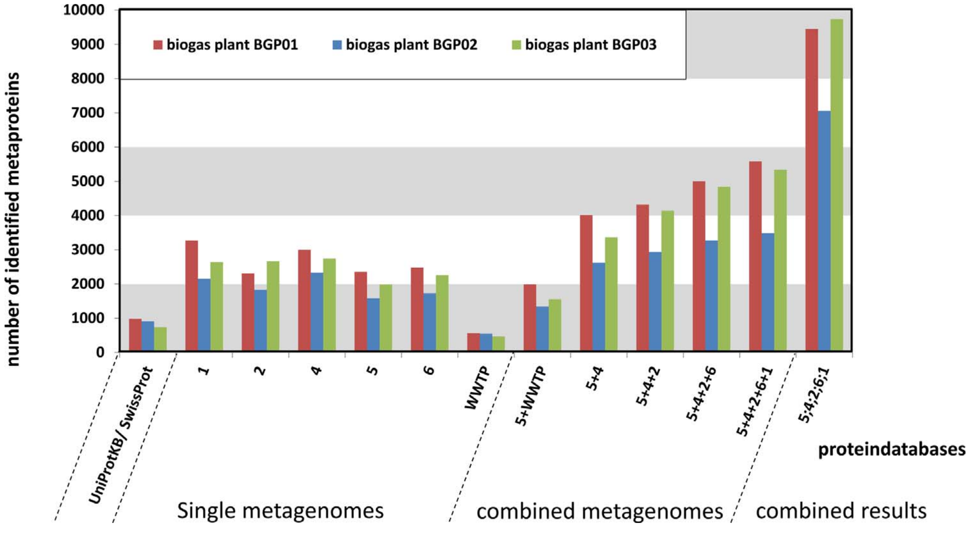

Fig. 3. Impact of different metagenomes and their combination on the number of identified metaproteins.

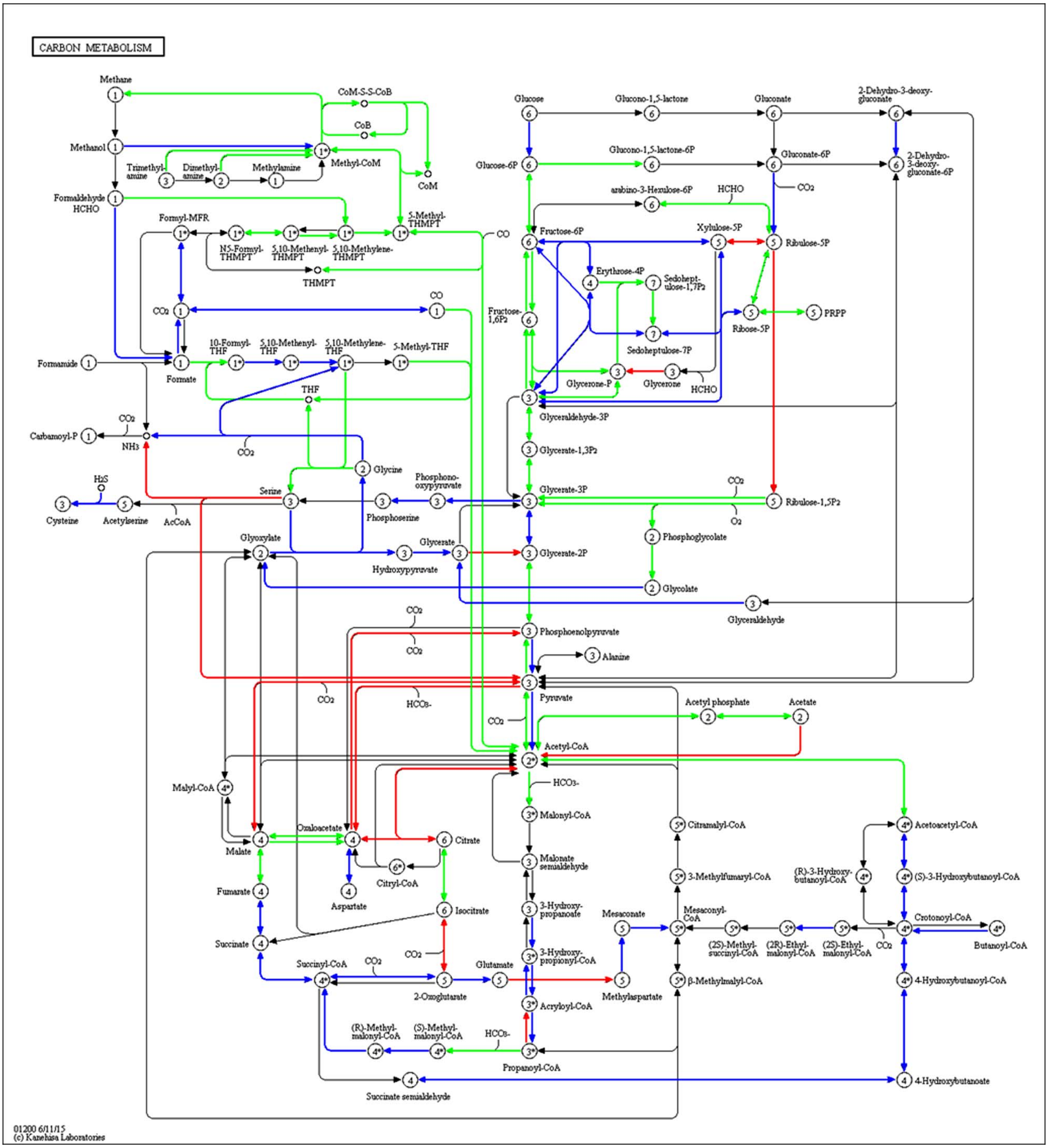

Fig. 4. This figure shows the identified metaproteins of sample BGP01 after protein database search against different databases mapped against the KEGG map 1200 (central carbon metabolism. Green: metaproteins identified by protein database search against UniProtKB/SwissProt; blue: metaproteins identified additionally by protein database search against the combined metagenomes (1 + 2 + 4 + 5 + 6); red: metaproteins identified additionally by protein database search against the single metagenomes (1;2;4;5;6).

on central metabolism and plotting the metaproteins into KEGG map 1200 clearly shows a higher coverage of pathways using the combined single searches (Fig. 4). This strategy avoided the increase of the FDR due to the bigger database, but the statistical correctness of this approach is questionable. However, it circumvents the accumulation of redundant sequence data in a combined database contributing to increased database size and FDR. Therefore, the removal of redundancy using peptide based databases could be a strategy to combine databases without increasing the FDR. Furthermore, the fact that combined metagenomes outcompete single corresponding metagenomes points out that many metagenome sequences do not comprehensively represent the microbial communities.

6. Protein inference problem and the grouping of proteins into “metaproteins”

Redundant identifications arising from homologous proteins share identical peptides and are therefore indistinguishable from each other. This hampers result evaluation and sample comparison within metaproteomic studies.

For pure culture proteomics Nesvizhskii et al., (2003) proposed to use the least number of proteins to explain all peptides. But this neglects

Table 2
Strategies for grouping of redundant homologous proteins to metaproteins.

Rule	Principle	Explanation	Reference
Protein rule	1. UniRef-Cluster	Grouping of proteins when they have 50%, 90% or 100% sequence similarity. Protein clustering provided by UniRef Cluster [Suzek2007].	Lu et al. (2014), Suzek et al. (2007)
	2. KEGG Ontologies	Grouping of proteins when they are similar to functional classified genes within KEGG Ontology [Mai 2005]. KEGG Ontologies are provide by UniProtKB databases [JAPI PAPER].	Gotelli et al. (2012), Kanehisa et al. (2016)
Peptide rule	1. Shared peptide set	Group proteins when they share the same peptides.	Keiblinger et al. (2012), Kolmeder et al. (2012), Schneider et al. (2011)
	2. One shared peptide	Group proteins when they have one identified peptide in common	Kohrs et al. (2014), Lu et al. (2014)
	3. One shared peptide + Levenshtein, distance < 2	Group proteins when they share the same peptides, but not if they have two similar peptides with less than 2 point mutations differences. This tracks the production of one protein by different microorganisms.	Muth et al. (2015a)
Taxonomy rule	1. Phylogenetic affiliation	Extends other rules by a certain phylogenetic affiliation.	Muth et al. (2015a)

the presence of protein isoforms or proteins from unsequenced microorganisms (Hettich et al., 2013) often found in analyses of metaproteomics data. To solve this issue the metaproteomics community started to develop concepts for grouping of redundant protein identifications [Table 2]. The metaprotein concept, introduced by Muth et al. (2015a), provides a good summary on protein grouping. Similar amino acid sequences (protein rules) or shared peptide identifications (peptide rules) constitute suitable criteria for grouping of homologous protein identifications into metaproteins. Conveniently, UniRef Clusters (Lu et al., 2014; Suzek et al., 2007) and KEGG Ontologies (Gotelli et al., 2012; Kanehisa et al., 2016) already classify most proteins on their sequence similarity. An easy retrieval of these classifications is enabled by the UniProtKB database, which is accessible through the UniProtJAPI library (Patient et al., 2008). Alternatively, proteins can be grouped when they share at least one identified peptide (Kohrs et al., 2014; Lu et al., 2014) or an identical peptide set (Keiblinger et al., 2012; Kolmeder et al., 2012; Schneider et al., 2011). It should be noted that for peptide comparison, the isobaric amino acids leucine and isoleucine are not distinguishable from each other.

All these strategies reduce the redundancy of the protein identifications successfully. However, only grouping based on identified peptides considers different conservation levels of the protein sequences. Thus, it enables a better taxonomic classification. Unfortunately, sample comparison using the peptide rule requires the protein grouping across all samples. Furthermore, the grouping may change as soon as additional samples are added. In consequence, grouping according to sequence similarity, such as UniRef clusters, is better suited for sample comparisons (Heyer et al., 2016; Kohrs et al., 2017).

In some instances it is desirable to consider the production of homologous proteins by different species. Homologous proteins often share peptides, which only differ in one or two amino acids. This indicates that these proteins should not be grouped together. To consider this bioinformatically, the Levenshtein distance (Levenshtein, 1966) between peptides of a protein group can be calculated (Muth et al., 2015a). Taxonomic foreknowledge is another option to improve metaprotein grouping. Protein groups can be restricted to certain phylogenetic affiliations, e.g. only proteins from the same genus.

7. Taxonomic and functional result evaluation

Comprehensive metaproteomics studies aim to describe the taxonomies and functions of complete microbial communities. In particular, the functions performed by each taxon should be elucidated.

Protein taxonomy [Table 3] is usually defined according to the NCBI Taxonomy (Federhen, 2012). It comprises the classification for all taxonomic levels into the phylogenetic tree starting from species, genus and family via class, order and phylum to the kingdom and superkingdom levels.

In contrast to pure culture proteomics, a large portion of identified peptides in metaproteomics may belong to several proteins from different species. Thus, the taxonomic value of an identified peptide is estimated using the lowest common ancestor (LCA) of the protein taxonomies where this peptide occurs. Protein taxonomy is then defined as the LCA of the peptide identifications (Huson et al., 2011; Jagtap et al., 2012) or on the basis of unique peptides (Karlsson et al., 2012; Rooijers et al., 2011). Certain taxa have a much larger number of unique peptides, which biases the taxonomic profile towards these taxa. In general, unique peptides are fairly uncommon, as the analyses by UniPept demonstrate (Mesuere et al., 2015). The LCA approach is imprecise as well, because peptide taxonomy is often assigned on the order level and not on the species level. To refine the taxonomy profile Huson et al. (2016) propose to weigh the identified peptides and their LCA taxonomy by the amount of unique peptides. Another approach to improve the precision of the taxonomic profile is to weigh identified peptides by their spectral count and their occurrence in reference proteomes (Penzlin et al., 2014). Still, evaluation and comparison of taxonomic profiles is often challenging due to the high complexity of the data. This has led to several new approaches for data evaluation and visualization. The Krona plot (Ondov et al., 2011) clearly visualizes the taxonomy profile of a sample over all taxonomic levels. Furthermore, calculating community indices such as richness and evenness can give a general overview about the taxonomic profile of different samples (Heyer et al., 2016; Marzorati et al., 2008). In addition, specific interactions between single taxa can be examined by co-occurrence networks (Heyer et al., 2016; Huson et al., 2016; Jenssen et al., 2001).

Several approaches with varying degree of specificity exist to assign functions to proteins [Table 3]. The protein acetyl-coenzyme A synthetase (P27550) is selected as example. It belongs to the acetate catabolism, which is sufficient to classify this proteins function. In other cases however, it is necessary to know that this protein transfers a coenzyme or contributes to chemotaxis. Originally, researchers studied the function of proteins separately through biochemical assays. Later their results were compiled, standardized and stored in databases. Recently, the functions of proteins from new species are derived from sequence similarity to functionally classified proteins. Functional classification of proteins with similar sequences is provided by databases such as KEGG ontology (KO) (Kanehisa et al., 2016), cluster of orthologous groups (COG) (Tatusov et al., 2000) and evolutionary genealogy of genes: non-supervised orthologous (eggNOG) (Huerta-Cepas et al., 2016).

Proteins of the same function possess differences in their amino acid sequence, but the sequences of their functional domains are highly conserved. Accordingly, the PFAM (Finn et al., 2016), the TIGRFAM database (Haft et al., 2013), the SMART database (Letunic et al., 2015) and the InterPro database (Finn et al., 2017) provide a functional classification based on similar functional domains. For example, acetyl-coenzyme A synthetase (P27550) possesses an acetyl-coenzyme A synthetase domain and an AMP-binding enzyme domain.

Table 3
Strategies for taxonomic and functional annotation of proteins.

Issue	Name/principle	Explanation	Reference
Taxonomic classification	1. Lowest common ancestor	Define taxonomy as the lowest common ancestor into the phylogenetic tree.	Huson et al. (2011), Jagtap et al. (2012)
	2. Weighted lowest common ancestor	Adjust the lowest common ancestor by unique identification for the single taxa.	Huson et al. (2016)
	3. Peptide similarity estimation and expression level weighting	Weight taxonomy of identified peptides by their spectra abundance and their occurence in a reference proteome.	Penzlin et al. (2014)
	4. Unique peptides	Define taxonomy and taxonomy profiles only based on unique peptides.	Rooijers et al. (2011), Karlsson et al. (2012)
Functional classification	1. KEGG Orthologies (KO)	Grouping of genes with same function by sequence similarity.	Kanehisa et al. (2016)
	2. Cluster of orthologues genes (COG)	Grouping of genes with same function by sequence similarity.	Tatusov et al. (2000)
	3. Evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG)	Extension off COG by non-supervised orthologous groups constructed from numerous organisms.	Huerta-Cepas et al. (2016)
	4. PFAM	Database of conserved functional units, represented by a set of aligned sequences with their probabilistic representation (hidden Markov model).	Finn et al. (2016)
	5 TIGRFAM	Database of conserved functional units, represented by a set of aligned sequences with their probabilistic representation (hidden Markov model). In contrast to PFAM TIGRFAM emphasize protein function and enables a more precise functional classification.	Haft et al. (2013)
	6. SMART	Functional domain database based on manually curated hidden Markov models.	Letunic et al. (2015)
	7. InterPro	Functional analyses of protein sequences by classifying them into families and predicting the presence of domains and important sites. Signatures are provided by 14 different member databases (among others PFAM, TIGRFAMS, SMART).	Finn et al. (2017)
	8. Enzyme Comission number (EC)	Numerical classification scheme for enzymes, based on the chemical reactions they catalyze	Bairoch, (2000)
	9. UniProt Keywords	Hierachical classification of protein functions.	UniProt, (2015)
	10. Gene ontologies	Hierachical classification of protein functions.	Ashburner et al. (2000)
Pathway mapping	1. MetaCyc	Curated database of experimentally confirmed metabolic pathways.	Caspi et al. (2016)
	2. KEGG pathways	Collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks.	Kanehisa et al. (2016)
	3. Reactome	Pathway database.	Fabregat et al. (2016)
	4. Interactive Pathways Explorer (iPath)	Web-based tool for the visualization, analysis and customization of pathways maps.	Yamada et al. (2011)
	5. CellNetAnalyzer	MATLAB toolbox providing computational methods and algorithms for exploring structural and functional properties of metabolic, signaling, and regulatory networks.	Klamt et al. (2007)
Calculation of sequence similarity	1. BLAST	Calculation of sequence similarities.	Altschul et al. (1990)
	2. DIAMOND	Calculation of sequence similarities. Up to 20,000 faster as BLAST.	Buchfink et al. (2015)
	3. MS-BLAST	Calculation of sequence similarities optimized for peptides.	Shevchenko et al. (2001)

It is important to note that functional annotation of proteins can be divided into categories such as molecular function, biological process or ligand, which are organized hierarchically. This is achieved by gene ontologies (GO) (Ashburner et al., 2000) and UniProtKB keywords (UniProt, 2015). For acetyl-coenzyme A synthetase (P27550) the UniProtKB keyword of the category ligand is ATP-binding protein; which belongs to the group of nucleotide-binding proteins. Enzyme commission numbers (EC) are another functional characterization of proteins (Bairoch; 2000). They use a four digit number code to classify enzymes depending on the catalyzed biochemical reaction. The EC for acetyl-coenzyme A synthetase (P27550) is 6.2.1.1; where 6 classifies it as a ligase; 6.2 as forming carbon sulfur bonds; 6.2.1. as acid-thiol ligase and 6.2.1.1. as acetate Co A ligase.

Conveniently, access to this taxonomic and functional metainformation is already provided by well annotated databases, such as UniProtKB. The entire database is available via the UniProt webpage and can be accessed programmatically via connectors such as the UniProtJAPI (Patient et al., 2008). Metagenomes miss taxonomic and functional annotation. Therefore, metagenome sequences are annotated by BLAST (Altschul et al., 1990) to link them to sequences of annotated proteins. Contigs may contain several genes with different functions, which can lead to false annotations. Moreover, the best BLAST hit is not always the correct one (Timmins-Schiffman et al., 2017) and for searches with short sequences, such as peptides, parameters for the BLAST should be adapted (MS-BLAST (Shevchenko et al., 2001)). Moreover, BLAST requires extensive computational time, which was addressed by development of the time-saving DIAMOND tool (Buchfink et al., 2015).

Another aim of metaproteomics studies is the analysis of certain metabolic pathways. Therefore, identified proteins can be visualized in the different metabolic and interaction pathways, using the pathway repositories MetaCyc (Caspi et al., 2016), KEGG pathways (Kanehisa et al., 2016) and Reactome (Fabregat et al., 2016). For KEGG pathways the web-based Interactive Pathways Explorer (iPath) (Yamada et al., 2011) provides an improved visualization and supports pathway analysis. Mapping of proteins to pathways is provided via the EC and KO numbers. Unfortunately, metabolic networks are incomplete, since many pathways are still unknown or specific for a minority of species. To overcome this limitation researcher started to create their own metabolic pathway maps. To achieve this, biochemical reactions, represented by EC numbers of identified proteins, were connected (Tobalina et al., 2015). A similar approach was chosen by Roume et al. (2015) aiming to identify key functions within a microbial community. Metabolic networks were modelled as a graph, where proteins (KO number) represented nodes and metabolites represented edges. Finally they defined key functions as nodes with high neighborhood connectivity. In future, networks based on metaproteome data could be used to predict metabolic fluxes, using software tools such as the CellNetAnalyzer (Klamt et al., 2007).

8. Quantitative data analysis in metaproteome studies

Protein quantification is crucial for comparative metaproteomics studies. Indeed different approaches for quantitative proteomics exist, e.g. isotopic chemical labelling of peptides (Vaudel et al., 2010). But due to interference of these approaches with contaminating compounds many metaproteomics studies simply rely on the estimation of protein amount by counting identified peptides or spectra and normalizing these results (Ishihama et al., 2005), (Zybailov et al., 2007). Depending on data-dependent selection of precursor ions and successful peptide identification these approaches are inaccurate and possess a small dynamic range [Tabb2009]. The quantification of the peptide peak intensity or area (Griffin et al., 2010) using tools such as Progenesis QI (http://www.nonlinear.com/progenesis/qi-for-proteomics/) or MaxQuant (Tyanova et al., 2016) is preferable. Alternatively, data-independent acquisition of MS/MS data (SWATH, MS^E) combines peptide identification and quantification capturing all possible fragment information of all precursors for subsequent protein quantification from complex data (Bilbao et al., 2015). The most accurate quantification can be achieved by targeting only a single peptide ("single reaction monitoring") or a limited selection of peptides of a certain protein ("single reaction monitoring"). For example, Saito et al. (2015) used this approach to quantify two nitrogen regulatory proteins for cyanobacterial taxa within microbial samples from the Central Pacific Ocean. The addition of isotopically labeled peptide for absolute quantification and the application of the Skyline software (MacLean et al., 2010) further improve this approach.

However, selection of peptides for targeted metaproteomics is more challenging than in pure culture proteomics, because a peptide may belong to multiple proteins from different taxa. Thus, the Unique Peptide Finder of the UniPept webservice (Mesuere et al., 2016) was developed to facilitate the selection of unique peptides for a certain taxa.

9. Strategies for storing and deployment of huge data

Metaproteomics experiments comprise a massive amount of data including MS spectra, identified peptides and proteins as well as taxonomic and functional information. Our latest large-scale metaproteomics study produced about two Terabyte of data comprising roughly 15 million spectra and 23,000 identified metaproteins (data not shown). Consequently, appropriate data storage using a database management system (DBMS) is beneficial. Key challenges for DBMS are high speed for writing and reading data as well as efficient data storage. Since MS acquisition and search algorithms are relatively slow, writing speed has a negligible impact. In contrast, reading speed can be limiting, because researches want to evaluate all data at once. Furthermore, lists of thousands of proteins are unfeasible when inspecting results. Instead, researchers favor meaningful summaries, comparisons and intuitive visualizations. But this requires demanding database queries.

Relational database management systems, which use the "Structured Query Language" (SQL), have been the norm to manage data in the past. In recent years, alternatives to SQL have gained popularity and are aggregated under the term NoSQL ("Not only SQL"). Relational database management systems store data in separate tables, which are connected via unique relations. NoSQL database management systems use other concepts to store data like key-value associations (Berkeley DB (http://www.oracle.com/technetwork/database/database-technologies/berkeleydb/overview/index.html, retrieved: 09-02-2017)), columns (Apache Cassandra (http://cassandra.apache.org/, retrieved: 09-02-2017)), documents (MongoDB (https://www.mongodb.com, retrieved: 09-02-2017)) or graphs (Neo4j, (www.neo4j.com, retrieved: 09-02-2017)).

NoSQL databases where motivated by the disadvantage in SQL databases to store all data in one place. In an analogy SQL databases can be imagined as a large building, which only a limited number of persons at a time can enter. An SQL query would be a person searching the building and collecting the information requested. If too many people search the building at a time, they will hinder each other and slow down the query process. NoSQL databases aim to address this issue of scalability. For instance, in our analogy Apache Cassandra creates a new identical building as soon as too many people try to enter. In consequence, NoSQL databases can handle more and more complex data requests. The disadvantage of NoSQL databases is reduced data consistency and large hard disc requirements due to multiple instances of the databases.

In sum NoSQL databases are highly beneficial for metaproteomics data. In line Chatterjee et al., (2016) used MongoDB for storing sequence information and Muth et al. (2015a) Neo4j for flexible result queries. Additionally, Mesuere et al., (2015) are planning to use Berkeley databases to store the taxonomic value of each tryptic peptide.

Another trend of data storing and deployment which could be useful to increase the speed of data processing in metaproteomics is fast data (Braun et al., 2015). The fast data approach makes it possible to stream single spectra data to the cloud and process the data in real time for storing the results into the database. In other words, it parallelizes the data processing step and the measurement step to reduce experiment time. For example already the software MaxQuant Real-Time (Graumann et al., 2012) picks up this idea and processes the MS data in real time.

10. Future challenges, perspectives and demands

Predictions about the future of metaproteomics software need to anticipate future applications for metaproteomics. Foreseeable trends are an increase in MS resolution and therefore more data that will be acquired. Since metaproteomics is still an emerging field, an increase in the number of research studies about complex microbial communities is expected. A great potential for the application of metaproteomics are process control in technical applications as well routine diagnostics of fecal samples. So far it is known that microbial communities in the human gut system are linked with autoimmune and allergic diseases, obesity, inflammatory bowel disease (IBD), and diabetes (Clemente et al., 2012). Consequently, the number of samples in clinical settings could rise to several thousand per day. Such an increase in sample numbers requires software tools that can handle huge data amounts. For routine diagnostics the total computation time may not exceed a few hours, so that a complete metaproteomics analysis may require less than one day. Another aspect is that software for medical applications has to conform to high quality standards and specific privacy regulations. Moreover, medical staff without a special bioinformatic background should be able to operate such software tools. Although the routine usage of metaproteomics is still in question, the development may proceed quickly. For example, MALDI-MS based identification of microbial isolates became a standard procedure in clinical laboratories.

Strategies to facilitate software usage are to provide it via Docker (e.g. Bioconda https://bioconda.github.io/, retrieved: 09-02-2017) or web services to avoid problems with the installation and configuration of complex software frameworks. For example, developers of the MPA are planning to provide their software platform as web service within the de.NBI project. Most users with a medical or biological background would favor a graphical ready-to-use software tool. In contrast, bioinformaticians prefer modular software packages operated from the command line. The latter strategy enables flexible assembly of workflows and an easy improvement of single modules. The challenge for future development of metaproteomics software is to satisfy both sides.

Because metaproteomics is still a developing field, universal standards still have to be adopted by the community. Implementation of ring trials for metaproteomics data processing could further insights into the comparability of software tools, and enable the introduction of quality standards.

Further improvement requires the validation of protein identifications by the FDR estimation. In contrast to pure culture proteomics the estimated FDR is not always correct since the protein sequences for the investigated samples are often unknown. A solution might be the usage of semi-supervised machine learning algorithms such as the software tools Percolator or Nokoi (Gonnelli et al., 2015).

The use of protein databases could be standardized as well. While some researchers use comprehensive protein databases, others use diverse metagenomes, which differ in the processing state and origin. A solution might be the generation of non-redundant (May et al., 2016), fusion metagenomes for each type of microbial community. Thereby, this fusion metagenome should be assembled as far as possible.

Additionally, the binning of metagenomes may also improve the protein database quality. Proteins of the same function or metabolic pathway are often located adjacent on a contig or operon. Thus, they should feature equal expression patterns.

The key to handle the increased amount of data is the real-time processing of all arising MS data as well as the scalability of the software and the database. This means that the single computational steps operate in parallel and hardware resources can be allocated on demand, e.g. by cloud computing (Mell and Grance, 2010). To guarantee the long term maintenance and support for such systems, it is reasonable to follow the latest trends from the industry instead of developing own solutions. Suitable frameworks, among others, are Apache Spark (http://spark.apache.org/, retrieved: 09-02-2017) for analyzing data distributed in the cloud and OpenStack (https://www.openstack.org/, retrieved: 09-02-2017) to manage the instances running on the cloud.

Another strategy to decrease computation time is the smart deployment of hardware resources. Graphical processing units (GPU) can perform specific calculations in parallel. On the other hand central processing units (CPU) are suited for general tasks, but work serially. Identification of MS/MS spectra is a calculation that can be parallelized. In line, the protein database search algorithm X!Tandem was recently adopted to utilize a GPU (He and Li, 2015).

Beside adaptation of metaproteomics to bigger data volumes and the decrease of computation time, improved bioinformatic strategies are required to increase the number of identified spectra. State-of-the-art metaproteomics studies only achieve identification of 5–30% spectra. An estimated 30% of all spectra belong to solvent and background components (Griss et al., 2016).This means at least another 30% spectra remain unidentified. Better metaproteomics software should contribute to overcome this issue. The generation of more suitable metagenomes for protein identification may increase the amount of identified spectra significantly. Inversely, assembly of metagenomes can be validated using peptides identified in metaproteomics studies (Nesvizhskii, 2014). Spectral libraries represent another strategy to handle unidentified spectra (Lam et al., 2007). They could store and cluster spectra from any sample. Samples can be also compared based on their unidentified spectra. Interesting spectra can be annotated later using protein database search algorithms. Due to the drastic reduction of candidates, manual *de novo* sequencing is also possible (Frank and Pevzner, 2005). Function and taxonomy of *de novo* peptides can be derived by MS-BLAST search (Shevchenko et al., 2001). However, *de novo* sequencing of peptides is hampered by the short length of tryptic peptides which impede MS-BLAST identification. Better *de novo* and MS-BLAST results could be achieved by other proteases such as Lys-C (Jekel et al., 1983) or Arg-C, which result in longer peptides. Due to increased computational power and more precise MS it may become possible to search against a database containing all theoretical peptides for a specific mass (Sadygov, 2015). This would also solve problem with the database size dependency of the FDR estimation.

Finally, metaproteomics software can benefit from the incorporation of data from other multi-omics techniques (Brink et al., 2016; Heintz-Buschart et al., 2016), e.g. metabolome data. For a detailed overview on multi-omics data processing, please refer to Franzosa et al. (2015).

11. Conclusions

Metaproteomics represents a powerful tool for the taxonomic and functional characterization of complex microbial communities from environmental samples. In the future it has the potential to become a valuable tool for routine diagnostics, e.g. analysis of human feces. However, success of metaproteomics studies depends on dedicated software tools. These tools must be capable to handle big data, but also need to be useable by people with no background in bioinformatics. To achieve these goals, web services and software tools capable of parallel computing are reasonable (e.g. cloud computing). This would decrease computational costs and enables small laboratories to perform metaproteomics studies. Moreover, metaproteomics studies will benefit from software supporting the taxonomic and functional interpretation of results. Even if it is obvious, the close cooperation of bioinformaticians and biologists should also be considered during software development.

Competing interests

The authors declare that they have no competing interest.

Funding

This work was supported by the Federal Ministry of Food, Agriculture and Consumer Protection (BMELV) communicated by the Agency for Renewable Resources (FNR), grant no. 22404115 ("Biogas Measurement Program III") and the de.NBI network ("MetaProtServ de-NBI-039").

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Author's contributions

The manuscript was written by Robert Heyer (RH), Dirk Benndorf (DB), Kay Schallert (KS), Beatrice Becher (BB), Udo Reichl (UR) and Günther Saake (GS). All authors read and approved the final manuscript.

Additional files

Not applicable.

Availability of data and material

Not applicable.

Acknowledgement

Not applicable.

References

Abram, F., Enright, A.M., O'Reilly, J., Botting, C.H., Collins, G., O'Flaherty, V., 2011. A metaproteomic approach gives functional insights into anaerobic digestion. J. Appl. Microbiol. 110, 1550–1560.

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Gruning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., Goecks, J., 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 44, W3–W10.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Bairoch, A., 2000. The ENZYME database in 2000. Nucleic Acids Res. 28, 304–305.

Barnouin, K., 2011. Guidelines for experimental design and data analysis of proteomic mass spectrometry-based experiments. Amino Acids 40, 259–260.

Barsnes, H., Vaudel, M., Colaert, N., Helsens, K., Sickmann, A., Berven, F.S., Martens, L., 2011. compomics-utilities: an open-source Java library for computational proteomics. BMC Bioinf. 12, 70.

Becher, D., Bernhardt, J., Fuchs, S., Riedel, K., 2013. Metaproteomics to unravel major microbial players in leaf litter and soil environments: Challenges and perspectives. Proteomics 13 (18-19), 2895–2909. http://dx.doi.org/10.1002/pmic.201300095.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Cheetham, R.K., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X.H., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X.L., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Catenazzi, M.C.E., Chang, S., Cooley, R.N., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fajardo, K.V.F., Furey, W.S., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.

Bilbao, A., Varesio, E., Luban, J., Strambio-De-Castillia, C., Hopfgartner, G., Muller, M., Lisacek, F., 2015. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. Proteomics 15, 964–980.

Braun, L., Etter, T., Gasparis, G., Kaufmann, M., Kossmann, D., Widmer, D., 2015. Analytics in motion: high performance event-processing and real-time analytics in the same database. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data 251–264.

Brink, B.G., Seidel, A., Kleinbolting, N., Nattkemper, T.W., Albaum, S.P., 2016. Omics fusion – a platform for integrative analysis of omics data. J. Integr. Bioinf. 13, 296.

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

Campanaro, S., Treu, L., Kougias, P.G., De Francisci, D., Valle, G., Angelidaki, I., 2016. Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. Biotechnol. Biofuels 9.

Cappadona, S., Baker, P.R., Cutillas, P.R., Heck, A.J., van Breukelen, B., 2012. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. Amino Acids 43, 1087–1108.

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Karp, P.D., 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 44, D471–480.

Chatterjee, S., Stupp, G.S., Park, S.K., Ducom, J.C., Yates 3rd, J.R., Su, A.I., Wolan, D.W., 2016. A comprehensive and scalable database search system for metaproteomics. BMC Genomics 17, 642.

Clemente, J.C., Ursell, L.K., Parfrey, L.W., Knight, R., 2012. The impact of the gut microbiota on human health: an integrative view. Cell 148, 1258–1270.

Colangelo, C.M., Chung, L., Bruce, C., Cheung, K.H., 2013. Review of software tools for design and analysis of large scale MRM proteomic datasets. Methods 61, 287–298.

Coordinators, N.R., 2017. Database resources of the national center for biotechnology information. Nucleic Acids Res. 45, D12–D17.

Crosswell, L.C., Thornton, J.M., 2012. ELIXIR: a distributed infrastructure for European biological data. Trends Biotechnol. 30, 241–242.

Doerr, A., 2015. DIA mass spectrometry. Nat. Methods 12 (35-35).

Elias, J.E., Gygi, S.P., 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods 4, 207–214.

Erickson, A.R., Cantarel, B.L., Lamendella, R., Darzi, Y., Mongodin, E.F., Pan, C., Shah, M., Halfvarson, J., Tysk, C., Henrissat, B., Raes, J., Verberkmoes, N.C., Fraser, C.M., Hettich, R.L., Jansson, J.K., 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One 7, e49138.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2016. The reactome pathway knowledgebase. Nucleic Acids Res. 44, D481–487.

Federhen, S., 2012. The NCBI taxonomy database. Nucleic Acids Res. 40, D136–D143.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–285.

Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztanyi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G.L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D.A., Necci, M., Nuka, G., Orengo, C.A., Park, Y., Pesseat, S., Piovesan, D., Potter, S.C., Rawlings, N.D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P.D., Tosatto, S.C., Wu, C.H., Xenarios, I., Yeh, L.S., Young, S.Y., Mitchell, A.L., 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 45, D190–D199.

Frank, A., Pevzner, P., 2005. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem. 77, 964–973.

Franzosa, E.A., Hsu, T., Sirota-Madi, A., Shafquat, A., Abu-Ali, G., Morgan, X.C., Huttenhower, C., 2015. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nat. Rev. Microbiol. 13, 360–372.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., Bairoch, A., 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31, 3784–3788.

Gonnelli, G., Stock, M., Verwaeren, J., Maddelein, D., De Baets, B., Martens, L., Degroeve, S., 2015. A decoy-free approach to the identification of peptides. J. Proteome Res. 14, 1792–1798.

Gonzalez-Galarza, F.F., Lawless, C., Hubbard, S.J., Fan, J., Bessant, C., Hermjakob, H., Jones, A.R., 2012. A critical appraisal of techniques, software packages, and standards for quantitative proteomic analysis. OMICS 16, 431–442.

Gotelli, N.J., Ellison, A.M., Ballif, B.A., 2012. Environmental proteomics, biodiversity statistics and food-web structure. Trends Ecol. Evol. 27, 436–442.

Graumann, J., Scheltema, R.A., Zhang, Y., Cox, J., Mann, M., 2012. A framework for intelligent data acquisition and real-time database searching for shotgun proteomics. Mol. Cell. Proteom. 11.

Griffin, N.M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., Koziol, J.A., Schnitzer, J.E., 2010. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat. Biotechnol. 28, 83–89.

Griss, J., Jones, A.R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G.G., Salek, R.M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q.W., Del Toro, N., Perez-Riverol, Y., Ghali, F., Bandeira, N., Xenarios, I., Kohlbacher, O., Vizcaino, J.A., Hermjakob, H., 2014. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol. Cell. Proteom. 13, 2765–2775.

Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D.L., Dianes, J.A., Del-Toro, N., Rurik, M., Walzer, M.W., Kohlbacher, O., Hermjakob, H., Wang, R., Vizcaino, J.A., 2016. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nat. Methods 13, 651–656.

Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., Beck, E., 2013. TIGRFAMs and genome properties in 2013. Nucleic Acids Res. 41, D387–395.

Hanreich, A., Heyer, R., Benndorf, D., Rapp, E., Pioch, M., Reichl, U., Klocke, M., 2012. Metaproteome analysis to determine the metabolically active part of a thermophilic microbial community producing biogas from agricultural biomass. Can. J. Microbiol. 58, 917–922.

He, P., Li, K., 2015. MIC-tandem: parallel X! tandem using MIC on tandem mass spectrometry based proteomics data. Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium 717–720.

Heintz-Buschart, A., May, P., Laczny, C.C., Lebrun, L.A., Bellora, C., Krishna, A., Wampach, L., Schneider, J.G., Hogan, A., de Beaufort, C., Wilmes, P., 2016. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat. Microbiol. 2, 16180.

Henry, V.J., Bandrowski, A.E., Pepin, A.S., Gonzalez, B.J., Desfeux, A., 2014. OMICtools: an informative directory for multi-omic data analysis. Database (Oxford) 2014.

Herbst, F.A., Lunsmann, V., Kjeldal, H., Jehmlich, N., Tholey, A., von Bergen, M., Nielsen, J.L., Hettich, R.L., Seifert, J., Nielsen, P.H., 2016. Enhancing metaproteomics—the value of models and defined environmental microbial systems. Proteomics 16, 783–798.

Hettich, R.L., Pan, C.L., Chourey, K., Giannone, R.J., 2013. Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. Anal. Chem. 85, 4203–4214.

Heyer, R., Kohrs, F., Benndorf, D., Rapp, E., Kausmann, R., Heiermann, M., Klocke, M., Reichl, U., 2013. Metaproteome analysis of the microbial communities in agricultural biogas plants. New Biotechnol. 30, 614–622.

Heyer, R., Kohrs, F., Reichl, U., Benndorf, D., 2015. Metaproteomics of complex microbial communities in biogas plants. Microb. Biotechnol. 8, 749–763.

Heyer, R., Benndorf, D., Kohrs, F., De Vrieze, J., Boon, N., Hoffmann, M., Rapp, E., Schluter, A., Sczyrba, A., Reichl, U., 2016. Proteotyping of biogas plant microbiomes separates biogas plants according to process temperature and reactor type. Biotechnol. Biofuel 9, 155.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., Jensen, L.J., von Mering, C., Bork, P., 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44, D286–293.

Huson, D.H., Mitra, S., Ruscheweyh, H.J., Weber, N., Schuster, S.C., 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 21, 1552–1560.

Huson, D.H., Beier, S., Flade, I., Gorska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.J., Tappu, R., 2016. MEGAN Community edition – interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput. Biol. 12, e1004957.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., Mann, M., 2005. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol. Cell. Proteom. 4, 1265–1272.

Jagtap, P., McGowan, T., Bandhakavi, S., Tu, Z.J., Seymour, S., Griffin, T.J., Rudney, J.D., 2012. Deep metaproteomic analysis of human salivary supernatant. Proteomics 12, 992–1001.

Jagtap, P., Goslinga, J., Kooren, J.A., McGowan, T., Wroblewski, M.S., Seymour, S.L., Griffin, T.J., 2013. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics 13, 1352–1357.

Jagtap, P.D., Blakely, A., Murray, K., Stewart, S., Kooren, J., Johnson, J.E., Rhodus, N.L., Rudney, J., Griffin, T.J., 2015. Metaproteomic analysis using the Galaxy framework. Proteomics 15, 3553–3565.

Jekel, P.A., Weijer, W.J., Beintema, J.J., 1983. Use of endoproteinase Lys-C from Lysobacterenzymogenes in protein sequence analysis. Anal. Biochem. 134, 347–354.

Jenssen, T.K., Laegreid, A., Komorowski, J., Hovig, E., 2001. A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. 28, 21–28.

Jones, A.R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S.J., Selley, J.N., Searle, B.C., Shofstahl, J., Seymour, S.L., Julian, R., Binz, P.A., Deutsch, E.W., Hermjakob, H., Reisinger, F., Griss, J., Vizcaino, J.A., Chambers, M., Pizarro, A., Creasy, D., 2012. The mzIdentML data standard for mass spectrometry-based proteomics results. Mol. Cell. Proteom. 11 M111 014381.

Junemann, S., Sedlazeck, F.J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., Mellmann, A., Goesmann, A., von Haeseler, A., Stoye, J., Harmsen, D., 2013. Updating benchtop sequencing performance comparison. Nat. Biotechnol. 31, 294–296.

Junemann, S., Prior, K., Albersmeier, A., Albaum, S., Kalinowski, J., Goesmann, A., Stoye, J., Harmsen, D., 2014. GABenchToB: a genome assembly benchmark tuned on bacteria and benchtop sequencers. PLoS One 9, e107014.

Kall, L., Canterbury, J.D., Weston, J., Noble, W.S., MacCoss, M.J., 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods 4, 923–925.

Kallmeyer, J., Pockalny, R., Adhikari, R.R., Smith, D.C., D'Hondt, S., 2012. Global distribution of microbial abundance and biomass in. Proc. Natl. Acad. Sci. U. S. A. 109, 16213–16216.

Kan, J., Hanson, T.E., Ginter, J.M., Wang, K., Chen, F., 2005. Metaproteomic analysis of Chesapeake Bay microbial communities. Saline Systems 1, 7.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462.

Karlsson, R., Davidson, M., Svensson-Stadler, L., Karlsson, A., Olesen, K., Carlsohn, E., Moore, E.R.B., 2012. Strain-level typing and identification of bacteria using mass spectrometry-based proteomics. J. Proteome Res. 11, 2710–2720.

Keiblinger, K.M., Wilhartitz, I.C., Schneider, T., Roschitzki, B., Schmid, E., Eberl, L., Riedel, K., Zechmeister-Boltenstern, S., 2012. Soil metaproteomics – comparative evaluation of protein extraction protocols. Soil Biol. Biochem. 54, 14–24.

Keller, A., Shteynberg, D., 2011. Software pipeline and data analysis for MS/MS proteomics: the trans-proteomic pipeline. Methods Mol. Biol. 694, 169–189.

Klamt, S., Saez-Rodriguez, J., Gilles, E.D., 2007. Structural and functional analysis of cellular networks with CellNetAnalyzer. BMC Syst. Biol. 1, 2.

Kohrs, F., Heyer, R., Magnussen, A., Benndorf, D., Muth, T., Behne, A., Rapp, E., Kausmann, R., Heiermann, M., Klocke, M., Reichl, U., 2014. Sample prefractionation with liquid isoelectric focusing enables in depth microbial metaproteome analysis of mesophilic and thermophilic biogas plants. Anaerobe 29, 59–67.

Kohrs, F., Heyer, R., Bissinger, T., Kottler, R., Schallert, K., Püttker, S., Behne, A., Rapp, E., Benndorf, D., Reichl, U., 2017. Proteotyping of laboratory-scale biogas plants reveals multiple steady-states in community composition. Anaerobe. https://www.ncbi.nlm.nih.gov/pubmed/28189830.

Kolmeder, C.A., de Been, M., Nikkila, J., Ritamo, I., Matto, J., Valmu, L., Salojarvi, J., Palva, A., Salonen, A., de Vos, W.M., 2012. Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. PLoS One 7, e29913.

Kolmeder, C.A., Salojarvi, J., Ritari, J., de Been, M., Raes, J., Falony, G., Vieira-Silva, S., Kekkonen, R.A., Corthals, G.L., Palva, A., Salonen, A., de Vos, W.M., 2016. Faecal metaproteomic analysis reveals a personalized and stable functional microbiome and limited effects of a probiotic intervention in adults. PLoS One 11, e0153294.

Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., Aebersold, R., 2007. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics 7, 655–667.

Letunic, I., Doerks, T., Bork, P., 2015. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 43, D257–260.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Phys. Dokl. 10, 707–710.

Locey, K.J., Lennon, J.T., 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. 113, 5970–5975.

Lu, F., Bize, A., Guillot, A., Monnet, V., Madigou, C., Chapleur, O., Mazeas, L., He, P., Bouchez, T., 2014. Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity. ISME J. 8, 88–102.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., MacCoss, M.J., 2010. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics 26, 966–968.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Ropp, A., Neumann, S., Pizarro, A.D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.A., Deutsch, E.W., 2011. MzML-a community standard for mass spectrometry data. Mol. Cell. Proteom. 10.

Marzorati, M., Wittebolle, L., Boon, N., Daffonchio, D., Verstraete, W., 2008. How to get more out of molecular fingerprints: practical tools for microbial ecology. Environ. Microbiol. 10, 1571–1581.

May, D.H., Timmins-Schiffman, E., Mikan, M.P., Harvey, H.R., Borenstein, E., Nunn, B.L., Noble, W.S., 2016. An alignment-Free metapeptide strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. J. Proteome Res. 15, 2697–2705.

Mell, P., Grance, T., 2010. The NIST definition of cloud computing. CommunAcm 53 (50-50).

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

Mesuere, B., Van der Jeugt, F., Devreese, B., Vandamme, P., Dawyndt, P., 2016. The unique peptidome: taxon-specific tryptic peptides as biomarkers for targeted metaproteomics. Proteomics 16, 2313–2318.

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

Muth, T., Behne, A., Heyer, R., Kohrs, F., Benndorf, D., Hoffmann, M., Lehteva, M., Reichl, U., Martens, L., Rapp, E., 2015a. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. J. Proteome Res. 14, 1557–1565.

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

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

Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R., 2003. A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646–4658.

Nesvizhskii, A.I., 2014. Proteogenomics: concepts, applications and computational strategies. Nat. Methods 11, 1114–1125.

Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinf. 12, 385.

Püttker, S., Kohrs, F., Benndorf, D., Heyer, R., Rapp, E., Reichl, U., 2015. Metaproteomics of activated sludge from a wastewater treatment plant – a pilot study. Proteomics 15, 3596–3601.

Patient, S., Wieser, D., Kleen, M., Kretschmann, E., Jesus Martin, M., Apweiler, R., 2008. UniProtJAPI: a remote API for accessing UniProt data. Bioinformatics 24, 1321–1322.

Penzlin, A., Lindner, M.S., Doellinger, J., Dabrowski, P.W., Nitsche, A., Renard, B.Y., 2014. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. Bioinformatics 30, i149–156.

Ram, R.J., Verberkmoes, N.C., Thelen, M.P., Tyson, G.W., Baker, B.J., Blake, R.C., 2nd Shah, M., Hettich, R.L., Banfield, J.F., 2005. Community proteomics of a natural microbial biofilm. Science 308, 1915–1920.

Rodriguez-Valera, F., 2004. Environmental genomics, the big picture? FEMS Microbiol. Lett. 231, 153–158.

Rooijers, K., Kolmeder, C., Juste, C., Dore, J., de Been, M., Boeren, S., Galan, P., Beauvallet, C., de Vos, W.M., Schaap, P.J., 2011. An iterative workflow for mining the human intestinal metaproteome. BMC Genomics 12, 6.

Roume, H., Heintz-Buschart, A., Muller, E.E.L., May, P., Satagopam, V.P., Laczny, C.C., Narayanasamy, S., Lebrun, L.A., Hoopmann, M.R., Schupp, J.M., et al., 2015. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. NPJ Biofilms Microbiomes 1.

Sachsenberg, T., Herbst, F.A., Taubert, M., Kermer, R., Jehmlich, N., von Bergen, M., Seifert, J., Kohlbacher, O., 2015. MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. J. Proteome Res. 14, 619–627.

Sadygov, R.G., 2015. Using SEQUEST with theoretically complete sequence databases. J. Am. Soc. Mass Spectrom. 26, 1858–1864.

Saito, M.A., Dorsk, A., Post, A.F., McIlvin, M.R., Rappe, M.S., DiTullio, G.R., Moran, D.M., 2015. Needles in the blue sea: sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. Proteomics 15, 3521–3531.

Schlüter, A., Bekel, T., Diaz, N.N., Dondrup, M., Eichenlaub, R., Gartemann, K.H., Krahn, I., Krause, L., Kromeke, H., Kruse, O., Mussgnug, J.H., Neuweger, H., Niehaus, K., Pühler, A., Runte, K.J., Szczepanowski, R., Tauch, A., Tilker, A., Viehover, P., Goesmann, A., 2008. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. J. Biotechnol. 136, 77–90.

Schneider, T., Schmid, E., de Castro Jr., J.V., Cardinale, M., Eberl, L., Grube, M., Berg, G., Riedel, K., 2011. Structure and function of the symbiosis partners of the lung lichen (Lobariapulmonaria L. Hoffm.) analyzed by metaproteomics. Proteomics 11, 2752–2756.

Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., Standing, K.G., 2001. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal. Chem. 73, 1917–1926.

Stolze, Y., Bremges, A., Rumming, M., Henke, C., Maus, I., Puhler, A., Sczyrba, A., Schluter, A., 2016. Identification and genome reconstruction of abundant distinct taxa in microbiomes from one thermophilic and three mesophilic production-scale biogas plants. Biotechnol. Biofuels 9.

Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., Kohlbacher, O., 2008. OpenMS – an open-source software framework for mass spectrometry. BMC Bioinf. 9, 163.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., Wu, C.H., 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23, 1282–1288.

Tanca, A., Palomba, A., Deligios, M., Cubeddu, T., Fraumene, C., Biosa, G., Pagnozzi, D., Addis, M.F., Uzzau, S., 2013. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. PLoS One 8, e82981.

Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., Muth, T., Rapp, E., Martens, L., Addis, M.F., Uzzau, S., 2016. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome 4.

Tang, H., Li, S., Ye, Y., 2016. A graph-Centric approach for metagenome-guided peptide and protein identification in metaproteomics. PLoS Comput. Biol. 12, e1005224.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33–36.

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

Tobalina, L., Bargiela, R., Pey, J., Herbst, F.A., Lores, I., Rojo, D., Barbas, C., Pelaez, A.I., Sanchez, J., von Bergen, M., Seifert, J., Ferrer, M., Planes, F.J., 2015. Context-specific metabolic network reconstruction of a naphthalene-degrading bacterial community guided by metaproteomic data. Bioinformatics 31, 1771–1779.

Tyanova, S., Temu, T., Cox, J., 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat. Protoc. 11, 2301–2319.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F., 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37–43.

UniProt, C., 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–212.

Vaudel, M., Sickmann, A., Martens, L., 2010. Peptide and protein quantification: a map of the minefield. Proteomics 10, 650–670.

Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A., Martens, L., 2011. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 11, 996–999.

Vizcaino, J.A., Csordas, A., Del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.W., Wang, R., Hermjakob, H., 2016. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 44, 11033.

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

Wilmes, P., Andersson, A.F., Lefsrud, M.G., Wexler, M., Shah, M., Zhang, B., Hettich, R.L., Bond, P.L., VerBerkmoes, N.C., Banfield, J.F., 2008. Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. ISME J. 2, 853–864.

Wilmes, P., Heintz-Buschart, A., Bond, P.L., 2015. A decade of metaproteomics: where we stand and what the future holds. Proteomics 15, 3409–3417.

Wohlbrand, L., Trautwein, K., Rabus, R., 2013. Proteomic tools for environmental microbiology–a roadmap from sample preparation to protein identification and quantification. Proteomics 13, 2700–2730.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., Bork, P., 2011. iPath2. 0: interactive pathway explorer. Nucleic Acids Res. 39, W412–W415.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Giron, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S.J., Cunningham, F., Aken, B.L., Zerbino, D.R., Flicek, P., 2016. Ensembl 2016. Nucleic Acids Res. 44, D710–716.

Zhang, X., Ning, Z., Mayne, J., Moore, J.I., Li, J., Butcher, J., Deeke, S.A., Chen, R., Chiang, C.K., Wen, M., Mack, D., Stintzi, A., Figeys, D., 2016. MetaPro-IQ: a universal metaproteomic approach to studying human and mouse gut microbiota. Microbiom 4, 31.

Zybailov, B.L., Florens, L., Washburn, M.P., 2007. Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. Mol. Biosyst. 3, 354–360.