CHAPTER 17

Comparative Metaproteomics to Study Environmental Changes

Sabine Matallana-Surget*, Pratik D. Jagtap†, Timothy J. Griffin†, Mélanie Beraud‡, Ruddy Wattiez‡

**Division of Biological and Environmental Sciences, Faculty of Natural Sciences, University of Stirling, Stirling, Scotland, †Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN, United States ‡Department of Proteomics and Microbiology, Research Institute for Biosciences, Université de Mons, Mons, Belgium*

17.1 Introduction

17.1.1 Towards a Better Understanding of the Functioning of an Ecosystem

To understand the functional response of an organism to an environmental disruption, stress, or pollution, many studies have merely characterized one strain of interest, and studied its physiology and proteome in laboratory-controlled conditions. However, the specific case of the response of one bacterial strain can never predict that which would occur in the natural environment. Indeed, microorganisms function as a community within their natural environment, thus allowing them to benefit from syntrophism and from cell-to-cell communication involving genetic exchanges. Such interactions can either be beneficial or detrimental and influence the structure and functioning of the community, as well as the environment itself [1–3]. Understanding microbial population dynamics in a consortium (natural or artificial) will benefit the study of complex ecosystems in response to environmental changes.

Metagenomics is aimed at studying the genomes of all organisms sampled in the environment and has contributed to important knowledge regarding the genetic diversity of uncultured microorganisms. This tool has allowed a big step to be taken in moving from the so-called structural ecology, towards a "functional" ecology. However, a simple understanding of the genetic potential of a microbial community does not inform us about the part of the genetic material actually expressed by the organisms. In fact, the same genome can result in a number of different proteomes. Metatranscriptomics and metaproteomics aim to focus on the genes that are expressed and may reveal the mechanisms by which bacteria function in their ecosystems. Although metatranscriptomics is a powerful tool to determine gene expression, it relies on mRNA stability, which is extremely low in prokaryotes in comparison to protein

Metagenomics. https://doi.org/10.1016/B978-0-08-102268-9.00017-3

stability. Furthermore, transcriptome analysis does not necessarily represent the genes that are ultimately translated into proteins [4], emphasizing the need to study regulation at the translational level. In this way, proteome-based analyses can be expected to provide a better view of an ecosystem's functioning.

Metaproteomics enables an effective functional study of microbial communities [5,6], allowing us, in association with metagenomics, to further our understanding of the relationships between microbial diversity, the functioning of an ecosystem and the networks of interaction, without any preconceived ideas about the functions that could be affected. The so-called shotgun approach provides the most extensive coverage of the dominant metabolic processes within a microbial community, and can highlight changes in the expression profiles of some key biogeochemical processes. Since 2005, when one of the first environmental metaproteomics studies was carried out on isolated microbial communities in the Chesapeake Bay [7], metaproteomics has undergone considerable development and now involves a series of challenging steps in its workflow that we will present and discuss in this chapter.

17.1.2 Environmental Metaproteomics: Past and Future Trends

Metaproteomics seeks to identify the protein profile of complex microbial communities within an ecosystem at a time point. Environmental metaproteomics studies, using a descriptive or comparative analysis, have focused on four major ecosystems (Table 17.1):

- Aquatic—Seawater
- Aquatic—Freshwater
- Terrestrial—Soil
- Terrestrial—Sediment

As illustrated in Table 17.1, aquatic metaproteomics has attracted considerable attention over the past 12 years and this specific ecosystem will be further introduced in Section 17.1.3.

Some metaproteomics studies have used proteomics as a descriptive tool, providing a comprehensive protein catalog of complex environments (Fig. 17.1, Table 17.1). Such a descriptive approach allows elucidating the metabolic activities employed by the microbial community isolated from a specific ecosystem at the moment of sampling. These studies provide an excellent starting point to evaluate the functioning potential of a complex community, to develop and optimize advanced quantitative proteomic methods for more complex experimental designs.

Metaproteomics has developed rapidly in the past 12 years and nowadays it is almost exclusively used in comparative studies (Table 17.1; Fig. 17.1).

The increased number of available metagenomes and the development of high accuracy mass spectrometers have increased tremendously the complexity of the proteomic datasets, ranging

Table 17.1: Environmental metaproteomics: descriptive vs.comparative studies

Reference	Ecosystem	Aim of the Study	Proteomics Approach	Database
Descriptive Analysis				
[8]	Biofilm	Proteomic analysis of AMD biofilms	2D LC-MS/MS	Specific Metagenome
[9]	Aquatic	Proteomic analysis of an oligotrophic environment - Sargasso Sea	2D LC-MS/MS	In-house database from selected genomic and metagenomic data (GOS)
[10]	Aquatic	Proteomic analysis of cold and nutrient limited waters (Ace lake)	SDS PAGE (1D) LC-MS/MS	Specific Metagenome + Public databases
[11]	Sediment	Proteomic analysis of arsenic-rich sediments	SDS PAGE (1D) LC-MS/MS	Specific Metagenome
[12]	Aquatic	Proteomic analysis of coastal upwelling waters	2D-LC-MS/MS	In-house database from selected genomic and metagenomic data (GOS)
[13]	Soil	Proteomic analysis of rice roots	SDS PAGE (1D) LC-MS/MS	Specific Metagenome
[14]	Soil	Proteomic analysis of semiarid soil	SDS PAGE (1D) LC-MS/MS	Public databases
[15]	Aquatic	Proteomic analysis of the South China Sea	SDS PAGE (1D) LC-MS/MS	GOS, MIMAS
[16]	Aquatic	Targeted proteomic analysis of Pacific Ocean	MRM LC-MS	In silico tryptic digestion of selected genomes
[17]	Soil	Proteomic analysis of the rhizosphere from serpentine soil	2D-LC-MS/MS	In-house database selected on 16S rRNA metagenomic analysis
Comparative Analysis				
[7]	Aquatic	Comparison of two stations (upper and lower Bay)	2D gels LC-MS/MS MALDI-TOF MS	General/de *novo* sequencing
[18]	Aquatic	Comparison of metaproteomes along the nutrient gradient	LC-MS/MS	In-house database from marine metagenomes (from GOS)
[19]	AMD biofilm	Comparison of biofilms in various environments	2D LC-MS/MS	Specific Metagenome
[20]	Aquatic	Comparison of different depths of a meromictic lake	SDS PAGE (1D) LC-MS/MS	In-house metagenomic state specific
[21]	Aquatic	Comparison of metaproteomes in response to an algal bloom	2D LC-MS/MS	MIMAS
[22]	Aquatic	Seasonal comparison of bacterioplankton	SDS PAGE (1D) LC-MS/MS	In-house database from marine metagenomes

(Continued)

Table 17.1 Environmental metaproteomics: descriptive vs.comparative studies—cont'd

Reference	Ecosystem	Aim of the Study	Proteomics Approach	Database
[23]	Aquatic	Seasonal comparison of bacterioplankton	SDS PAGE (1D) LC-MS/MS	In-house database from marine metagenomes (from GOS)
[24]	Aquatic	Comparison of metaproteomes along the redox gradient within the oxygen minimum zone	2D-LC-MS/MS	Specific Metagenome
[25]	Biofilm	Comparison of microbial fouling from 2 destroyers	SDS PAGE (1D) LC-MS/MS	Specific Metagenome + Public databases
[26]	Soil	Comparison of soil communities facing deforestation	LC-MS	Prophane
[27]	Aquatic	Comparison of different depths of St Lawrence stratified estuary	SDS PAGE (1D) LC-MS/MS	In-house database from marine metagenomes and metatranscriptomes
[28]	Sediment	Comparison of different polluted sites impacted by industrial activity	SDS PAGE (1D) LC-MS/MS	Specific Metagenome + Public databases

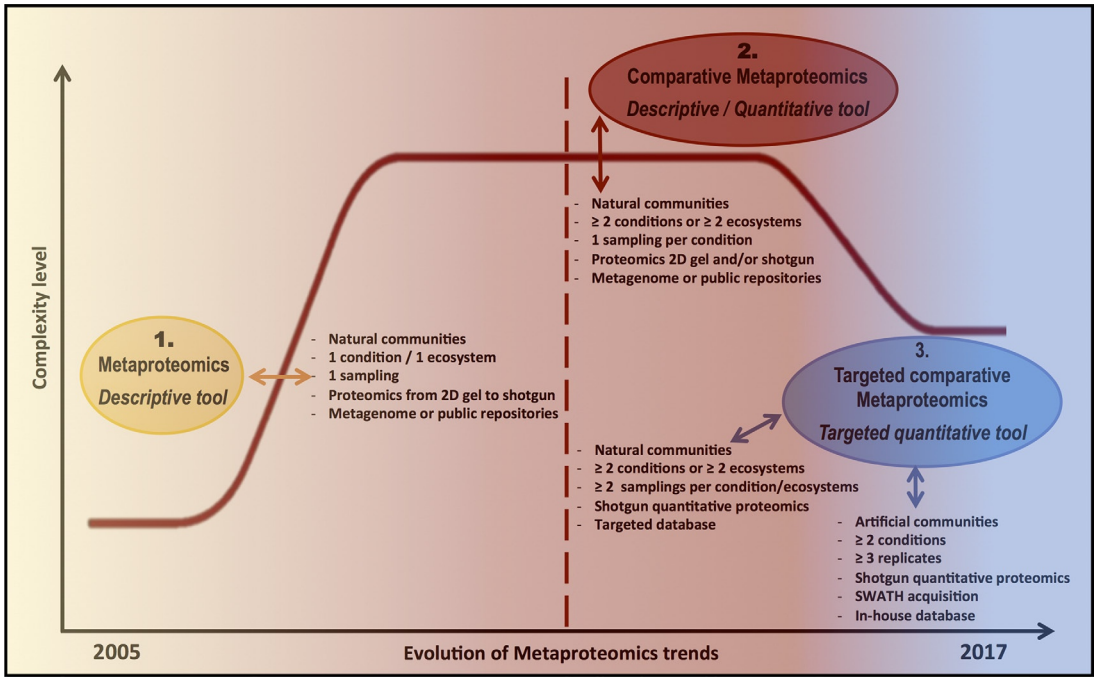


Fig. 17.1
Evolution of the metaproteomics trends and their complexity level over the past 12 years.

from a few identified proteins in the earliest studies [7] to more than 5000 proteins nowadays [21,23,24]. Paradoxically, the success of such big datasets might also suffer from their complexity when it comes to biological interpretation, especially in comparative metaproteomics. Simplifying the complexity level of the proteomic dataset should also be considered, as it can help to understand the functioning of complex communities exposed to environmental change.

A new trend in metaproteomics consists of using quantitative proteomics on simplified microbial communities. Quantitative metaproteomics aims at accurately determining the relative abundances of thousands of proteins in a microbial community. This approach should be used to understand comprehensively the regulation of key pathways of representative bacteria in microbial communities and uncover significant changes in protein expression between communities at different developmental stages, environment types or in response to different perturbations. Interestingly, very few studies have used the quantitative proteomics (label-free or label-based) approach in their comparative studies [29–31].

We would like to emphasize here the importance of identifying a clear biological question in order to set up accordingly a suitable experimental design. The future new trend of the "targeted comparative metaproteomics" approach will aim at:

- Working on a *complex natural community* but targeting a limited number of microorganisms/proteins; in other words, fishing for data.

- Working on *simplified artificial communities*, and assessing the impact of environmental changes by quantitative proteomics using several replicates.

Both strategies will be discussed in this chapter, presenting common metaproteomics workflow, challenging steps, advantages, and limitations of both strategies.

17.1.3 Aquatic Metaproteomics

Marine bacteria are the most abundant organisms on Earth; indeed, the world's oceans contain as many as 1.18×10^{29} prokaryotes and play a key role in climate regulation [32]. Despite the vital role of microorganisms (photosynthetic and heterotrophic bacteria) in marine ecosystems, crucial in the major biogeochemical balances of the planet, their functioning is still not well-defined. This arises from the fact that the vast majority of marine bacteria are nonculturable using standard culturing techniques (99%), which introduces a major bias in our understanding of the functioning of ecological communities and ecosystems. The functions and interactions within microbial communities thus remain unclear, since physiological, biochemical and genetic approaches are impossible to carry out on the majority of bacteria. Environmental metagenomics, which does not involve culturing bacteria, has enabled marked conceptual advances in informing not only microbial diversity, but also the genetic potential of marine bacteria. A first study using a metaproteomics investigation was performed by Kan and colleagues in the Chesapeake Bay in 2005 [7]. Since then, metaproteomics studies on aquatic ecosystems (marine ecosystems and freshwaters) have become numerous and are listed in Fig. 17.2 (for a review on marine metaproteomics, see Ref. [33]).

Most metaproteomics studies on aquatic ecosystems have used a comparative approach to analyze spatial and temporal changes [7,18,21–23] (Table 17.1, Fig. 17.2).Alternatively, natural communities have also been isolated and studied in mesocosms or microcosms, allowing quantitative metaproteomics approaches (Fig. 17.1). This intermediate approach presents several advantages. Mesocosms and microcosms are easy to manipulate in controlled conditions, facilitate replicates, and most importantly allow in vivo labeling, such as the isotope probing (protein-SIP) method, for quantitative proteomics applications[29]. Three studies have focused on natural communities in artificial ecosystems to study the response to cadmium in wastewater [30], species-specific assimilation of amino acids in microcosms [29], and the comparison of oligotrophic vs.copiotrophic freshwater mesocosms [31].

Until now, only one published study performed comparative metaproteomics on aquatic artificial communities, comparing the response of two thermophiles cultivated in isolation or in coculture [34]. The authors observed different levels of antiviral resistance between single-species and pairwise cultures [34].Beraud and coworkers studied the impact of metals on the community structure and functioning of an artificial community composed of ninemarine bacterial species. Comparative metaproteomics of the community between stressed vs. nonstressed cocultures revealed that the community was upregulating key oxidative stress-related proteins, thus allowing it to cope with toxic heavy metal (Beraud et al., in prep). We present in Box 17.1 our metaproteomics study focusing on the day/night cycle of natural microbial communities.

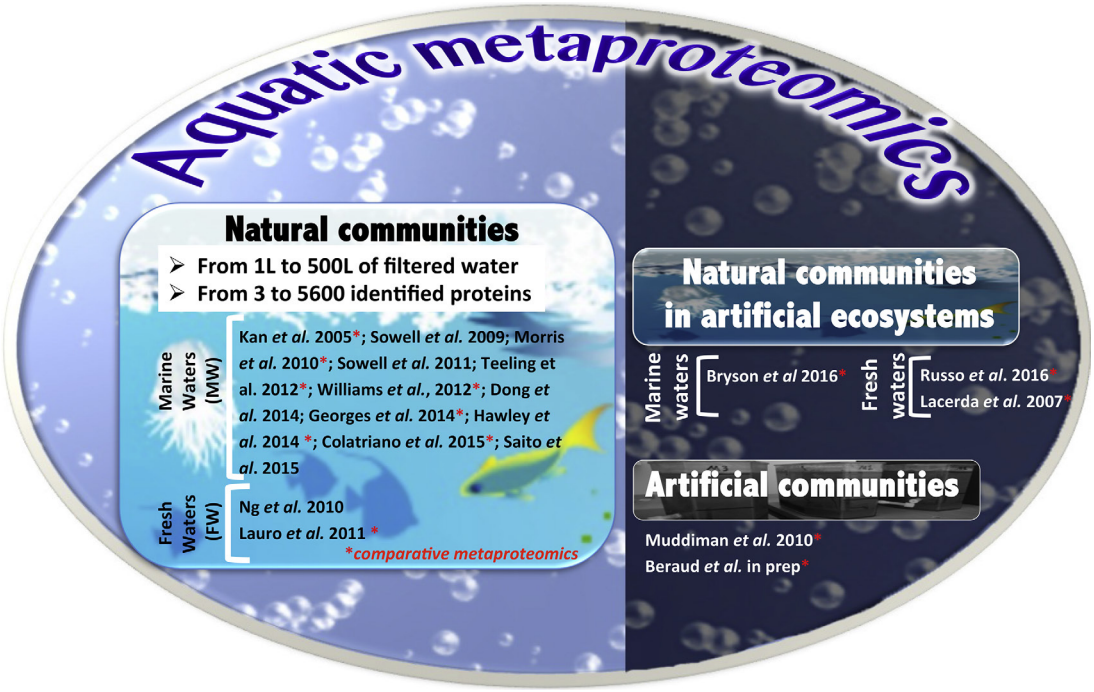


Fig. 17.2
Aquatic metaproteomics studies.

Box 17.1 Example of Targeted Comparative Metaproteomics on Natural Communities

We would like to emphasize here the importance of clearly stating the biological questions and the objectives of a study before undertaking any further steps in the metaproteomics workflow. This section presents an example of a targeted quantitative metaproteomics approach (Matallana-Surget et al., in prep).

The day/night cycle experienced by virtually all life on Earth is of enormous biological importance in that it imposes temporal structure on ecosystem productivity. Although the topic of light/dark cycles has been extensively studied, there are only certain phenomena known to be significantly impacted by the light/dark cycle, such as nycthemeral migrations of zooplankton, daytime photosynthesis in phytoplankton, and nitrogen fixation by nitrogen-fixing bacteria during the night (Fig. 17.3). In the world's oceans about half of the photosynthesis and the bulk of life-sustaining nutrient cycling are carried out by the smallest organisms: the picoplankton. This diverse community includes cyanobacteria known to have highly regulated circadian rhythms and also heterotrophic microbes not suspected to display diel rhythmicity, but which are nonetheless dependent on the primary producers. To what extent picoplankton communities

(Continued)

Box 17.1 Example of Targeted Comparative Metaproteomics on Natural Communities—cont'd

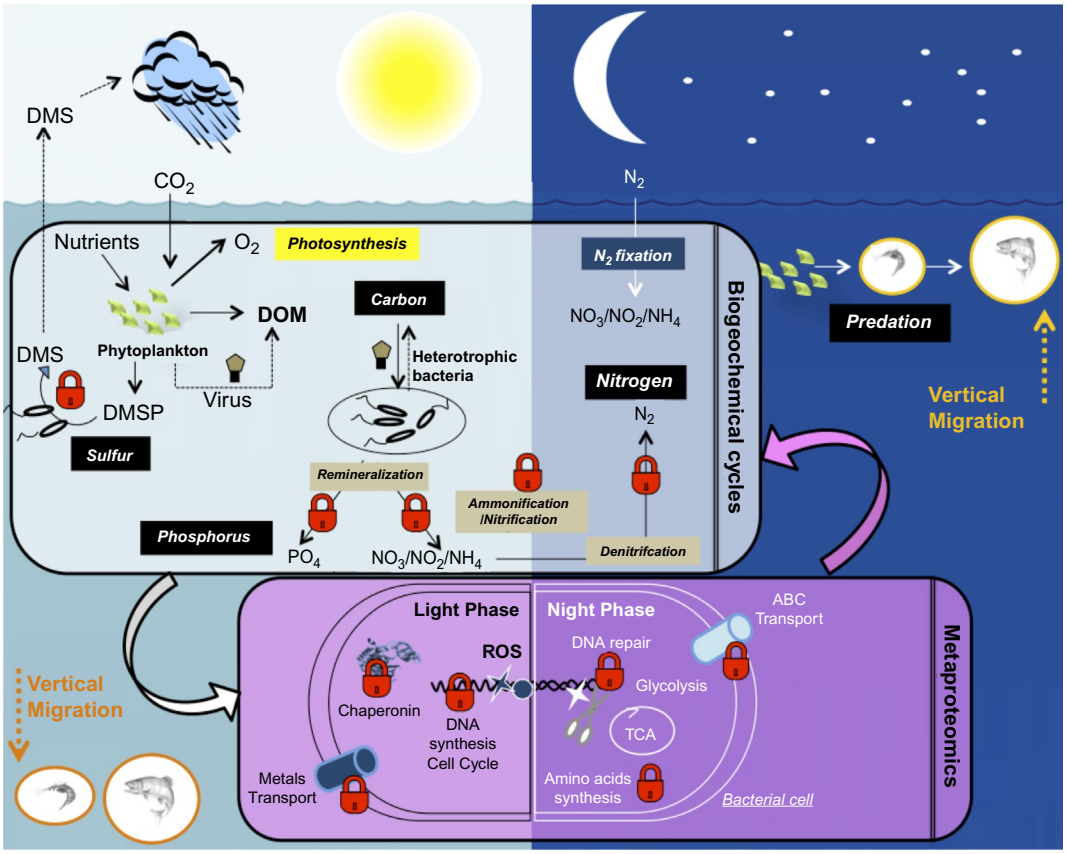


Fig. 17.3
Day/night rhythms at the ocean surface. *Red* padlocks indicate biogeochemical cycles and cellular metabolic pathways that are not yet known to display diel rhythmicity (Matallana-Surget et al., in prep).

are collectively entrained by day/night cycles, and how this influences their population structure, regulates their physiologies, and impinges on species interactions, are questions of immediate urgency.

We conducted a comprehensive investigation of the impact of day/night cycle on microbial communities, using comparative metaproteomics, enabling protein expression in photosynthetic and heterotrophic bacteria to be quantified during the day and the night, for three consecutive days. This study answered for the first time the ecological question: "Who is doing what, and when?" Resulting proteomics mass spectra were searched against an in-house database comprising proteomes of the most abundant microorganisms representative from the studied ecosystem. This is the first quantitative metaproteomics, performing sampling in triplicate, ensuring the reproducibility of the comparative study (Matallana-Surget et al., in prep).

17.2 Metaproteomics on Complex Natural Communities

17.2.1 Metaproteomics Workflow

Metaproteomics uses standard proteomics workflow, extensively reviewed [35–38], but also presents challenging steps that are specifically related to the complex nature of environmental samples. The different steps of the metaproteomics workflow, using aquatic samples, are summarized in Fig. 17.4 and can be grouped into four main parts as follows:

- Water sampling
- Sample preparation

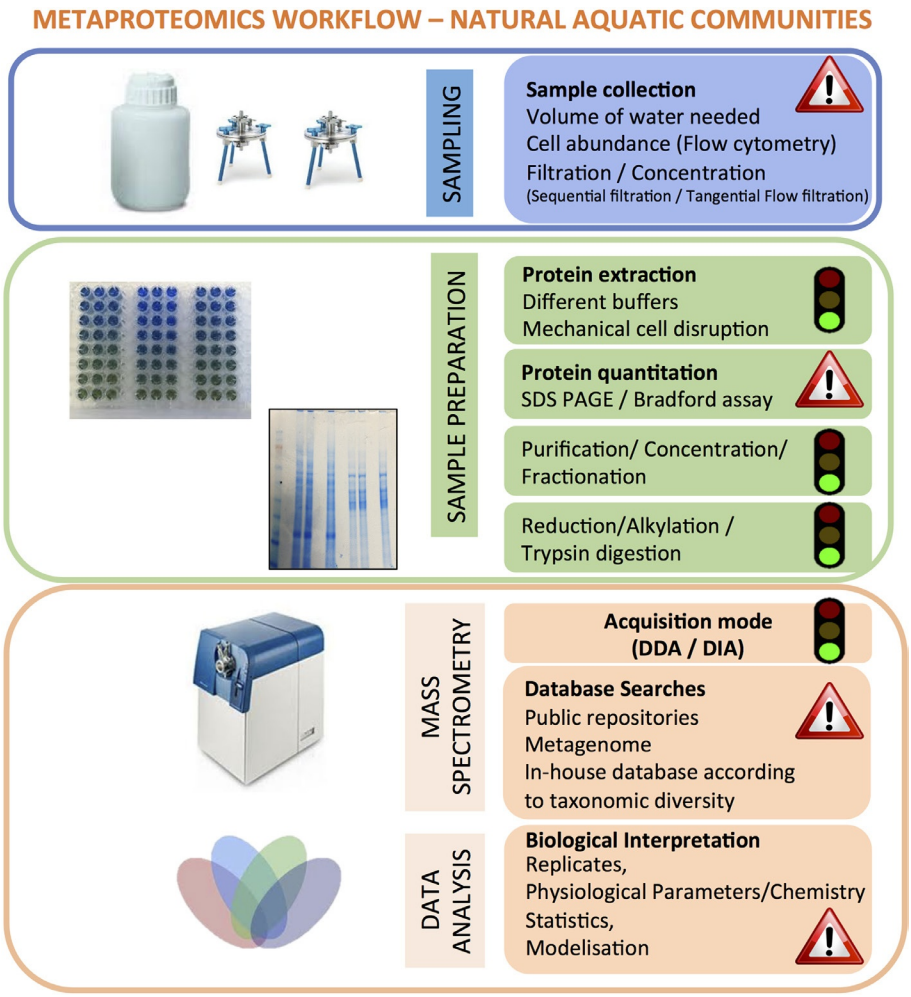


Fig. 17.4
Different steps in the metaproteomics workflow using natural microbial communities isolated from aquatic ecosystems.

- Mass spectrometry
- Data analysis

We will present and discuss in this section the steps that are specific to each ecosystem (represented by green lights in Fig. 17.4) and the ones that should be standardized between all metaproteomics studies, to allow future comparison (represented by caution signs in Fig. 17.4).

17.2.1.1 Sampling

Although aquatic metaproteomics is a growing field for both marine and freshwater ecosystems, until now no standardization has been done in regards to the number of cells needed for downstream proteomics applications. Over the past 12 years, metaproteomics studies have used a different volume of water, from 1 to 500 L, without reporting the number of cells that were extracted. Only two studies provided the initial cell count before and/or after sampling, as well as the total amount of protein extracted per condition ([7,27]; Table 17.2). As illustrated in Table 17.2, no correlation between the volume of filtered/

Table 17.2: Correlation between the volume of water sampled and number of identified proteins for the aquatic metaproteomics studies

Reference	Volume of Water (L)	Number of Cells (T_0)	Protein Extracted	Proteins Identified
Descriptive Analysis				
[9]	230	NC	NC	1042
[10]	1–10	NC	1 mg/filter	504
[12]	100	NC	4.7 mg	481
Comparative Analysis				
[7]	15	T_0: 2.5×10^{06} cell/mL After concentration: 2.5×10^{08} cell/mL (Flow Cytometry)	140–192 µg	250 spots, 41 analyzed by MS, 3 proteins well-identified
[18]	100–200	1×10^{08} cell/mL (FISH, DAPI)	NC	2273
[20]	5	[9.00×10^{05}–1.60×10^{07}] (SYBR-Microscopy)	NC	NC
[21]	500	[4.50×10^{05}–3.0×10^{06}] (CARD FISH)	NC	3128–7278
[22]	200	NC	NC	1061
[23]	10	NC	NC	5627
[24]	20	NC	NC	5019
[27]	10	[3.6×10^{05}–2.0×10^{05}] (Flow Cytometry)	NC	2282–2522

concentrated water and the number of identified proteins can be established, demonstrating the urgent need for providing accurate cell count information by flow cytometry. We certainly acknowledge the fact that the discrepancy between the amount of proteins obtained from different volumes of water samples is not only dependent on the initial cell concentration but also relies on the efficiency of the protein extraction and the accuracy of the mass spectrometer. Nevertheless, we strongly believe that this parameter should be systematically provided in future experiments (Fig. 17.4).

Water can either be filtered or concentrated by using tangential flow filtration. A sequential filtration using filters with different pore sizes enablesisolation of free-living bacteria from algae-attached bacteria ([31]; Matallana-Surget et al., in prep). Only one study used chemical fixation to stop further protein expression after water sampling [7]. Most of the metaproteomics studies have used liquid nitrogen flash freezing [9,18] or have frozen the concentrated pellets directly to −80°C until protein isolation.

17.2.1.2 Sample preparation

Protein extraction can be tricky depending on the nature of the ecosystem. Optimization steps are often required to overcome the generally low bacterial biomass available and possible interfering substances such as humic acid, cell exudation, various degradation products, lignin, organic and inorganic materials [37]. Indeed,there is no standard and universally efficient protein isolation protocol that could apply to all environmental samples, because of the species-specific distribution of an ecosystem, the wide range of protein abundance levels, and possible interfering substances that are specific to one given ecosystem (Fig. 17.4). In this way, we strongly advise to do a mock experiment when possible, before conducting the real sampling in order to determine the volume of water required and optimize the protein isolation steps.It is essential to try a combination of extraction protocols and protein quantification methods to select the ones that give the best results (see Ref. [37]) (Figs. 17.4 and 17.5).

Protein isolation efficiency relies on the association of chemical reagents (detergent, reducer, chaotropic agent, enzyme) and mechanical disruption (sonication, boiling, glass beads grinding, French press, freeze-thaw cycles). Among the various mechanical methods, sonication is the most commonly used method for cell disruption in metaproteomics.

Protein quantitation is a crucial step in the sample preparation, as inhibiting substances could be coextracted together with proteins and are known to interfere with protein quantification, separation and identification [39]. When the amount and volume of protein isolates allow it, both Bradford assay and SDS PAGE should be performed to ensure the same amount of protein for analysis in mass spectrometry (using, or not, a downstream gel-based approach). Indeed,gels can also be used to compare protein abundances [7]. Protein isolate can be fractionated using 1D- and 2D-gels or various chromatography steps to reduce the sample

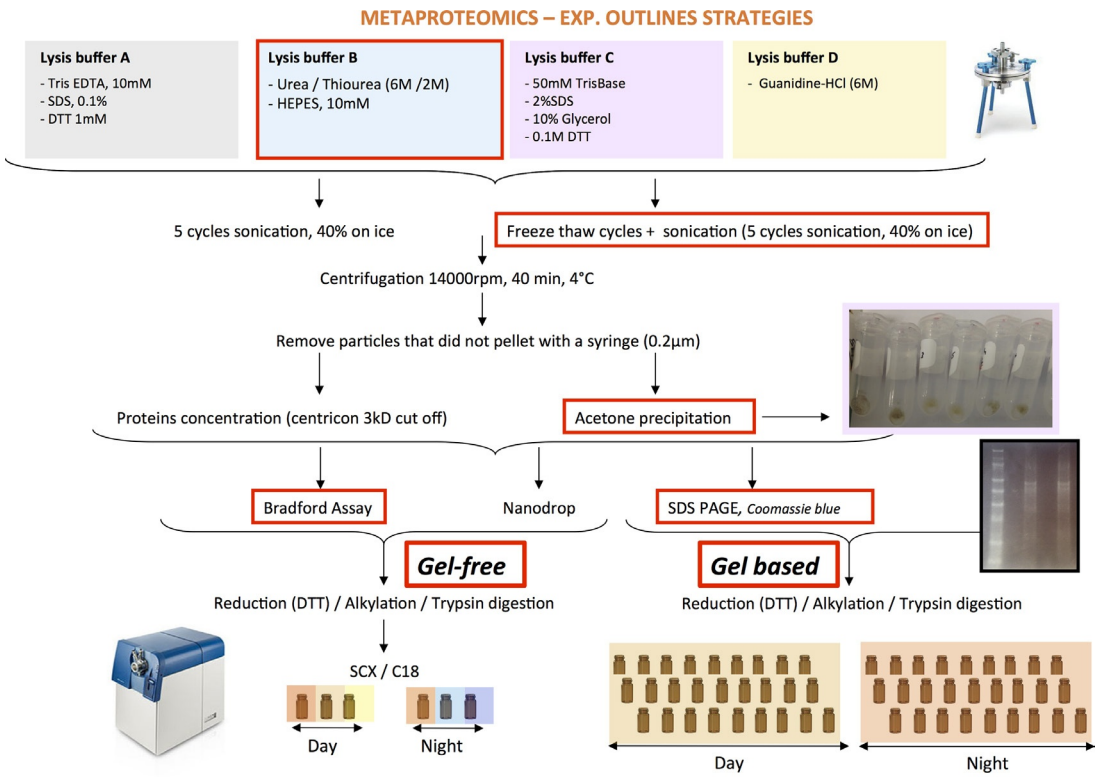


Fig. 17.5
Different steps in the metaproteomics approach on natural microbial communities during day/night cycle. Different experimental strategies tested and the one kept with thicker red boxes (Matallana-Surget et al., in prep).

complexity [35]. Finally, quantitative proteomics can be performed on natural communities by using either in vitro labeling or a label-free method. While in vitro labeling uses tags that specifically react with some amino acids, label-free quantitation reliability depends on standardized and reproducible liquid chromatography (LC) conditions [40,41].

17.2.1.3 Mass spectrometry and data analysis

Due to the high complexity of the sample, peptides are commonly analyzed using tandem MS/MS. Quantitation can be performed using either signal intensity (area under the curve, AUC) or the number of fragmentation spectra (spectral counts, SC) (for review, see Ref. [40]). The AUC method is based on the detection of peptide ion abundances at specific retention times [42]. Protein quantification can only be considered statistically significant when several replicates are performed. Briefly, spectral counting consists of counting the number of peptides per protein to infer protein abundance. The most commonly used quantification methods are (i) the normalized spectra abundance factor (NSAF), and (ii) the

exponentially modified protein abundance index (emPAI). A comprehensive comparison of spectral counting indexes was done by McIlwain and coworkers [43]. The different acquisition methods used in quantitative metaproteomics are discussed in Section 17.3.3. Finally, the quality of identification mainly relies on the database used and this is further described in Section 17.4 (Fig. 17.4).

17.2.2 Example of a Targeted Comparative Metaproteomics Study

To illustrate the optimization phase of the sample preparation presented in the previous section, we present in Box 17.2, the optimization steps of our study, focusing on the day/night cycle presented previously in Box 17.1.

17.3 Metaproteomics on Simplified Artificial Communities

Metaproteomics on synthetic communities is expected to bring about a new wave of findings responding efficiently to major environmental changes (climate change, pollution). Artificial microbial communities represent a simplified level of complexity in comparison to natural microbial communities. Complex artificial communities have proven to be successful inanswering many important ecological questions regarding the role of interspecies interactions in stressful conditions [45], community evolution [46–48] or community functioning [48–52]. Previous studies demonstrated that higher microbial diversity (richness, evenness) can contribute to better tolerance to stresses and species invasion [47,52–55]. In this section, we will present and discuss the main challenges that are specific to the design of synthetic communities, ranging from the engineering of the artificial community itself to the biological interpretation. Although bacterial cocultures can be easily manipulated and provide a suitable approach for comparative studies, only two studies using metaproteomics on artificial communities have been published [34,56]. The challenging steps specific to a synthetic community are presented in Fig. 17.7.

17.3.1 Design of an Artificial Community

The main challenge in the design of an artificial community is to identify a group of bacteria that can fulfill most of the criteria listed in Table 17.3. First of all, the obvious criterion is the relevance of the selected bacteria regarding a specific ecosystem and/or a specific biological question. One straightforward way would be to select candidates from the literature or from known relevant metagenomic studies of natural environments. Selected microorganisms need to be culturable in laboratory conditions. Their genomes need to be entirely sequenced allowing proteomics application. Selected bacteria should present similar growth rates in the conditions of interest. In the context of a stress study, where CFU (colony forming units) could be performed, selected bacterial strains will have to provide different morphotypes and/or different pigmentations, to allow easy colony counting (Table 17.3). Moreover, it will be

Box 17.2 Optimization phase of the sample preparation

We conducted a comprehensive investigation of the impact of the day/night cycle on microbial communities, using comparative metaproteomics as introduced in Box 17.1. Prefiltered surface seawater was collected during the day and night, by sequential filtration (onto 142-mm 0.8-µm and 0.2-µm polyethersulfone filters) for three consecutive days.

We tested different extraction buffers as well as different concentration steps using either centricon centrifugal filter or acetone precipitation. The steps that have proven the most successful are highlighted with thicker boxes in Fig. 17.5.

Regarding the protein quantification steps, both Bradford assay and SDS PAGE confirmed the same results. The profiles obtained for both communities were found to be reproducible between replicates and very distinct between both 0.8-µm and 0.2-µm communities (Fig. 17.6) (Matallana-Surget et al., in prep).

We performed both gel-free and gel-based approaches and we interestingly observed that only 30% of the identified proteins were found to be common between both approaches. Indeed, it has already been demonstrated that both gel-free and gel-based approaches are complementary [36,44]. We were able to analyze the day/night oscillation by using an in-house database targeting specific organisms of interest, abundant and representative of the studied ecosystem. The main functions identified were as follows: proteins involved in photosynthesis, cell division, motility, DNA repair, translation regulator, and viral proteins. Only 20% of proteins were found to be common between the day and the night conditions, out of three replicates (Matallana-Surget et al., in prep). The abundance of key proteins was found to be rhythmically regulated and this was confirmed in the three replicates. Performing replicates of sampling enabled strengthening of biological interpretations.

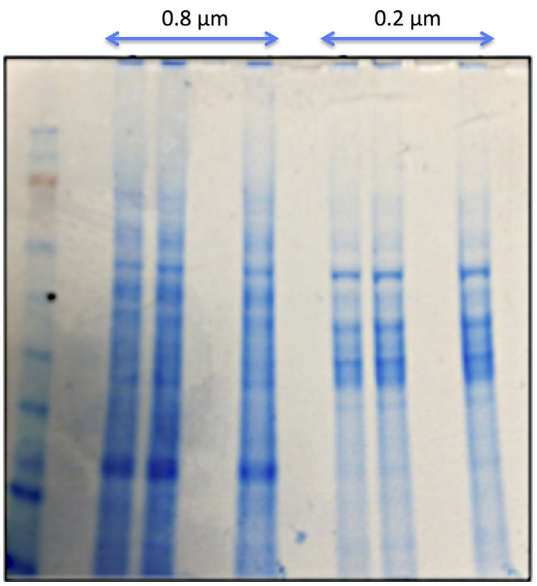


Fig. 17.6
SDS PAGE (Bis-Tris, 4%–12%), showing the soluble proteomes isolated from the natural microbial communities (0.8-µm and 0.2-µm filters) (Matallana-Surget et al., in prep).

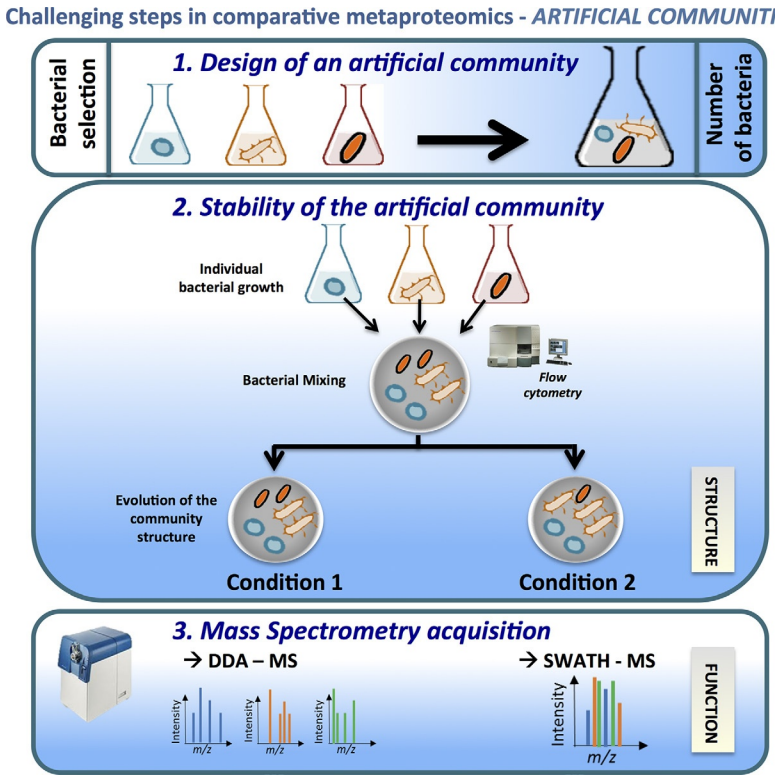


Fig. 17.7
Challenging steps in quantitative metaproteomics using artificial communities.

Table 17.3: List of criteria for the strain selection of an artificial community

1. Representative of the studied ecosystem
2. Culturable bacteria
3. Fully sequenced genomes
4. Similar growth rate
5. Phenotypically different (different morphotypes/pigmentation)
6. Compatibility to grow in coculture (none synthesizing antimicrobial substance(s))

essential to test the growth of the bacterial candidates in community because some strains could synthesize antimicrobial substances, inhibiting the growth of other bacteria and thus introducing a bias in our study.

Artificial communities are synthetic assemblages usually made up of less than 20 bacterial strains. They are cultured in artificial ecosystems, i.e., in controlled conditions mimicking their natural environment. The ideal number of selected species would be 10 to 20 species, preserving the complexity of the system. Several studies on artificial communities demonstrate an improvement of community functioning and stability with species richness

[49,50]. However, working with more than 20 strains can be very complex and fastidious. Growth characteristics and stress behavior (if relevant) should be analyzed for each strain before undertaking any experiments on the final artificial ecosystem. To ensure reproducibility between replicates, each strain should be individually adapted to artificial ecosystem conditions prior to mixing [57]. Artificial communities can be frozen after mixing, making high-throughput studies easier [58].

17.3.2 Stability of the Artificial Community

The stability of the artificial community must be carefully monitored during the entire experiment. The proportion of inoculated cells in the artificial community must be accurately measured by flow cytometry, or optical density if preliminary correlation to CFUs has been performed [34]. Initial community evenness is important for the functioning of the community [53,55]. The artificial community will evolve much more rapidly than a natural community and this represents one of the main drawbacks of this approach. Indeed, complex interaction networks produce greater stability than simple interactions existing in a synthetic community. An initial consortium made of nine bacterial species can lead to a final community containing only four detectable bacteria (Beraud et al., in prep). In that respect, structural analysis of the community is a crucial step to determine relative abundance of each species. The evolution of the community structure can be measured by CFU, QPCR [46,47] or shotgun sequencing [47]. Recently, Rubbens and coworkers have proposed a new method using flow cytometry to assess the structure of synthetic bacterial communities based on distinct strain characteristics [59].

17.3.3 Sample Preparation, Mass Spectrometry Acquisition and Data Analysis

The sample preparation for proteomics is similar to that described previously for natural communities. Several reviews are already available on sample preparation [35]. Synthetic communities offer two advantages, which are (i) the diversity of quantitative proteomics methods (label-free, in vivo metabolic labeling, in vitro labeling), and (ii) the diversity of mass spectrometry acquisition methods.

Two main acquisition methods are commonly used: data-dependent acquisition (DDA) or data-independent acquisition (DIA) (Fig. 17.4). The DDA method uses the MS spectrum to target a subset of peptides for downstream MS/MS acquisition. Consequently, the number of peptides sampled is limited by the MS/MS sampling speed, despite the dynamic range and peak capacity of the mass analyzer [60]. On the other hand, the DIA method allows a systematic acquisition, independently of precursor ion information. This method also presents many variations such as (i) collecting fragmentation data without precursor ion selection and (ii) using wide isolation windows [61]. Even though the DDA-MS method has been successfully used for artificial community studies, this approach cannot be applied to a low-diversity bacterial community, dominated by only one or two species (Beraud et al., in

prep). Indeed, protein identification from low-abundance bacteria is still one of the major challenges in mass spectrometry-based metaproteomics. To overcome this issue, a Sequential Windows Acquisition of All Theoretical ions approach (SWATH-MS) was developed [62].

The advantage of the SWATH-MS method is that it allows all data to be acquired (DIA method) as well as a postacquisition targeted analysis. Targeted peptides of interest are searched among all acquired MS/MS spectra, using a spectral library obtained in a data-dependent acquisition mode [62]. In this way, the quality and the coverage of the library are of crucial importance. Indeed, only the proteins present in the spectral library could be identified and quantified [63]. In regards to the metaproteomics approach using artificial communities, the spectral library can be obtained using DDA-MS analysis of single-species cultures (Fig. 17.8). Nevertheless, in order to be consistent and allow the highest protein coverage, the spectral library should contain proteomes from at least two species and include different bacterial growth conditions. SWATH-MS presents numerous advantages in comparison to DDA, such as high sensitivity and high accuracy of quantification [62].

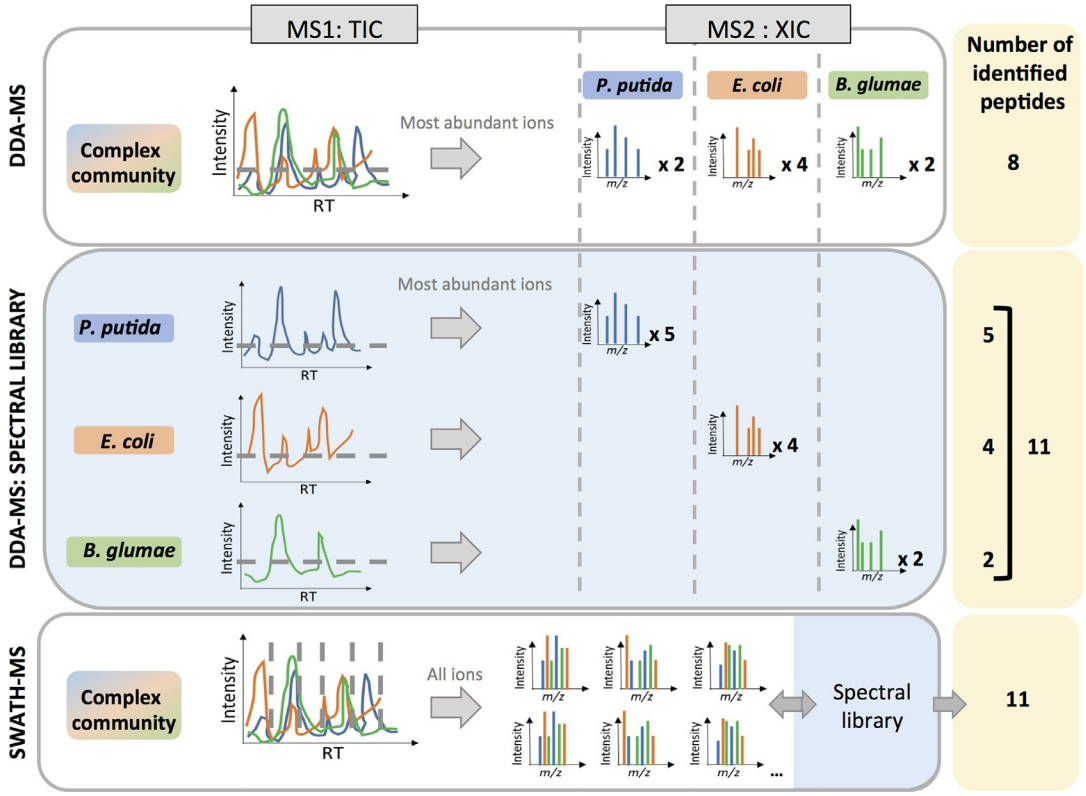


Fig. 17.8
Mass spectrometry acquisition in quantitative metaproteomics. In this example, theartificial community is composed of threespecies (*E. coli, P. putida* and *B. glumae*) (Beraud et al., in prep).

Regarding data analysis, working with artificial communities simplifies this step, which is challenging for natural communities. Indeed, it is possible to create a species-specific database, with all the proteomes from the bacteria present in the synthetic consortium. Interestingly, Tanca and coworkers used an artificial community composed of nine species to compare the potential of metagenomic vs. species-specific databases. The authors demonstrated that only 37% of the peptides were common to all databases [64].

As discussed earlier, a synthetic community is generally not as stable as the one observed in the environment. For these reasons, proteomics data needs to be carefully interpreted, because the differential protein abundance between two conditions can be (i) the result of a differential protein regulation in itself, or (ii) a modification of the community composition over time during cultivation or between the tested conditions, as illustrated in Fig. 17.9. It is thus crucial to determine the community structure at the beginning and the end of the experiment.

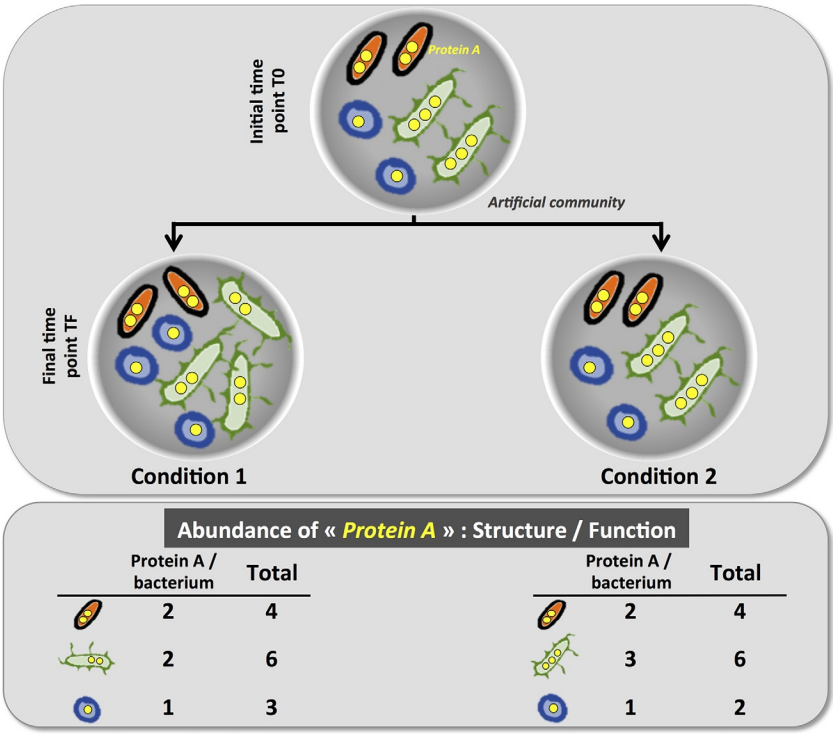


Fig. 17.9

Quantitative metaproteomics analysis using artificial communities. This example presents the differential abundance of the protein A for three bacterial species between the initial time (T_0) and final time (T_F) for two different conditions. This figure highlights that both the number of bacterial cells and abundance of proteins will evolve over time. The protein abundance can be analyzed at two different levels: the cellular level or the community level.

17.4 Data Analysis

Metaproteomics data analysis can present an analytical challenge mainly due to the large and complex nature of the metaproteomic protein sequence database, which in turn compounds false discovery rate (FDR) statistics [65] and the protein inference problem. Moreover, the analysis also necessitates extraction of taxonomic and functional information from identified peptides and proteins. This can only be accomplished by the use of multiple analytical tools to generate outputs that eventually facilitate biological interpretation. Despite these challenges, some approaches have been developed in the last 13 years. The analytical approaches are mainly composed of: (a) database generation; (b) database search; (c) functional analysis; and (d) taxonomy analysis (Fig. 17.10).

17.4.1 Database Generation

Many factors determine the composition of the protein sequence database, also referred to as the protein search database, whichcan be used to search the mass spectra. These include factors such as the source of the sample, sample preparation methods and the focus of investigation. Studies with a dataset with known microbiota and using variable compositions of the protein search database have highlighted the importance of the choice of protein search database [66]. One major challenge presented by multiorganism proteome databases is that they can be of large size, thus affecting the sensitivity of identifications due to its effect on

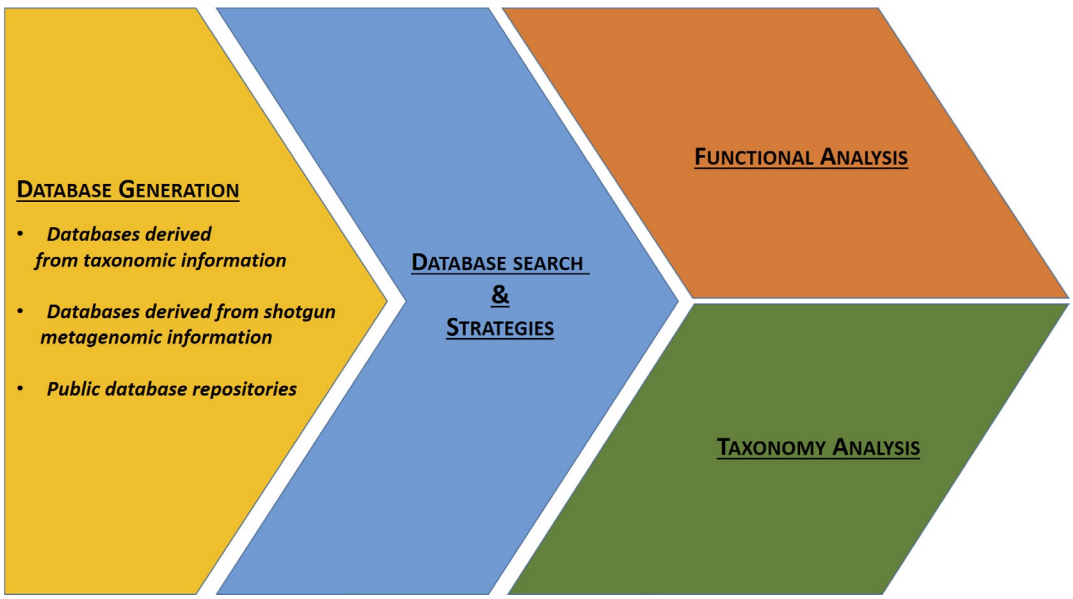


Fig. 17.10
Analytical steps involved in metaproteomics data analysis.

FDR statistics [65]. While keeping these challenges in mind, we highlight here a few options so as to arrive at a composition of the protein search database that is as accurate as possible.

17.4.1.1 Databases using taxonomy information

A majority of metagenomic analyses are carried out using 16S rRNA sequencing in order to assign operational taxonomic units (OTUs) in the form of species, genera or phyla. However, this metagenomic analysis is restricted to determining phylogenetic composition and the information can be used for constructing metaproteomic databases. It should also be noted that the 16S rRNA method is well suited for analysis of a large number of microbiome samples, but suffers from limited taxonomical and functional resolution. Taxonomic information available via the OTU table can be used to generate UniProt-based partial or complete proteomes using a UniProt application programming interface (API). In some cases, a list of genera is available through publications and this can be useful in generating a protein database for metaproteomic search.

Recently, a Galaxy-compatible tool (see Section 17.4.5) and workflow was made available that takes in an OTU table or list of taxon identifiers or genera names as an input and generates a merged metaproteome sequence database for search (https://github.com/galaxyproteomics/tools-galaxyp/issues/86).

Taxonomy information–derived databases have the advantage of being more specific to datasets under study as compared to the public repositories (see Section 17.4.1.3) that are commonly used in metaproteomics research. However, the large databases still suffer from limitations with respect to sensitivity of identifications, as described earlier. Tanca et al. have investigated the impact of database selection on metaproteomic identifications regarding depth and reliability of results and have proposed using iterative searches and suitable filters for reporting metaproteomics results [66].

17.4.1.2 Databases from shotgun metagenomic approach using taxonomy information

Shotgun metagenomics, which uses a whole metagenome sequencing method, offers increased resolution (over 16S rRNA sequencing), thus enabling more accurate taxonomic and functional categorization of identified sequences [67]. Shotgun metagenomics is, however, more expensive than 16S RNA sequencing. Recently, May et al. published a method that generated a "metapeptide database" from shotgun metagenomics sequencing. This approach was used on two large ocean metagenomics samples and they demonstrated that this led to significant increase in the number of identifications (presumably due to a more accurate and compact database) as compared to an assembled predicted metaproteome and NCBInr [68].

Interestingly, the authors also followed up with another publication on the effect of the quality and size of the search database on taxonomic and functional analysis, which has a profound

effect on biological conclusions [69]. The authors recommend a general best-practices guide that could be useful before undertaking a large metaproteomics study.

Omega (*o*verlap-graph *me*ta*g*enome *a*ssembler) is another tool that is available for assembly of shotgun metagenome data. Omega uses the graph overlap approach to provide continuous assembly of metagenomics data. This can be used along with the SIPROS search method described in Section 17.4.2.

Lastly, Tang et al. published a graph-centric approach that uses the *de bruijn* graph structure in metagenome assembly algorithms to improve peptide and protein identification in metaproteomics [70]. Most of these approaches have been recently proposed and will need to be evaluated by users so that they can be optimized.

17.4.1.3 Public repositories

For the last 13 years, the most commonly used protein databases for metaproteomics searches have been from publically available proteomic databases, such as NCBInr, UniProt, etc. [14,71–80], or from generation of protein databases based on literature surveys and using reference proteomic databases [18,76,81–85]. In the last three years, researchers have started to use metagenome information-derived databases (as described in Sections 17.4.1.1 and 17.4.1.2) in their metaproteomics [13,17,27,68,69,84,86]. In the field of marine metaproteomics, researchers have also used the Global Ocean Sampling (GOS) database, which provides access to 584 annotated genomes [87–89].

In addition to this, environmental metagenomics databases are available on the EBI Metagenomics website (https://www.ebi.ac.uk/metagenomics). Tools described in Sections 17.4.1.1 and 17.4.1.2 can be used on these public repositories to generate environment-specific metaproteomic databases. Researchers are highly recommended to read the suggestions from Timmins-Schiffman et al. [69], Tanca et al. [64] and Muth et al. [90], so as to choose the best approach for constructing databases for metaproteomics searches.

17.4.2 Database Search

Multiple database search algorithms are available for proteomics research [91] and algorithms that are able to search large databases have been used in metaproteomics research (see Table 17.1). In particular, search algorithms such as Sequest, Mascot, OMSSA and ProteinPilot have been used due to their fast searching speed and the ability to generate outputs that are compatible with downstream processing steps.

In addition to this, a newer generation of search algorithms is being developed that has in mind the needs of metaproteomics applications, namely: a) ability to search large databases; b) speed; and c) peptide or PSM output with robust FDR threshold calculations. A few approaches are noteworthy and would be worth testing and optimizing before they are

routinely used in metaproteomics research. Chatterjee et al. published an approach using the ComPIL (Comprehensive Protein Identification Library) database generation method and BlazMass was used to search this database. The combination of ComPIL with Blazmass allows larger database proteomic searches than were previously possible. The authors claim that this method can be used on complex metaproteomics datasets whose microbial composition is unknown [92]. This could potentially lead to identification of metaproteins from datasets without any metagenomics, metaproteomic or public repository databases.

Wang et al. published SIPROS, a search algorithm that was originally used for quantitative analysis and single amino acid polymorphisms [93]. One of the advantages of using SIPROS is the scalability with respect to the number of computational cores that it can use, thus making it possible to search a large number of spectra against extremely large databases. Muth et al. have suggested use of de novo search algorithms to identify microbial peptides from metaproteomics samples [90]. They have also suggested using multiple search algorithms in order to increase the percentage of peptide spectral matches in a dataset. Our laboratory has been using SearchGUI/PeptideShaker within the Galaxy platform (see Section 17.4.5 and Ref. [94]),in whichat least eight open-source search algorithms can be used for search against large databases.

Most of the database generation strategies suggested in Section 17.4.1 generate large databases, which in turn affect the sensitivity of identifications. Multiple strategies have been suggested to increase peptide identifications. This includes the two-step method for searching large databases [65,74,77,95,96] wherein a first-step search is used to generate a FASTA file for a rigorous FDR based second-step search. A cascaded search method has also been demonstrated to increase the number of identifications of rare, low-abundance peptides [97]. Muth et al. have recommended using a database sectioning approach so that searches against subsets of a large database may increase the number of identifications [90]. Regardless of the choice of search algorithms, the goal is to generate outputs that are compatible with the next steps of taxonomic analysis (Section 17.4.3), functional analysis (Section 17.4.4) and subsequent targeted validation.

17.4.3 Taxonomy Analysis

Although metagenomics approaches can be used to decipher the taxonomic composition of the microbiome, peptide level identifications from metaproteomic investigations can also be used either independently or to compare and confirm metagenomics findings.

In metaproteomic studies, microbial peptides identified by searching mass spectrometry data by using database search methods (Section 17.4.2) against an appropriate search database (Section 17.4.1) can be used to determine the taxonomic composition of the dataset. For this, the list of identified peptides is searched against a UniProt database (using the UniPept tool)

[98] or the NCBInr database (using BLAST-P). The list of taxonomic identifications is then subjected to least common ancestor (LCA) analysis to yield a list of taxon identifications (kingdom, phylum, genus or species). UniPept generates outputs such as CSV tabular output and a phylogenetic tree so that data can be visualized (Fig. 17.11). Given that the BLAST-P step is the rate-limiting step and with the recent changes to the NCBI database structure, developers have been exploring alternative approaches using DIAMOND and UniPept module changes along with tools in MEGAN (such as MEGANIZER) to generate inputs that can be processed using MEGAN.

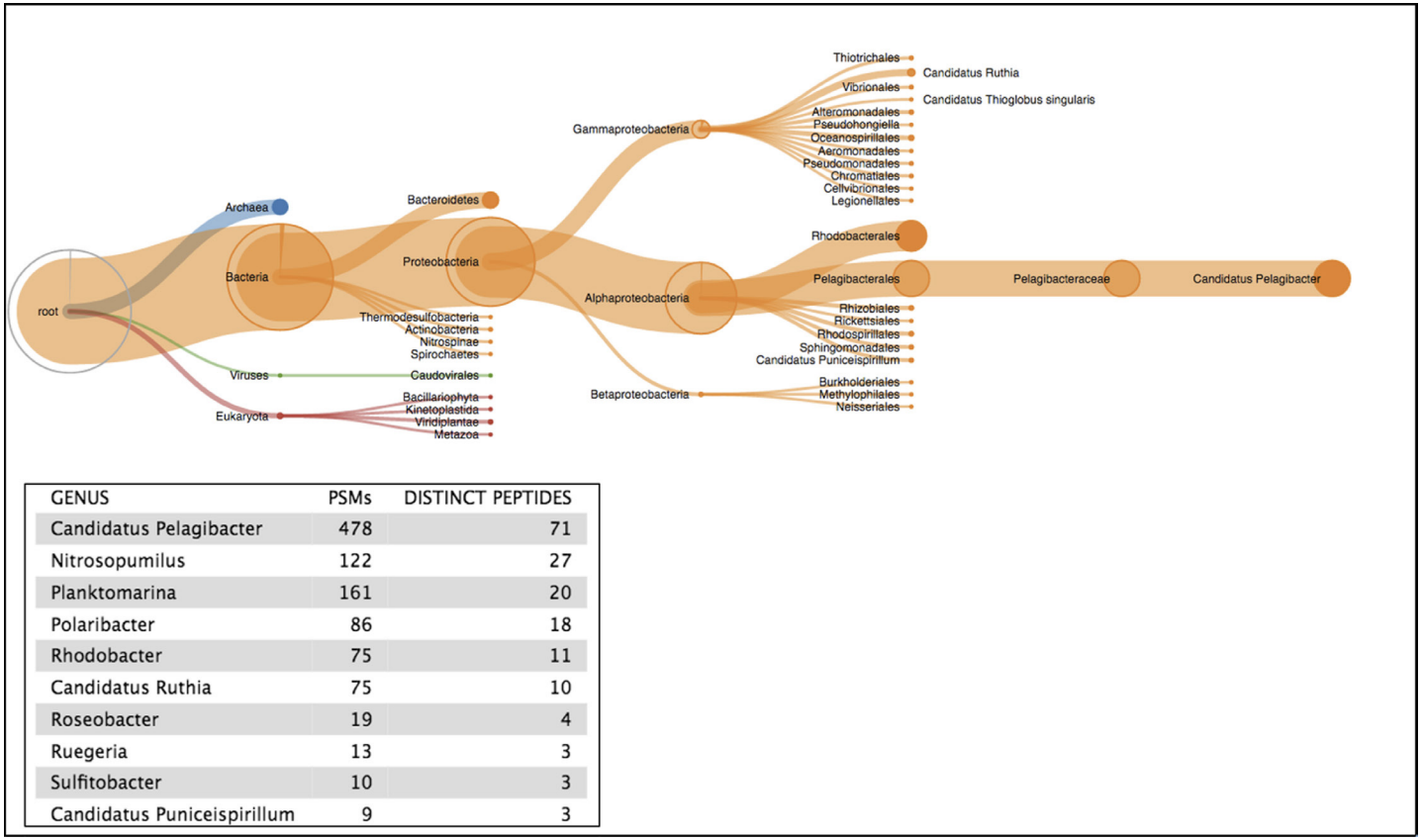
Prophane is one of the suite of tools that is available for taxonomic analysis. Data generated after database searching is processed by CLUSTALW and annotated using either BLAST-P or annotation finder, and the results are used to perform taxonomical annotation. Taxonomical lineage is determined by grouping all related members to a taxonomic unit [99]. Muth et al. have published a tool named MetaProteomeAnalyzer that uses outputs from SearchGUI/PeptideShaker (described in Section 17.4.2). The software developers used graph database method to assign the identified peptides to proteins, and the proteins were then used to identify species using the common ancestor method [100].

17.4.4 Functional Characterization

While metagenomics studies can perform predictive functional analysis using shotgun metagenomics data [101] or 16S rRNA [102], metaproteomics has a distinct advantage in determining the protein functional expression by a microbial community [77]. For example, there was variability in the taxonomy distribution of human microbiome samples among body habitat and individuals. However, most metabolic pathways are prevalent and distributed evenly across individuals and body habitats [103]. This indicates that the metabolic pathways are much more stable and could potentially be used as a measure of baseline state of the microbiome.

Prominent among these are Gene Ontology analysis, COG analysis, MEGAN analysis and EggNOG analysis. MEGAN6 in particular can be used to carry out InterPro2GO, KEGG, SEED and EggNOG analysis to determine the distribution of functions amongexpressed proteins in the microbiome [104].

The Prophane suite of tools (described for taxonomy analysis in Section 17.4.3) can also be used to determine the distribution of functions in a microbiome sample. Prophane uses RPSBLAST or HMMER3 outputs to perform functional prediction for each protein, based on either COG classification or TIGRFAMS (for prokaryotes) and KOG classification or PFAMS (for eukaryotic proteins) (http://www.prophane.de/index.php). MetaProteome Analyzer (described in Section 17.4.3) provides enzyme and pathway display options where proteins aggregated by E.C. numbers and KEGG pathways can be visualized [100].



GENUS	PSMs	DISTINCT PEPTIDES
Candidatus Pelagibacter	478	71
Nitrosopumilus	122	27
Planktomarina	161	20
Polaribacter	86	18
Rhodobacter	75	11
Candidatus Ruthia	75	10
Roseobacter	19	4
Ruegeria	13	3
Sulfitobacter	10	3
Candidatus Puniceispirillum	9	3

Fig. 17.11

Phylogenetic tree view and genera information about taxonomy of Bering Strait dataset. Metaproteomics analysis using Unipept Peptide to LCA analysis (Pept2LCA) provides a phylogenetic treeview of the Bering Strait dataset [68]. The Pept2LCA application identifies the taxonomic lowest common ancestor(s) for a batch of tryptic peptides. The outputs from Pept2LCA can also be used to parse out information about genera, unique peptides and PSMs associated with the genera using the Galaxy framework (see Section 17.4.5).

There is a greater need in data interrogation and visualization options in the taxonomy and functional analysis of metaproteomics datasets. Exciting developments are anticipated via collaboration among mass spectrometrists/analytical chemists (data acquisition), developers (algorithm development) and users (project implementation). The next section describes features and progress using the Galaxy bioinformatics infrastructure that offers such a platform for collaborative development.

17.4.5 Galaxy Platform: Metaproteomic and Taxonomic Data Analysis

As described previously, metaproteomic data analysis comprises numerous steps and a number of separate software tools requiring integration into operational workflows. To achieve a goal of automated and reproducible data analysis in metaproteomics, there is a necessity for a single environment where these tools can be deployed, integrated into workflows, and made usable by both data scientists and wet-bench researchers. How can such a need be addressed? Fortunately, an answer to this question exists in the form of the Galaxy bioinformatics platform.

Galaxy is a freely available, open, web-based bioinformatics platform [105]. Originally created with a focus on genomics research, Galaxy was developed to provide a flexible environment to deploy disparate software and integrate these into workflows within a user-friendly and scalable platform. A web-based interface is user-friendly and allows tools and workflows tobe easily accessed. Galaxy is built with the user experience in mind, automatically saving parameters and intermediate data generated in complex analysis pipelines, and guiding users in questions of input file format requirements and compatibility of data files with software tools. The platform can also be implemented on large-scale, high-performance computing infrastructure, providing the storage, memory and processor power that is necessary in many contemporary areas of'omics data analysis – including for metaproteomics.

An additional strength of Galaxy as a metaproteomics bioinformatics solution is the active community of developers and users. The core Galaxy team maintains and upgrades the framework in response to feedback from the community. The core team has also developed and made available numerous training tutorials for new users (e.g., see https://wiki.galaxyproject.org/Learn#Galaxy_101). In addition to the core development team, an active world-wide community of developers helps to enhance Galaxy by contributing new software from a variety of domains that can be used by the entire Galaxy community. This collaborative team approach has greatly benefited the community of researchers using Galaxy, as researchers have collectively shared the burden of extending the functionalities of the software, without dependence on a single lab to be responsible for all new developments. Accordingly, metaproteomics tools in Galaxy have also benefited from a similar team approach by a network of researchers from across the globe.

Galaxy operates as a web-based interface, viewable in any web browser. Fig. 17.12 shows a screenshot of the Galaxy interface. The basic layout of the interface includes a Tool

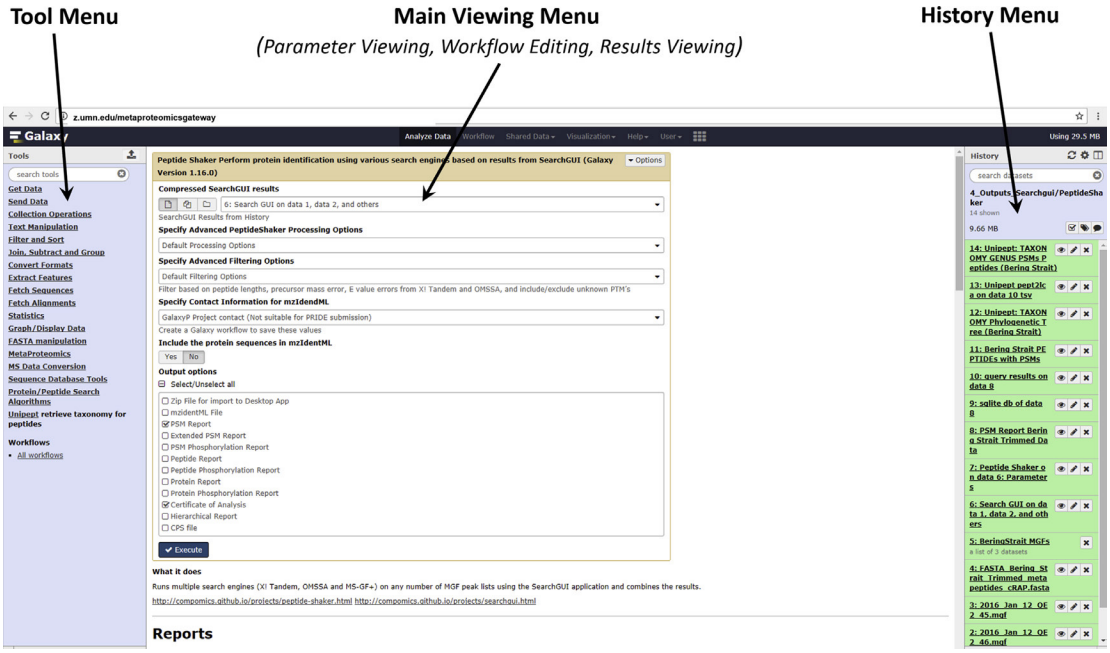


Fig. 17.12

Galaxy platform interface. The Galaxy platform interface contains (a) *tool menu* for various software tools available for analysis; (b) *Main viewing pane* for viewing tool parameters, editing workflows and viewing results; and (c) *History menu* that contains input data and generated output files after data processing.

menu, which contains the list of available software within the instance being used, which can be customized to the application. The main viewing pane offers a larger space to view parameters for tools, edit workflows, and open visualization tools to view results.

The History menu is a real-time record of active software operations and their status. The History menu contains a record of all data used as inputs for the analysis, as well as any intermediate or final result files that are produced. Users can build Histories from scratch, by uploading data that will act as input, and selecting appropriate software that act together within a data analysis pipeline. When the software tools are executed, Galaxy records all steps and archives the data analysis event as distinct History. This complete record of an analysis is saved in the user's account, and is only removed if actively deleted by the user.

A saved History can be reaccessed at any time. Outputted results can be downloaded. New software tools and analysis steps can be added to the existing History if further analysis of results files is desired. Any step in the History can also be rerun, using either the original parameters or adjusted parameters. The History can be renamed and saved as a different record using adjusted parameters, if desired.

Workflows are the other main operational unit in Galaxy, and are used commonly for multiomic applications such as metaproteomics. Workflows consist of the processing steps and software tool parameters that make up a History, with the exception that the Workflow does not contain any input or output data. As such, a Workflow can be thought of as a scaffold containing all relevant software and parameters needed to run a data analysis pipeline. Once appropriate input data is uploaded to an active History, the Workflow will recognize this input and can be run. Once run, the Workflow, along with all input and output data, becomes an archived History. Workflows using multiple software tools in sequential steps can be easily built in Galaxy. The Workflow canvas enables users to graphically build customized workflows from software in the Tools menu. The Workflow editor (Fig. 17.13) guides users through the development of multistep pipelines by only allowing tools with compatible data outputs to be linked to the next software tool.

Workflows can also be created from a completed History. A user can simply use the "Extract Workflow" function, which extracts the software tools and parameters that make up the data analysis pipeline, while leaving behind any specific data inputs or results from the History.

A very powerful feature of Galaxy is its amenability to sharing complete Histories and Workflows with other users. This enables not only dissemination of software and pipelines, but also promotes reproducibility and transparency in complex data analysis operations. Sharing can be done in multiple ways. Histories or Workflows can be published, such that they will show up in a shared folder available to all other users of the Galaxy instance. A URL link can also be created that points to a specific History or Workflow. This URL can be shared with selected users of the Galaxy instance to give them access. Histories and Workflows can also be downloaded as a Galaxy file type, which can be imported and used in other Galaxy instances.

Given these collective features available through Galaxy, it is well-suited as a platform to enable the sophisticated analysis pipelines required in metaproteomics. A paper describing a blueprint using Galaxy for metaproteomics was recently published by our group [95], which highlights the advantages of this platform. Since this publication, a number of enhancements have been made, with tools in place that cover the main aspects of metaproteomics data analysis (see Fig. 17.14 outlining some of the Galaxy metaproteomics tools). These include tools for generating customized protein sequence databases, derived from a variety of sources. Tools are in place to translate protein sequences *insilico* from genomic information derived from microbial communities. For studies of microbial communities residing in a host-organism, tools are in place to combine the database from the host with sequences for the microbes.

For sequence database searching, the SearchGUI [107] and PeptideShaker [108] tools have been deployed. SearchGUI bundles multiple popular and freely available sequence database searching programs for matching tandem mass spectrometry (MS/MS) data to peptide sequences. Use of

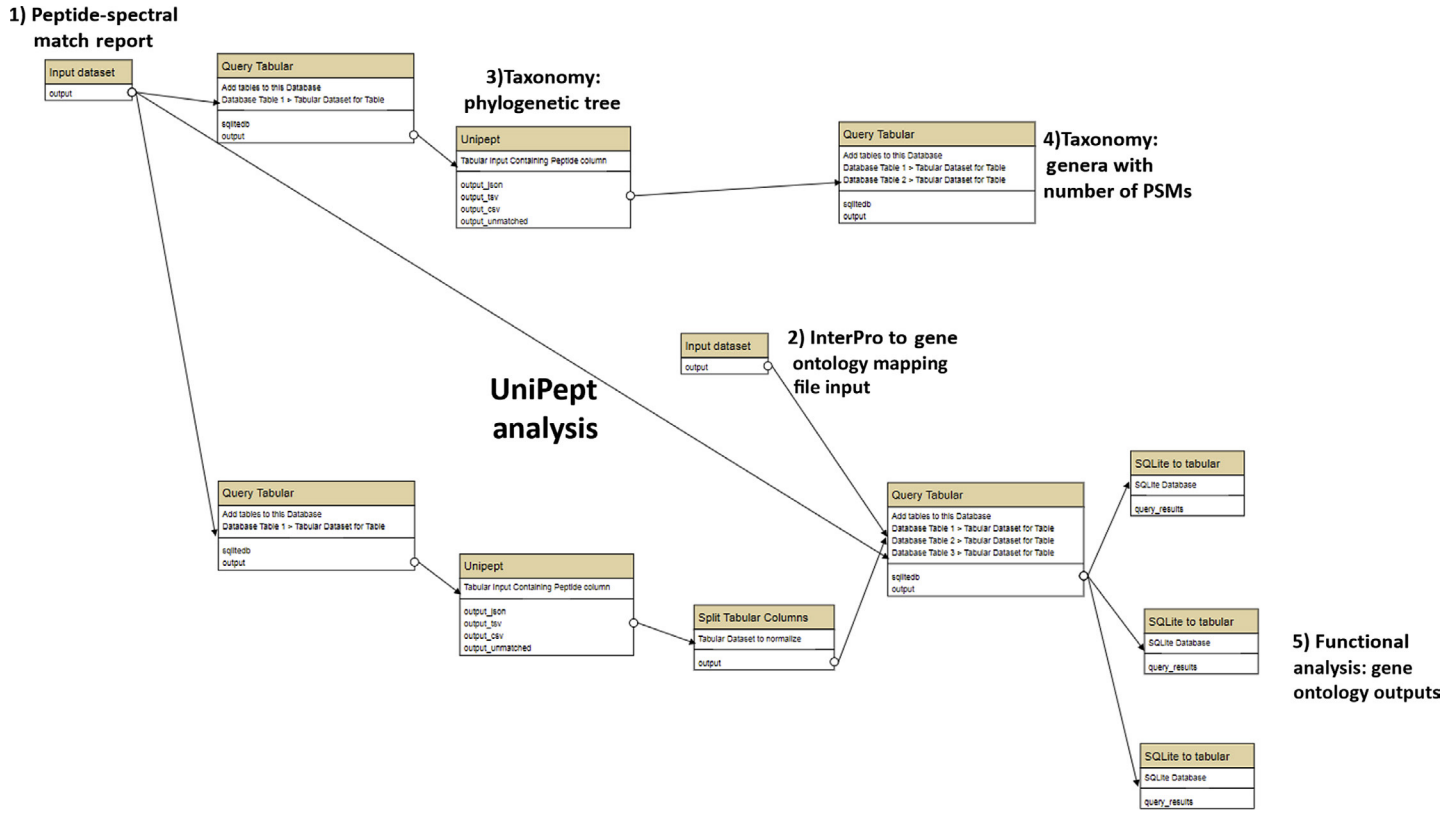


Fig. 17.13

Galaxy workflow to generate outputs for taxonomic analysis outputs and functional analysis. Representation of the "Edit mode" of a Galaxy workflow wherein PSM reports generated from SearchGUI and PeptideShaker (see Section 17.4.2) are used as an input along with a Gene Ontology Mapping file to generate outputs for taxonomic analysis (see Fig. 17.11) and functional analysis using Unipept software [106].

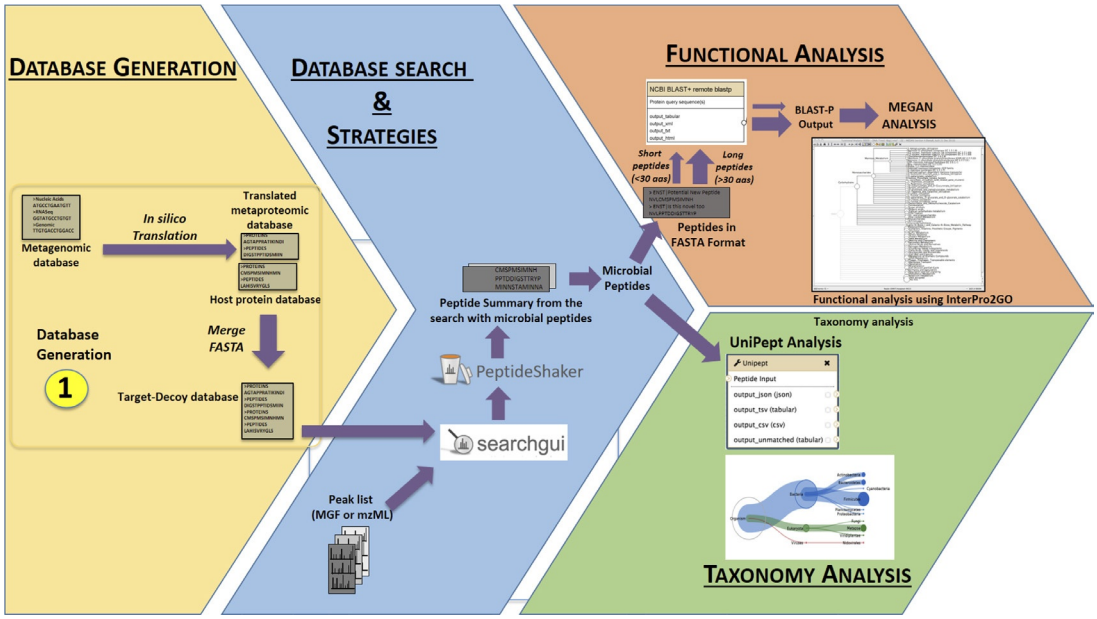


Fig. 17.14
Analytical modules available for metaproteomics data analysis using the Galaxy platform (Section 17.4.5).

complementary software programs helps increase the depth and confidence of peptide sequence matches [109]. PeptideShaker organizes the outputted PSMs, applies statistical analyses to establish confidence of results, and infers protein identities from the peptide data.

The identified peptide sequences outputted from the PeptideShaker program require further processing to determine functional and taxonomic characteristics of the sample being analyzed. We have described the use of BLASTP in Galaxy [95] to automatically screen peptide sequences putatively derived from microbes. The BLASTP filtering step helps to confirm that these sequences indeed derive from microbes, and not the host organism (e.g., human). The confirmed peptides can be downloaded and sent to the MEGAN bioinformatics tool for further functional and taxonomic analysis. Our group is also prototyping the use of InterPro2go web service tool [110], where a Galaxy tool queries this database with peptide sequences and associated protein accession numbers to retrieve functional annotation in the form of Gene Ontology (GO) terms. The returned GO annotated data is provided as a tabular data output in Galaxy, which can also be visualized using a number of simple graphing tools available within Galaxy.

For taxonomic analysis, the Unipept tool [98], also described above, offers a web service for querying taxonomy profiles from identified microbial peptide sequences. Unipept determines the level of taxonomic specificity for any given peptide sequence (e.g., genus, species) and builds visualizations for these. A prototype Galaxy tool is currently in development for

automatically querying the Unipept software and returning taxonomic information, and data files needed to open web-based interactive visualizations for the data file.

17.5 Conclusions

Before undertaking any steps of the metaproteomics workflow, the biological question and research objectives should be clearly defined. In conclusion, we summarize the pros and cons of performing metaproteomics on natural or artificial communities. Fig. 17.15 summarizes the advantages of both approaches: natural vs. artificial communities.

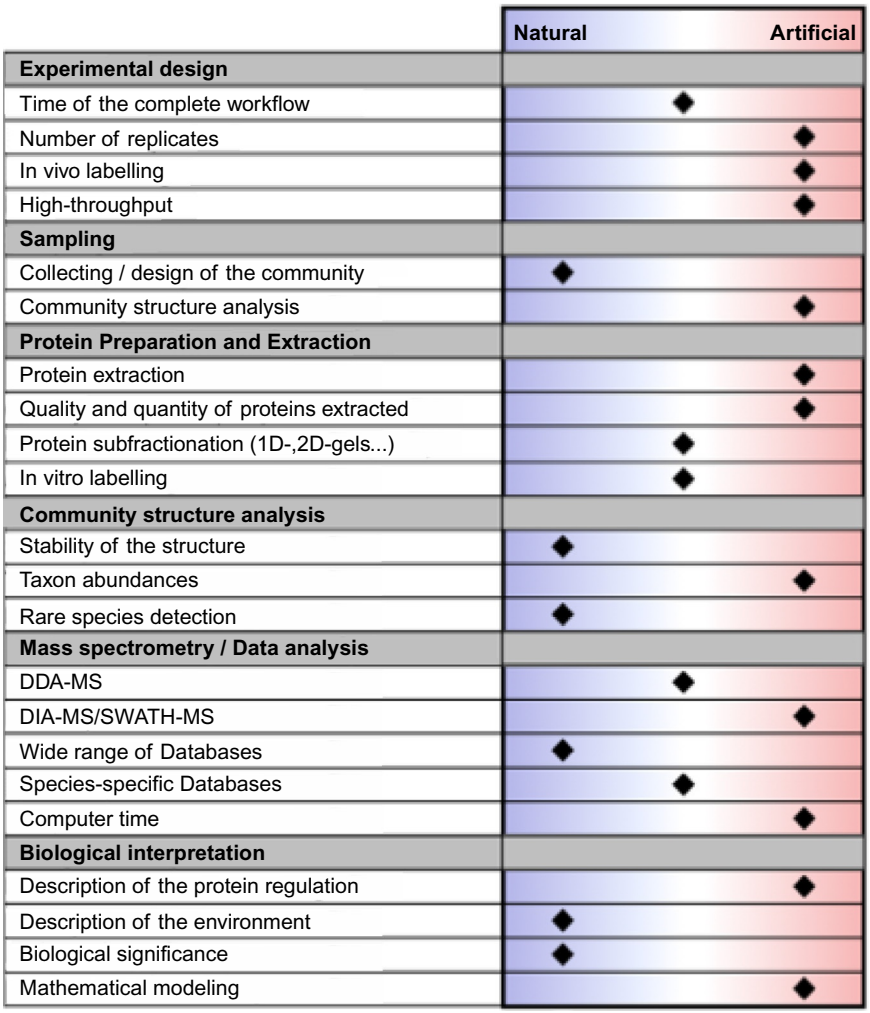


Fig. 17.15
Pros and cons of natural vs.artificial communities. Cursors, represented as black diamonds, show the easiest approach for each step of the comparative metaproteomics workflow.

Artificial versus natural communities: which is the best-suited strategy for each step of the metaproteomics workflow?

17.5.1 Pros for Natural Community Approach

Natural microbial communities are mainly represented by nonculturable bacteria, and for this reason better represent the reality of the response to environmental change, a perturbation in comparison to artificial community.

Even though sampling can be long and tedious, this first step of the experimental set-up appears to be less challenging than engineering a suitable synthetic community. Choosing the bacterial strains of a synthetic consortium requires a long optimization phase, as described in Section 17.3.1.

Another advantage of undertaking metaproteomics on natural communities is the stability that it offers the system. Indeed, complex interaction networks within natural dynamic communities comprising a wide diversity of microorganisms, with overlapping ecological functions, produce greater stability than artificial communities [111]. Once the diversity and/or metagenome have been described, metaproteomics can be performed using the same metagenome database because of the stability of a natural ecosystem.

Biological interpretation can be highly complex and tricky; however, it is important to remind one here that the level of complexity of the databases created only depends on the objectives of the study, and thus targeted comparative metaproteomics can offer an intermediate level of complexity.

17.5.2 Pros for Artificial Community Approach

Although artificial communities do not reflect the complexity and dynamics of natural microbial assemblages, this approach also offers several advantages. Indeed, an accurate functional description of protein regulation and thus metabolic interactions between bacterial species in simplified microbial communities will allow a better understanding of the natural environment from which key culturable bacteria are isolated.

Synthetic communities are easy to manipulate, and facilitate the production of several replicates in a high-throughput fashion [49,53,55]. Sample preparation for proteomics equals performing proteomics on a "metaorganism." In this way, extracted proteomes are more efficient both in terms of quantity and quality with artificial communities compared to natural communities. This approach allows in vivo metabolic labeling for quantitative metaproteomics application and/or taxa quantification and overcomes the issue of natural compounds present in the environment that might interfere with mass spectrometry.

Because synthetic microbial assemblages are less complex than natural microbial communities, therefore they are generally very well suited for mathematical modeling. This field might also open up new biotechnological applications.

Acknowledgments

T.J. Griffin, P.D. Jagtap and the University of Minnesota Galaxy-P research team are supported by grant 1U24CA199347 from the U.S. NCI Informatics for Cancer Research program, as well as grant 1458524 from the U.S. National Science Foundation. The Galaxy-P infrastructure is maintained by James Johnson, Thomas McGowan and other team members at the Minnesota Supercomputing Institute, Minneapolis, USA. M.B. gratefully acknowledges the financial support of the Belgian Science Policy Office (IUAP contract # P7/25).

References

[1] Großkopf T, Soyer OS. Synthetic microbial communities. Curr Opin Microbiol 2014;18:72–7.
[2] Little AEF, Robinson CJ, Peterson SB, Raffa KF, Handelsman J. Rules of engagement: interspecies interactions that regulate microbial communities. Annu Rev Microbiol 2008;62(1):375–401.
[3] Mayali X, Azam F. Algicidal bacteria in the sea and their impact on Algal Blooms1. J Eukaryot Microbiol 2004;51(2):139–44.
[4] Waldbauer JR, Rodrigue S, Coleman ML, Chisholm SW. Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. Lin S., editor. PLoS One 2012;7(8):e43432.
[5] Allen EE, Banfield JF. Community genomics in microbial ecology and evolution. Nat Rev Microbiol 2005;3(6):489–98.
[6] DeLong EF. Microbial community genomics in the ocean. Nat Rev Microbiol 2005;3(6):459–69.
[7] Kan J, Hanson TE, Ginter JM, Wang K, Chen F. Metaproteomic analysis of Chesapeake Bay microbial communities. Saline Syst 2005;1:7.
[8] Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC, et al. Community proteomics of a natural microbial biofilm. Science (New York, NY) 2005;308(5730):1915–20.
[9] Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, et al. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J 2009;3(1):93–105.
[10] Ng C, DeMaere MZ, Williams TJ, Lauro FM, Raftery M, Gibson JA, et al. Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. ISME J 2010;4(8):1002–19.
[11] Bertin PN, Heinrich-Salmeron A, Pelletier E, Goulhen-Chollet F, Arsène-Ploetze F, Gallien S, et al. Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics. ISME J 2011;5(11):1735–47.
[12] Sowell SM, Abraham PE, Shah M, Verberkmoes NC, Smith DP, Barofsky DF, et al. Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. ISME J 2011;5(5):856–65.
[13] Bao Z, Okubo T, Kubota K, Kasahara Y, Tsurumaru H, Anda M, et al. Metaproteomic identification of diazotrophic methanotrophs and their localization in root tissues of field-grown rice plants. Appl Environ Microbiol 2014;80(16):5043–52.
[14] Bastida F, Hernández T, García C. Metaproteomics of soils from semiarid environment: functional and phylogenetic information obtained with different protein extraction methods. J Proteome 2014;101:31–42.
[15] Dong H-P, Hong Y-G, Lu S, Xie L-Y. Metaproteomics reveals the major microbial players and their biogeochemical functions in a productive coastal system in the northern South China Sea: Shantou coast microbial metaproteomics. Environ Microbiol Rep 2014;6(6):683–95.
[16] Saito MA, Dorsk A, Post AF, McIlvin MR, Rappé MS, DiTullio GR, et al. Needles in the blue sea: sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. Proteomics 2015;15(20):3521–31.
[17] Mattarozzi M, Manfredi M, Montanini B, Gosetti F, Sanangelantoni AM, Marengo E, et al. A metaproteomic approach dissecting major bacterial functions in the rhizosphere of plants living in serpentine soil. Anal Bioanal Chem 2017;409:2327–39.

[18] Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G. Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. ISME J 2010;4(5):673–85.
[19] Mueller RS, Denef VJ, Kalnejais LH, Suttle KB, Thomas BC, Wilmes P, et al. Ecological distribution and population physiology defined by proteomics in a natural microbial community. Mol Syst Biol 2010;6.
[20] Lauro FM, DeMaere MZ, Yau S, Brown MV, Ng C, Wilkins D, et al. An integrative study of a meromictic lake ecosystem in Antarctica. ISME J 2011;5(5):879–95.
[21] Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM, et al. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science 2012;336(6081):608–11.
[22] Williams TJ, Long E, Evans F, DeMaere MZ, Lauro FM, Raftery MJ, et al. A metaproteomic assessment of winter and summer bacterioplankton from Antarctic Peninsula coastal surface waters. ISME J 2012;6(10):1883–900.
[23] Georges AA, El-Swais H, Craig SE, Li WK, Walsh DA. Metaproteomic analysis of a winter to spring succession in coastal northwest Atlantic Ocean microbial plankton. ISME J 2014;8(6):1301–13.
[24] Hawley AK, Brewer HM, Norbeck AD, Paša-Toli L, Hallam SJ. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. Proc Natl Acad Sci 2014;111(31):11395–400.
[25] Leary DH, Li RW, Hamdan LJ, Hervey WJ, Lebedev N, Wang Z, et al. Integrated metagenomic and metaproteomic analyses of marine biofilm communities. Biofouling 2014;30(10):1211–23.
[26] Bastida F, García C, von Bergen M, Moreno JL, Richnow HH, Jehmlich N. Deforestation fosters bacterial diversity and the cyanobacterial community responsible for carbon fixation processes under semiarid climate: a metaproteomics study. Appl Soil Ecol 2015;93:65–7.
[27] Colatriano D, Ramachandran A, Yergeau E, Maranger R, Gélinas Y, Walsh DA. Metaproteomics of aquatic microbial communities in a deep and stratified estuary. Proteomics 2015;15(20):3566–79.
[28] Gillan DC, Roosa S, Kunath B, Billon G, Wattiez R. The long-term adaptation of bacterial communities in metal-contaminated sediments: a metaproteogenomic study: bacteria in metal-contaminated sediments. Environ Microbiol 2015;17(6):1991–2005.
[29] Bryson S, Li Z, Pett-Ridge J, Hettich RL, Mayali X, Pan C, et al. Proteomic stable isotope probing reveals taxonomically distinct patterns in amino acid assimilation by coastal marine bacterioplankton. VerBerkmoes N., editor. mSystems 2016;1(2):e00027-15.
[30] Lacerda CMR, Choe LH, Reardon KF. Metaproteomic analysis of a bacterial community response to cadmium exposure. J Proteome Res 2007;6(3):1145–52.
[31] Russo DA, Couto N, Beckerman AP, Pandhal J. A metaproteomic analysis of the response of a freshwater microbial community under nutrient enrichment. Front Microbiol 2016;7:1172.
[32] Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 1998;95(12):6578–83.
[33] Wang D-Z, Xie Z-X, Zhang S-F. Marine metaproteomics: current status and future directions. J Proteome 2014;97:27–35.
[34] Muddiman D, Andrews G, Lewis D, Notey J, Kelly R. Part II: defining and quantifying individual and co-cultured intracellular proteomes of two thermophilic microorganisms by GeLC-MS2 and spectral counting. Anal Bioanal Chem 2010;398(1):391–404.
[35] Arsène-Ploetze F, Bertin PN, Carapito C. Proteomic tools to decipher microbial community structure and functioning. Environ Sci Pollut Res 2015;22(18):13599–612.
[36] Matallana-Surget S, Leroy B, Wattiez R. Shotgun proteomics: concept, key points and data mining. Expert Rev Proteomics 2010;7(1):5–7.
[37] Wang D-Z, Kong L-F, Li Y-Y, Xie Z-X. Environmental microbial community proteomics: status, challenges and perspectives. Int J Mol Sci 2016;17(8):1275.
[38] Wilmes P, Heintz-Buschart A, Bond PL. A decade of metaproteomics: where we stand and what the future holds. Proteomics 2015;15(20):3409–17.

[39] Bastida F, Moreno JL, Nicolás C, Hernández T, García C. Soil metaproteomics: a review of an emerging environmental science. Significance, methodology and perspectives. Eur J Soil Sci 2009;60(6):845–59.
[40] Neilson KA, Ali NA, Muralidharan S, Mirzaei M, Mariani M, Assadourian G, et al. Less label, more free: approaches in label-free quantitative mass spectrometry. Proteomics 2011;11(4):535–53.
[41] Otto A, Becher D, Schmidt F. Quantitative proteomics in the field of microbiology. Proteomics 2014;14(4–5):547–65.
[42] Podwojski K, Eisenacher M, Kohl M, Turewicz M, Meyer HE, Rahnenführer J, et al. Peek a peak: a glance at statistics for quantitative label-free proteomics. Expert Rev Proteomics 2010;7(2):249–61.
[43] McIlwain S, Mathews M, Bereman MS, Rubel EW, MacCoss MJ, Noble WS. Estimating relative abundances of proteins from shotgun proteomics data. BMC Bioinform 2012;13(1):308.
[44] Matallana-Surget S, Joux F, Wattiez R, Lebaron P. Proteome analysis of the UVB-resistant marine bacterium *Photobacterium angustum* S14. In: Gasset M., editor. PLoS One 2012;7(8):e42299.
[45] Helbling EW, Zagarese H, editors. UV effects in aquatic organisms and ecosystems. Cambridge: Royal Society of Chemistry; 2003.
[46] Andrade-Domínguez A, Salazar E, del Carmen Vargas-Lagunas M, Kolter R, Encarnación S. Eco-evolutionary feedbacks drive species interactions. ISME J 2014;8(5):1041–54.
[47] Kelvin Lee KW, Hoong Yam JK, Mukherjee M, Periasamy S, Steinberg PD, Kjelleberg S, et al. Interspecific diversity reduces and functionally substitutes for intraspecific variation in biofilm communities. ISME J 2016;10(4):846–57.
[48] Lawrence D, Fiegna F, Behrends V, Bundy JG, Phillimore AB, Bell T, et al. Species interactions alter evolutionary responses to a novel environment. In: Ellner S.P., editor. PLoS Biol 2012;10(5):e1001330.
[49] Bell T, Newman JA, Silverman BW, Turner SL, Lilley AK. The contribution of species richness and composition to bacterial services. Nature 2005;436(7054):1157–60.
[50] Gravel D, Bell T, Barbera C, Bouvier T, Pommier T, Venail P, et al. Experimental niche evolution alters the strength of the diversity–productivity relationship. Nature 2011;469(7328):89–92.
[51] Mee MT, Collins JJ, Church GM, Wang HH. Syntrophic exchange in synthetic microbial communities. Proc Natl Acad Sci 2014;111(20):E2149–56.
[52] Pande S, Merker H, Bohl K, Reichelt M, Schuster S, de Figueiredo LF, et al. Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. ISME J 2014;8(5):953–62.
[53] De Roy K, Marzorati M, Negroni A, Thas O, Balloi A, Fava F, et al. Environmental conditions and community evenness determine the outcome of biological invasion. Nat Commun 2013;4:1383.
[54] Lee KWK, Periasamy S, Mukherjee M, Xie C, Kjelleberg S, Rice SA. Biofilm development and enhanced stress resistance of a model, mixed-species community biofilm. ISME J 2014;8(4):894–907.
[55] Wittebolle L, Marzorati M, Clement L, Balloi A, Daffonchio D, Heylen K, et al. Initial community evenness favours functionality under selective stress. Nature 2009;458(7238):623–6.
[56] Giannone RJ, Huber H, Karpinets T, Heimerl T, Küper U, Rachel R, et al. Proteomic characterization of cellular and molecular processes that enable the Nanoarchaeum equitans-Ignicoccus hospitalis relationship. In: Randau L., editor. PLoS One 2011;6(8):e22942.
[57] Pagaling E, Strathdee F, Spears BM, Cates ME, Allen RJ, Free A. Community history affects the predictability of microbial ecosystem development. ISME J 2014;8(1):19–30.
[58] Kerckhof F-M, Courtens ENP, Geirnaert A, Hoefman S, Ho A, Vilchez-Vargas R, et al. Optimized cryopreservation of mixed microbial communities for conserved functionality and diversity. In: McCluskey K., editor. PLoS One 2014;9(6):e99517.
[59] Rubbens P, Props R, Boon N, Waegeman W. Flow cytometric single-cell identification of populations in synthetic bacterial communities. In: Larsen P.E., editor. PLoS One 2017;12(1):e0169754.
[60] Stahl DC, Swiderek KM, Davis MT, Lee TD. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. J Am Soc Mass Spectrom 1996;7(6):532–40.
[61] Chapman JD, Goodlett DR, Masselon CD. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrom Rev 2014;33(6):452–70.

[62] Gillet LC, Navarro P, Tate S, Rost H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 2012;11(6). O111.016717.
[63] Schubert OT, Gillet LC, Collins BC, Navarro P, Rosenberger G, Wolski WE, et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat Protoc 2015;10(3):426–41.
[64] Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, Biosa G, et al. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. In: Martens L., editor. PLoS One 2013;8(12):e82981.
[65] Jagtap P, Goslinga J, Kooren JA, McGowan T, Wroblewski MS, Seymour SL, et al. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. Proteomics 2013;13(8):1352–7.
[66] Tanca A, Palomba A, Fraumene C, Pagnozzi D, Manghina V, Deligios M, et al. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome 2016;4(1):51.
[67] Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, et al. Characterization of the gut microbiome using 16S or shotgun metagenomics. Front Microbiol 2016;7:459.
[68] May DH, Timmins-Schiffman E, Mikan MP, Harvey HR, Borenstein E, Nunn BL, et al. An alignment-free 'metapeptide' strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. J Proteome Res 2016;15(8):2697–705.
[69] Timmins-Schiffman E, May DH, Mikan M, Riffle M, Frazar C, Harvey HR, et al. Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. ISME J 2017;11(2):309–14.
[70] Tang H, Li S, Ye Y. A graph-centric approach for metagenome-guided peptide and protein identification in metaproteomics. PLoS Comput Biol 2016;12(12):e1005224.
[71] Belstrøm D, Jersie-Christensen RR, Lyon D, Damgaard C, Jensen LJ, Holmstrup P, et al. Metaproteomics of saliva identifies human protein markers specific for individuals with periodontitis and dental caries compared to orally healthy controls. PeerJ 2016;4:e2433.
[72] Garcia GD, Santos E de O, Sousa GV, Zingali RB, Thompson CC, Thompson FL. Metaproteomics reveals metabolic transitions between healthy and diseased stony coral Mussismilia braziliensis. Mol Ecol 2016;25(18):4632–44.
[73] Haange S-B, Oberbach A, Schlichting N, Hugenholtz F, Smidt H, von Bergen M, et al. Metaproteome analysis and molecular genetics of rat intestinal microbiota reveals section and localization resolved species distribution and enzymatic functionalities. J Proteome Res 2012;11(11):5406–17.
[74] Jagtap P, McGowan T, Bandhakavi S, Tu ZJ, Seymour S, Griffin TJ, et al. Deep metaproteomic analysis of human salivary supernatant. Proteomics 2012;12(7):992–1001.
[75] Klaassens ES, de Vos WM, Vaughan EE. Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. Appl Environ Microbiol 2007;73(4):1388–92.
[76] Kolmeder CA, de Been M, Nikkilä J, Ritamo I, Mättö J, Valmu L, et al. Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions. PLoS One 2012;7(1):e29913.
[77] Rudney JD, Jagtap PD, Reilly CS, Chen R, Markowski TW, Higgins L, et al. Protein relative abundance patterns associated with sucrose-induced dysbiosis are conserved across taxonomically diverse oral microcosm biofilm models of dental caries. Microbiome 2015;3:69.
[78] Rudney JD, Xie H, Rhodus NL, Ondrey FG, Griffin TJ. A metaproteomic analysis of the human salivary microbiota by three-dimensional peptide fractionation and tandem mass spectrometry. Mol Oral Microbiol 2010;25(1):38–49.
[79] Schaubeck M, Clavel T, Calasan J, Lagkouvardos I, Haange SB, Jehmlich N, et al. Dysbiotic gut microbiota causes transmissible Crohn's disease-like ileitis independent of failure in antimicrobial defence. Gut 2016;65(2):225–37.
[80] Wilmes P, Bond PL. The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. Environ Microbiol 2004;6(9):911–20.

[81] Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One 2012;7(11):e49138.
[82] Fouts DE, Pieper R, Szpakowski S, Pohl H, Knoblach S, Suh M-J, et al. Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. J Transl Med 2012;10:174.
[83] Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, et al. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. Proc Natl Acad Sci 2012;109(19):E1173–82.
[84] Kohrs F, Wolter S, Benndorf D, Heyer R, Hoffmann M, Rapp E, et al. Fractionation of biogas plant sludge material improves metaproteomic characterization to investigate metabolic activity of microbial communities. Proteomics 2015;15(20):3585–9.
[85] Wu J, Zhu J, Yin H, Liu X, An M, Pudlo NA, et al. Development of an integrated pipeline for profiling microbial proteins from mouse fecal samples by LC-MS/MS. J Proteome Res 2016;15(10):3635–42.
[86] Young JC, Pan C, Adams RM, Brooks B, Banfield JF, Morowitz MJ, et al. Metaproteomics reveals functional shifts in microbial and human proteins during a preterm infant gut colonization case. Proteomics 2015;15(20):3463–73.
[87] Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 2007;5(3):e77.
[88] Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI, et al. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS One 2008;3(1):e1456.
[89] Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 2007;5(3):e16.
[90] Muth T, Kolmeder CA, Salojärvi J, Keskitalo S, Varjosalo M, Verdam FJ, et al. Navigating through metaproteomics data: a logbook of database searching. Proteomics 2015;15(20):3439–53.
[91] Eng JK, Searle BC, Clauser KR, Tabb DL. A face in the crowd: recognizing peptides through database search. Mol Cell Proteomics 2011;10(11). R111.009522.
[92] Chatterjee S, Stupp GS, Park SKR, Ducom J-C, Yates JR, Su AI, et al. A comprehensive and scalable database search system for metaproteomics. BMC Genomics 2016;17(1):642.
[93] Wang Y, Ahn T-H, Li Z, Pan C. Sipros/ProRata: a versatile informatics system for quantitative community proteomics. Bioinformatics (Oxford, England) 2013;29(16):2064–5.
[94] Guerrero CR, Jagtap PD, Johnson JE, Griffin TJ. Chapter 13. Using galaxy for proteomics. In: Bessant C, editor. New developments in mass spectrometry. Cambridge: Royal Society of Chemistry; 2016. p. 289–320.
[95] Jagtap PD, Blakely A, Murray K, Stewart S, Kooren J, Johnson JE, et al. Metaproteomic analysis using the Galaxy framework. Proteomics 2015;15(20):3553–65.
[96] Jagtap PD, Johnson JE, Onsongo G, Sadler FW, Murray K, Wang Y, et al. Flexible and accessible workflows for improved proteogenomic analysis using the Galaxy framework. J Proteome Res 2014;13(12):5898–908.
[97] Kertesz-Farkas A, Keich U, Noble WS. Tandem mass spectrum identification via cascaded search. J Proteome Res 2015;14(8):3027–38.
[98] Mesuere B, Willems T, Van der Jeugt F, Devreese B, Vandamme P, Dawyndt P. Unipept web services for metaproteomics analysis. Bioinformatics (Oxford, England) 2016;32(11):1746–8.
[99] Lünsmann V, Kappelmeyer U, Benndorf R, Martinez-Lavanchy PM, Taubert A, Adrian L, et al. In situ protein-SIP highlights Burkholderiaceae as key players degrading toluene by para ring hydroxylation in a constructed wetland model. Environ Microbiol 2016;18(4):1176–86.
[100] Muth T, Behne A, Heyer R, Kohrs F, Benndorf D, Hoffmann M, et al. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. J Proteome Res 2015;14(3):1557–65.

[101] Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 2012;8(6):e1002358.
[102] Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 2013;31(9):814–21.
[103] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012;486(7402):207–14.
[104] Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res 2007;17(3):377–86.
[105] Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 2016;44(W1):W3–10.
[106] Mesuere B, Debyser G, Aerts M, Devreese B, Vandamme P, Dawyndt P. The Unipept metaproteomics analysis pipeline. Proteomics 2015;15(8):1437–42.
[107] Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. Proteomics 2011;11(5):996–9.
[108] Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, Sickmann A, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat Biotechnol 2015;33(1):22–4.
[109] Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. Combining results of multiple search engines in proteomics. Mol Cell Proteomics 2013;12(9):2383–93.
[110] Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, Mitchell A, et al. Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database (Oxford) 2012;2012:bar068.
[111] Konopka A, Lindemann S, Fredrickson J. Dynamics in microbial communities: unraveling mechanisms to identify principles. ISME J 2015;9(7):1488–95.