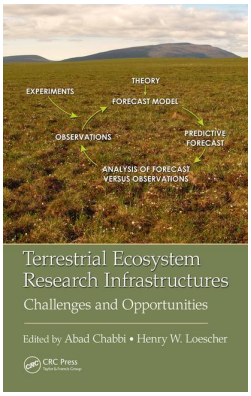

Terrestrial Ecosystem Research Infrastructures Challenges and Opportunities

Abad Chabbi, Henry W. Loescher

Large-Scale Sequence-Based Information

Publication details
https://www.routledgehandbooks.com/doi/10.1201/9781315368252-8
Achim Quaiser, Alexis Dufresne, Sophie Michon-Coudouel, Marine Biget, Philippe Vandenkoornhuyse
Published online on: 22 Feb 2017

7

Large-Scale Sequence-Based Information: Novel Understanding of Ecology and Novel Avenues to Test Ecological Hypotheses

Achim Quaiser, Alexis Dufresne, Sophie Michon-Coudouel, Marine Biget, and Philippe Vandenkoornhuyse

CONTENTS

Abstract

Ecology entered a new era with growing evidences of the huge impact of microorganisms on all existing ecosystems. Molecular approaches, such as the SSU rRNA gene analysis and environmental genomic approaches, identified an unexpected microbial taxonomic and functional diversity in all kinds

of habitats. This diversity and the low cultivability suggest functional interactions within microbial communities as well as between microbial communities and plants or animals in dependence of the environmental conditions. These findings shape the groundwork of new revolutionary ecological concepts away from isolated characteristics of organisms to broad *holistic**/ integrated views. Aiming to capture and to understand the functioning of interacting microbial communities, they must be considered as a whole, in constant evolution and in adaptation to environmental factors. Since there is no all-inclusive approach combining all targeted aspects of analysis, results obtained from particular molecular ecology approaches are still rather tesserae. Nevertheless, combining different approaches fills the mosaic leading to more comprehensive reproductions of in situ community functions.

In this chapter, we highlight the findings and limitations of environmental genomics that paved the way for new *holistic** concepts of *community assemblages** as well as perspectives. Combined with insights into molecular techniques, the reader should be well fitted to apply and develop adequate research strategies that imply environmental genomic approaches combined with high-throughput sequencing.

Keywords: Metagenomics, Metatranscriptomics, Metabarcoding, Taxonomic and functional diversity, Culture-independent ecology

7.1 Use of Sequence Data in Ecology: An Overview

The major aims of ecology are the study and understanding of the processes that foster biodiversity and that biodiversity contributes toward the understanding of ecosystem functioning. Biodiversity is defined as the species assemblage within a community, the genetic variability within the species, and the diversity of ecosystems formed by these species assemblages (Wilson 1989). Thus, the species level does not intrinsically summarize the biodiversity but corresponds to one key level. However, defining the species level is not straightforward. It depends on the organism observed and the definition of species that is adopted. A good *species concept** allows to group *evolutionary-related organisms**. This systematic view of life based on evolutionary relationships among organisms allows to formally describe diversity that can lead to modifications of taxonomy (e.g., the delineation a third domain of life, the Archaea, besides Bacteria and Eukarya). This systematic view is possible from the analysis of nucleic acid sequences, DNA and RNA, because these vertically inherited sequences (i.e., parents to offspring transmission) have recorded traces of evolution (i.e., mutation, insertion, deletion). Besides

* Definition of terms in *italic* with the sign "***": are found in the glossary.

the taxonomic criteria, which could be sometimes subjective, this *diachronic information**, contained in DNA and RNA, is a fundamental material to detect and classify the organisms and microorganisms, known or unknown. The use of DNA-/RNA-based methods has revolutionized our understanding of the extent of the diversity especially considering microorganisms.

Microorganisms are unicellular, and due to their small size, they are difficult to observe and describe. Microorganisms belong to the three domains of life: Eukarya, Bacteria, and Archaea. For a long time, microbiology was more regarded as a discipline of medical interest. Often, microbiologists analyzed one single cultured species from different characteristics such as phenotype, physiology, metabolism, and genetics. This culture-based approach is disconnected from the natural world and selects only a tiny fraction of the microorganisms. The development of molecular ecology approaches has changed the perception of the diversity of organisms and revealed that in the microbial world the cultivated and known microorganisms represent only a drop in an ocean of diversity. For example, 1 kg of soil contains about 10^{13} microorganisms in average (Whitman et al. 1998), thus, about a number 100 times higher than the number of stars in the Milky Way (Curtis and Sloan 2005), while waters (oceans + continental water) contain in total about $1.2{\cdot}10^{29}$ prokaryote cells (Whitman et al. 1998) and the human gut contains 100 times more unique genes than the human genome (Qin et al. 2010). This huge diversity of microorganisms plays a key role in ecosystem functioning and services (e.g., Vandenkoornhuyse et al. 2010).

Considering the historic background, one milestone was the development of culture-independent technologies, notably the SSU rRNA gene approaches allowing the detection and classification of uncultured microorganisms (Pace et al. 1986; Figure 7.1). The SSU rRNA gene is amplified from extracted DNA from environmental samples, cloned in an adequate *vector**, amplified in *Escherichia coli* cells, sequenced and classified by phylogenetic analysis (Figure 7.2). With this approach, the quantities of available SSU rRNA sequences, deposited in dedicated databases, were growing. The quality of the phylogenetic signal contained within the SSU rRNA gene has allowed its use as *molecular clock**, starting with the pioneering work of Woese (1987). Cloning, sequencing, and analysis of bulk genomic DNA without targeted amplification was applied shortly thereafter (Schmidt et al. 1991; Figure 7.1). Subsequently, the analysis of the first cloned and sequenced large genomic fragments (>35 kb) from marine plankton microbial assemblages further allowed to link taxonomic classification and functional potential together (Stein et al. 1996). Cultivation-independent surveys revealed that in natural ecosystems at most about 0.1% of the microbial species could be cultivated. Today, the SSU rRNA sequences generated by cultivation-independent molecular approaches combined with new sequencing technologies lead to the exponential growth of dedicated databases reaching more than five billions of SSU rRNA sequences in 2015 (e.g., www.arb-silva.de).

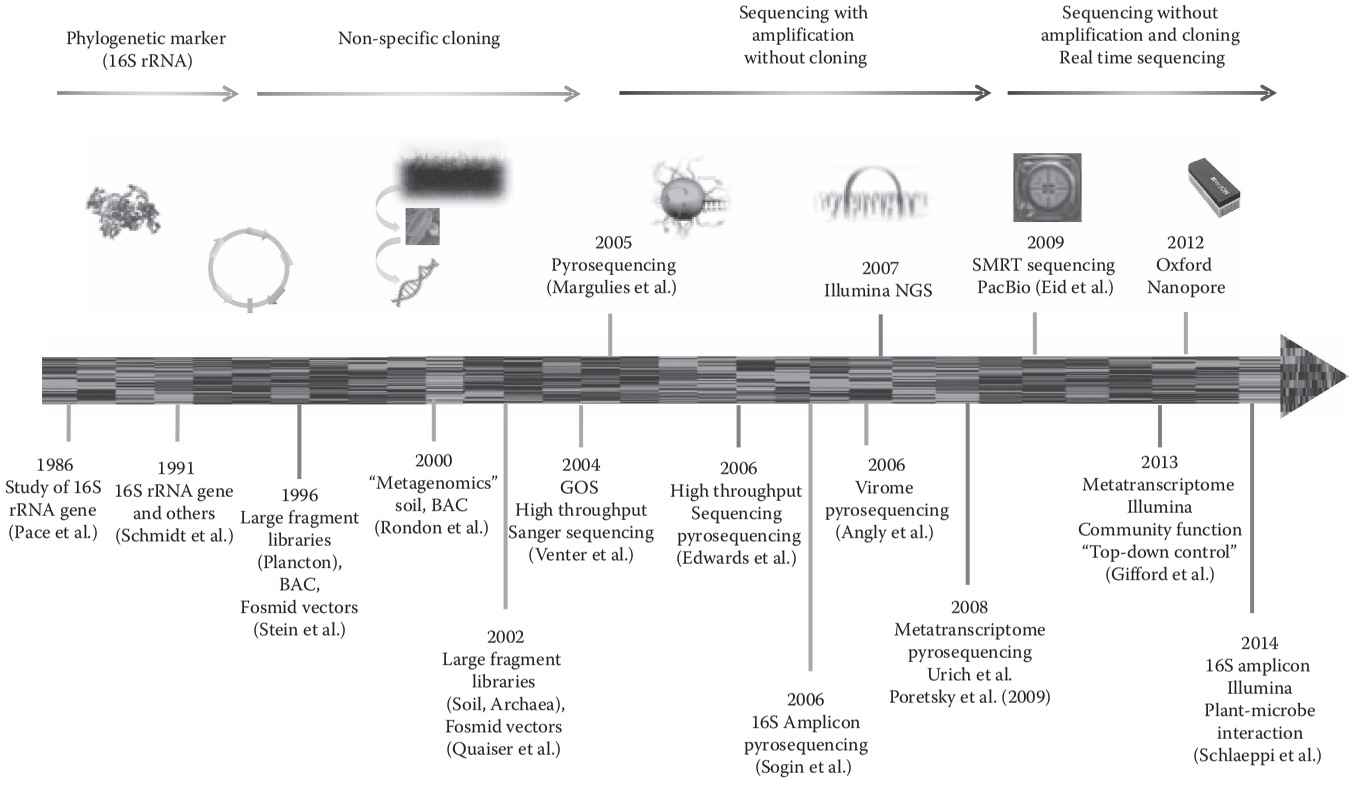

FIGURE 7.1
Historical overview of major advances in molecular ecology.

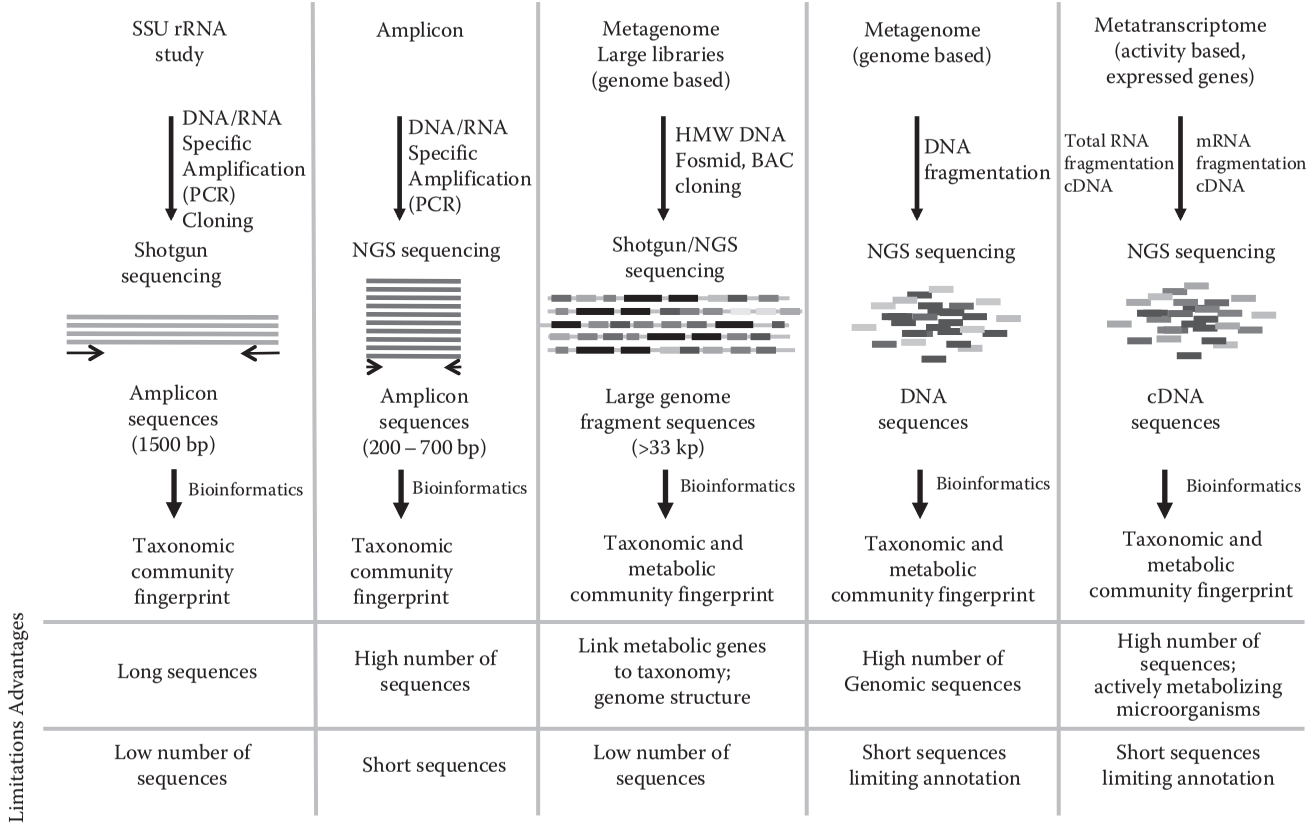

FIGURE 7.2
Overview of the different nucleic acid approaches.

Molecular ecology has entered the genomic era (e.g., Vandenkoornhuyse et al. 2010). New understanding of evolution from genomic comparisons is an important field of research supported by the increasing number of available genomes (i.e., in October 2015, 42,323 Bacteria, 997 Archaea, and 6,572 Eukarya genomes sequenced, https://gold.jgi.doe.gov). Environmental genomics currently drives important discoveries in biology and a cortege of new ideas, concepts, and theories (Vandenkoornhuyse et al. 2010). Developments of genomic and metagenomic approaches to microorganisms and microbial communities have demonstrated a high number of *horizontal gene transfers** even among evolutionarily distant microbial species, allowing the acquisition of new physiological functions (Nelson et al. 1999; Quaiser et al. 2003; Gladyshev et al. 2008; Keeling and Palmer 2008; Deschamps et al. 2014). In addition, viruses as mobile genetic elements involved in HGT were identified as the most abundant biological entity and are equipped with fast evolving genetic diversity (Fuhrman 1999; Stern and Sorek 2011). These attributes combined with the low cultivability of prokaryotes inevitably lead to new views of microbial *community assemblages** and evolutionary trajectories (Woese 2004; Goldenfeld and Woese 2007).

Current new frontiers in ecology are related to the recent knowledge and understanding of microorganisms in the environment. We underline urgent research questions considering the understanding of the complexity of the microbial communities and associated ecological functions. It is essential to understand the rules of microbial *community assembly** and the associated theories to describe these rules since the selection and genetic drift impact individual microorganisms, population and community, and expressed function.

The focus of this chapter is thus to provide an overview of ideas and frontiers together with past, current, and upcoming strategies related to environmental genomics.

7.2 Mass Sequencing: Current and Upcoming Technologies

The first sequencing technology was developed by Sanger (Sanger and Coulson 1975) and in parallel by Maxam and Gilbert using a more complex technology (Maxam and Gilbert 1977). Starting with fluorescence-based labeling of the four nucleotides and manual analysis of radiographies, the first automated sequencer was introduced in 1987 using Sanger technology. This technology was scaled up for the sequencing of the first human genome (Collins et al. 2003) costing ~US$2.7 billion. This was followed by second-generation sequencing using the sequencing-by-synthesis technology, thereby avoiding the cloning steps (Schuster 2008). Pyrosequencing improved the throughput over Sanger sequencing approaches (Margulies et al. 2005) with >99% accuracy and much lower cost per base pair. Complete microbial genomes could

be sequenced and assembled in several hours with sequence *coverage** allowing the assembly of the complete genomes. Due to the continuous development and improvement of pyrosequencing technology, the sequence length reached up to 700 base pairs in average and up to 1.5 million sequences per run (i.e., 454 GS flx Plus instrument, Roche, Switzerland). The current dominant technology of mass sequencing can produce in a single run 25 million reads of up to 300 bp per read (MiSeq instrument, Illumina Inc., United States). With HiSeq X series instrument (Illumina Inc., United States), even more spectacular numbers are obtained, reaching up to six billion sequences per run and 2 × 150 base pairs per read. Besides these methods, sequencing where nucleotides are read in real time without the bias-prone amplification step during sequencing is currently being developed (i.e., third-generation sequencing [TGS]) aiming to perform single-molecule sequencing (see Section 7.6.1). New technologies are currently developed. For example, the newest TGS technology targets the abilities to increase read length simplifying subsequent sequence data analyses. Since significantly lower number of reads and lower *coverage** is necessary, this goes along with reduced analytical computation and storage costs (see section in the following text for more detail). Among these methods, three technologies can be underlined: (1) single-molecule synthesis by DNA polymerase with real-time recording (Sequel instrument by Pacific Bioscience, United States); (2) nanopore technology, in which the nucleotides are identified by the modification of electric potential during the transfer of a single DNA molecule through a nanopore (minion by Oxford Nanopore Technologies, United Kingdom); and (3) direct imaging of labeled DNA molecules in real time (ongoing developments by ZS Genetics Inc., United States). These TGS developments have all the same focus, long sequences, and high accuracy at low cost.

The current dominant sequencing technologies allow only producing short sequences. These short stretches of data form a giant puzzle to reassemble. The reconstruction of complete genomes from these sequences necessitates strong computational and laboratory efforts. Major difficulties to assemble a full genome from short reads are related to occurring sequence repeats and natural genetic variations (*haplotypic variations**). The generation of long sequences, as targeted by TGS technologies, allows reducing these difficulties and thereby reducing the number of sequences needed. In other words, if the sequencing length would not be limited, only one read would be necessary to complete one *haploid** genome or chromosome.

7.3 Diversity, Molecular Barcoding, and New Prospects

Prior to addressing communities as a whole, the first key issue is to properly describe the organismal inventory (diversity) (Figure 7.3). For detailed

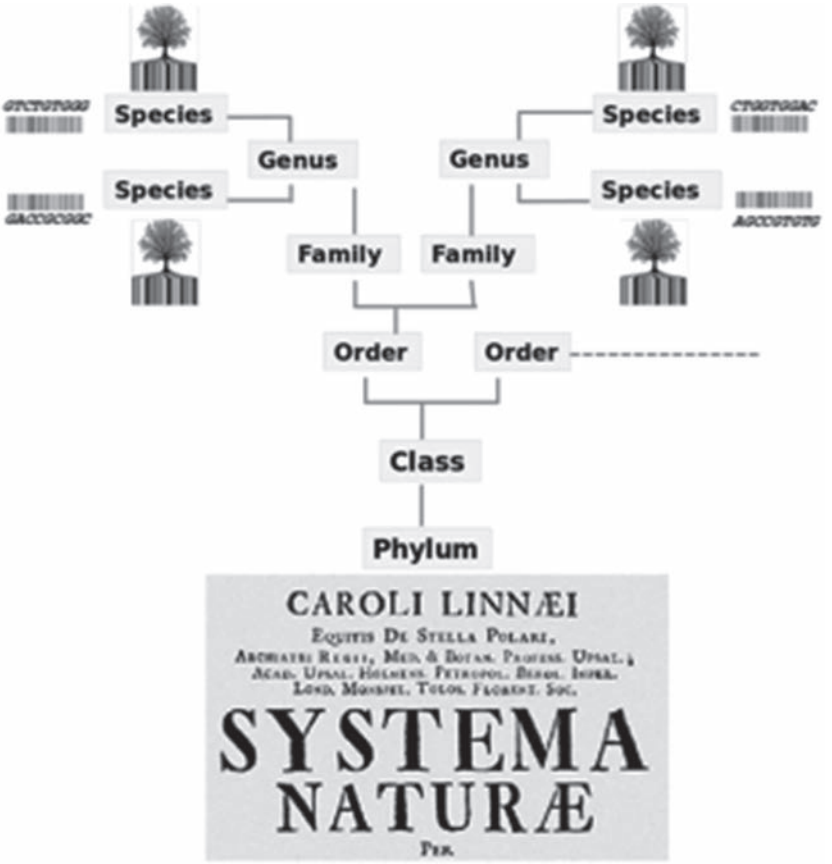

FIGURE 7.3
The hierarchy of biological classification by Carl Linneaus, and related data, metadata, and environmental factors to consider in a holistic perception of "systems ecology." The bar coding of life is based upon this taxonomic diversity.

distinction of the organisms and to describe evolutionary relationships, molecular identification of organisms has been developed by so-called molecular barcoding approaches. This important focus of current research is developed in the following text.

7.3.1 Molecular Barcoding of Plants and Animals

The DNA barcoding aims at identifying any animal or plant (i.e., multicellular organisms) using a short genetic sequence (i.e., the molecular barcode). Thus, molecular barcoding implicitly accept the existing Linnaean taxonomy (Figure 7.3). Molecular barcoding was developed with the broad rationale of (1) identification of all species, (2) rapid identification of species, (3) linkage

of the species to environmental conditions, (4) discovery of new species, and (5) increasing the use and utility of biological collections. For molecular comparison and distinction between organisms, adequate gene or sequence targets must be determined. To be considered as adequate, these DNA fragments should contain conserved regions and enough genetic variations to distinguish species. A consensual DNA barcoding region for animals, the mitochondrial Cytochrome Oxidase 1 (CO1) gene, seems to fit well with the aims of molecular barcoding. Conversely, the selection of a good DNA barcode was more challenging for plants. In this quest for a universal plant DNA barcode, two short coding regions of *plastid DNA**, *rbcLa* and *matK*, are used most often, while other combinations of molecular barcode, such as *trnH-psbA*, are successful in some cases (Ghahramanzadeh et al. 2013). A key difficulty is that there are relatively few nucleotides that are detectable that can determine robust differences among similar or related species (Chase and Fay 2009). It is known that the gene fragments of CO1, *rbc*L and *mat*K, contain an important amount of homoplasy (i.e., inherited homology due to the high rate of mutation within these genomic regions); thus, the quality of the phylogenetic information is moderate. Despite this fact, the successful development of the barcoding of life must be underlined (iBol and related BOLD database URL http://www.ibol.org/). The use of molecular barcoding has brought new depth to our understanding of traditional plant and animal systematic taxonomy and suggests insights that we otherwise would not have obtained. Considering the well-known example of two existing elephant species (i.e., African and Indian), DNA barcoding has allowed to reveal the likely coexistence of three African species (Blaxter 2003). In another example, invasive alien plant species threaten the structure and function of aquatic ecosystems. Barcoding tools have been able to distinguish these invasive plants, which is of particular importance when managing the trade of aquatic plants (Ghahramanzadeh et al. 2013). The DNA barcoding tools are also used to manage the conservation of animal species, such as the large and illegal influx of snapper turtle (*Macroclemys temminckii*) in the food chain (Roman and Bowen 2000). As demonstrated by these examples, robust barcoding tools are continually being developed to further diagnose species in the environment from DNA traces. From an applied aspect, this knowledge and associated tools are developing to better manage illegal use, trade, and will likely contribute to the conservation and preservation of species.

7.3.2 Microorganisms and Molecular "Barcoding"

For microorganisms, detection and taxonomy-based community descriptions within an environmental sample mostly rely on the analyses of the SSU rRNA gene fragment from an *amplicon** mass sequencing approach applied on all kind of ecosystems (e.g., Sogin et al. 2006; Lundberg et al. 2012; Ben Maamar et al. 2015; de Vargas et al. 2015; Nunes et al. 2015) (Figure 7.2). Other options are the recruitment of SSU rRNA sequences from

metatranscriptomes bypassing the bias-prone amplification step (e.g., Urich et al. 2008; Quaiser et al. 2014). From these SSU rRNA sequence analyses, a number of unknown phyla in Bacteria, Archaea, and Eukarya were revealed, and a large proportion of them are still without cultivated representatives (e.g., López-García and Moreira 2008). These SSU rRNA approaches can be assimilated to barcoding or metabarcoding methods, since the sequence analysis tends to the detection and identification of the microorganisms. It must be underlined that one key difference in comparison to plant or animal barcoding approaches is that the analyses of the SSU RNA gene sequences of microorganisms are not necessarily driven by the known taxonomy allowing the detection of undetermined lineages. The use of mass sequencing strategies to study SSU rRNA gene fragments from environmental samples has allowed major advances in the knowledge and understanding of the gut microbiota (e.g., Turnbaugh et al. 2009), the soil microbiota (e.g., Fierer et al. 2012), the plant microbiota (e.g., Bulgarelli et al. 2012; Lundberg et al. 2012), and the pico-eukaryotes in oceans (e.g., Rodríguez-Martínez et al. 2013; de Vargas et al. 2015).

Since the SSU rRNA gene sequence is highly conserved, albeit much less than eukaryotic barcoding targets, one of the limits of SSU rRNA *amplicon** approaches is the resolution level especially when using short sequence fragments. On the other hand, in contrast to PCR-independent approaches such as metagenomic and metatranscriptomic on bulk nucleic acids, *amplicon** approaches necessitate previous knowledge about the targets inevitably discriminating taxonomic lineages.

7.4 Meta-Omics to Assess the Different Dimensions of Diversity

Here, we define meta-omics as the culture-independent analyses beyond the SSU rRNA gene–based analyses of microbial assemblages. These meta-omics approaches are of different types. For a given environmental sample, the metagenomics use the extracted DNAs, while the metatranscriptomics focus on the extracted RNAs and multi-omics approaches aim to combine metagenomics, metatranscriptomics, and metaproteomics. In this section, we describe these different strategies and highlight the advantages and disadvantages.

7.4.1 Metagenomics

Environmental genomics (metagenomics) developed from the adoption of genomic techniques applied to complex microbial communities from

natural ecosystems. Typically, microbial sequencing studies targeted complete genomes, while in environmental genomics, due to the large diversity of microbial communities, this was not feasible. Large genomic fragments, above 32 kb, from marine (Stein et al. 1996; Béjà et al. 2000, 2001) and from soil (Rondon et al. 2000; Quaiser et al. 2002, 2003; Liles et al. 2003; Treusch et al. 2004) microbial communities were cloned, sequenced, and analyzed (Figure 7.2). Rondon et al. (2000) fixed the term "metagenome" for environmental genomic studies. These studies allowed getting first insights into the genomes from uncultured microorganisms. The large genomic fragments revealed the gene content and genome structure as well as the potential function mostly linked to the taxonomic origin contributing significantly to the understanding of microorganisms in the environment. For example, Beja et al. (2000) identified a novel type of bacterial *rhodopsin**, named *proteorhodopsin**, which has been shown to be a very abundant marine plankton (Béjà et al. 2001), and highlights an important previously unknown phototrophic ocean process. This functional identification has helped subsequent isolation and genome sequencing of phototrophic microorganisms from different bacterial phyla (Fuchs et al. 2007; Gómez-Consarnau et al. 2007). Analogical examples of fundamental functions discovered from a metagenomic approach also exist for soils. For instance, the identification of genes encoding for the key enzyme for *nitrification**, the ammonium monooxygenase on an archaeal genome (Treusch et al. 2005; Leininger et al. 2006) suggested key roles of *mesophilic** archaea in nitrogen cycle in soils. Combined with the detection of ammonium monooxygenases in marine plankton samples (Venter et al. 2004) and large genome fragment analyses of *Cenarchaeum symbiosum* (Hallam et al. 2006), mesophilic Archaea (Thaumarchaeota) are to date recognized as keystone players of the *nitrification** within the nitrogen cycle previously thought to be performed exclusively by a restricted number of Bacteria. These two prime examples revealed the strength of environmental genomics. Isolation or enrichment cultures can demonstrate direct information about phenotype, metabolism, physiology, and genomics, while culture-independent approaches can show the characteristics within the environment, such as distribution, diversity, dynamics, and interactions. These "classical" metagenomic approaches permitted also to get first insights into the genomes from species of other major phyla, such as Acidobacteria (Liles et al. 2003; Quaiser et al. 2003, 2008; Kielak et al. 2010), Poribacteria (Fieseler et al. 2006), and marine Thaumarchaeota (Martín-Cuadrado et al. 2008). Interestingly, evolutionary information provided findings about *horizontal gene transfers** among different groups, for example, between Alphaproteobacteria and Acidobacteria (Quaiser et al. 2003), and most impressively massive *horizontal gene transfers**, up to 29.7%, from Bacteria to marine Thaumarchaeota and Euryarchaeota (Deschamps et al. 2014).

The environmental DNA sequencing or metagenomic shotgun sequencing is a powerful method to analyze the microbial diversity. Shotgun *fosmid**

end sequencing was applied to reveal stratified microbial assemblages in the ocean (DeLong et al. 2006) or to explore the diversity and metabolic potential of soil communities (Treusch et al. 2004). *Fosmid** sequencing approaches (Sanger sequencing) provided, at this time, an affordable method to produce relatively high number (5,000–10,000 sequences in these examples) of environmental sequences representing a DNA fingerprint of the microbial community.

To the contrary of large genome fragment analysis, one of the major challenges applying high-throughput sequencing, which goes along with short sequences, is to link the diversity or taxonomic origin with function. The first study applying mass sequencing (high-throughput Sanger sequencing) was the Global Ocean Sequencing Project (Venter et al. 2004; Rusch et al. 2007; Yooseph et al. 2007) (Figure 7.1). This sequencing effort (about 700 Mbp), at the time, doubled the number of existing nucleotide and protein databases (NCBI), allowing the study of the genetic potential of surface ocean water. Several follow-up studies revealed interesting results about the ribotype diversity (Rusch et al. 2007), protein families (Yooseph et al. 2007), prokaryotic genomes (Biers et al. 2009), and others. Analogously, the comparison of whale fall (3 × 25 Mbp) and soil microbial communities (100 Mbp) (Tringe et al. 2005), applying cloning and short fragment Sanger sequencing, revealed habitat-specific fingerprints and, due to the complexity of the microbial communities, evidence for the impossibility to assemble individual genomes. Only about 1% of the reads had overlaps with other reads.

In contrast, the application of the same approach combined with *fluorescence in situ hybridization** was applied on a low-complexity microbial community from an acid mine drainage biofilm (Tyson et al. 2004). Interestingly, with about the same sequencing effort (76 Mbp), several near-complete draft genomes could be reconstituted, and based on the gene content, a metabolic map and pathways were proposed. A recent review details draft or complete genome reconstructions from metagenomic analyses (Gasc et al. 2015).

Higher number of sequences at lower cost can be obtained applying next-generation mass sequencing on bulk DNA. In contrast to gene-targeted *amplicon** approaches described earlier (i.e., Section 7.3), no prior knowledge about the target is necessary and potential *primer** biases are avoided. Otherwise, the bioinformatics analysis is more complex targeting in general numerous genes or gene families, while homologs of the majority of the sequences are missing. Several studies yielded important information about the metabolic potential of microbial communities representing unique fingerprints useful to highlight particular differences (e.g., Ghai et al. 2010; Mackelprang et al. 2011; Quaiser et al. 2011; Shi et al. 2011). For example, recent metagenomic analyses addressing the human microbiota have revealed the role of the microbiome in health and disease (Korem et al. 2015). Assuming that patterns of sequencing read *coverage** in metagenomic samples reflects the microbial growth rate within the microbiota, Korem et al. (2015) used the ratio of copy number to show differences between

virulent and avirulent strain dynamics and correlations to bowel disease. Gut microbiome datasets generated by next-generation sequencing are growing, allowing a comprehensive view of the functional diversity in humans, animals, and invertebrates (Yoon et al. 2015). These metagenomic analyses helped to get insights into the *pleiotropic** roles of the gut microbiome in the nutritional processing, the microbial metabolisms and interactions, the modulation of the host physiology, and the impact of the microbiota on the host defense system (e.g., Doré and Blottière 2015; Yoon et al. 2015). In addition, the likely mutualistic functions of the microbiota emerged and opened new perspectives for the understanding of metazoa along with plants as meta-organisms (Vandenkoornhuyse et al. 2015). Despite this growing understanding and knowledge, and considering the current high research effort to analyze the gut microbiota, an obvious limit of the metagenomic approach is that a large proportion of *open reading frames** are cataloged as uncharacterized and/or novel gene with unknown function (e.g., Qin et al. 2010; Yoon et al. 2015). Thus, the use of the metagenomic approaches has a lot of advantages revealing important discoveries that lead to modification of paradigms (i.e., discovery of archaeal *nitrification**). Nevertheless, it is clear that the more the metagenome is complex, the more the metagenomic analyses are descriptive. In this case, large parts of the metagenome, due to the huge microbial diversity, are not interpretable because a high number of genes and proteins are still unknown and uncharacterized, and in consequence no homologs are deposited in international databases. Altogether, all the genomes sequenced and all the sequences yet produced represent an estimation of about only 1×10^{-19}% of the total DNA sequenced, the ongoing "earth microbiome project" being included in this estimation (Microbiology by Numbers 2011).

7.4.2 Metatranscriptomics

Metatranscriptomics targets the expressed genes in a microbial community at the time of sampling. Compared to metagenomics, this approach targets only active modules from intact cells of the microbial community. Since the half-life of mRNA transcripts in general is short (i.e., seconds to minutes, in general), there is no doubt that these sequences are derived from actively metabolizing cells. Otherwise, gene expression adapts rapidly to changing environmental conditions, requiring rapid and elaborated sampling strategies, in order to fix cells as fast as possible assuring the cells' in situ conditions (Feike et al. 2011). In environmental metatranscriptomic studies, the ribosomal RNAs represent about 80%–97% of the total transcripts, while messenger RNAs count only for up to 20% depending on the ecosystem analyzed (He et al. 2010; Quaiser et al. 2014; Tveit et al. 2014). Therefore, new approaches for mRNA enrichment, through mRNA amplification and by rRNA depletion, were developed (Carvalhais et al. 2012). Other studies favor the extraction and analysis of total RNA, thereby preserving the possibility

to exploit the full potential of the metatranscriptome sequences that include mRNA and rRNA from all three kingdoms (e.g., Urich et al. 2008, 2014; Radax et al. 2012; Quaiser et al. 2014). The first metatranscriptome study analyzed freshwater bacterioplankton communities by mRNA enrichment, cloning of reverse-transcribed RNA, Sanger sequencing, and linking the transcripts to environmental processes (Poretsky et al. 2005). Even though the *sequencing depth** was limited, the first signature transcripts of an environmental community were obtained. In another study, pyrosequencing of metatranscriptome and metagenome from ocean surface water community revealed gene- and taxon-specific expression patterns and surprisingly a high number of unique sequences in the metatranscriptome absent in the metagenome were identified (Frias-Lopez et al. 2008).

The full potential of the metatranscriptomic approach was demonstrated on different ecosystems such as marine sponges (Radax et al. 2012), marine sediments (Urich et al. 2014), peat soils (Tveit et al. 2012, 2014), and iron-rich microbial mats (Quaiser et al. 2014). The stratification of an iron-rich microbial mat could be shown based on taxonomic relevant rRNA as well as mRNA analysis, highlighting the strong expression level of iron oxidizers affiliated to *Leptothrix* species particularly near the surface and surprisingly type I *methanotrophs** in deeper layers (Quaiser et al. 2014). The presence of *methanotrophs** coincided with about 10–100 times higher methane concentrations within the microbial mat compared to surrounding waters. The *methanotrophs*' key enzyme coincided with the pattern of rRNA sequences of *methanotrophs*,* confirming the stratification of the taxonomic analysis and linking diversity to function. Comparison of surface marine plankton metatranscriptomes over the seasons using Illumina sequencing suggested strong top-down ecological control for some faster-growing microbial groups (Gifford et al. 2013).

For the analysis of community diversity and functioning, several observations from metatranscriptome studies can be highlighted, especially as follows: (1) mRNA pools provide precious information about ongoing microbial community processes but only for the time of sampling, (2) replication and comparative metatranscriptome analysis are essential for emerging reliable results, (3) metatranscriptome analysis represents a highly sensitive tool to display the instantaneous metabolic state of the microbial community, (4) metatranscriptomes based on total RNA extraction reduces methodological biases but can limit the functional information included in mRNA, and (5) total RNA metatranscriptomics has a strong potential to link diversity to function including representatives from the three domains. However, the same limit as for metagenomic analyses also exists for metatranscriptomic approaches related to the high number of detected genes without homologs and with unknown function in databases.

The major advantage of metagenomic and metatranscriptomic analyses is the possibility to address the different levels of the ecological hierarchy at the same time (i.e., individual, population, community, ecosystem; Figure 7.4) as

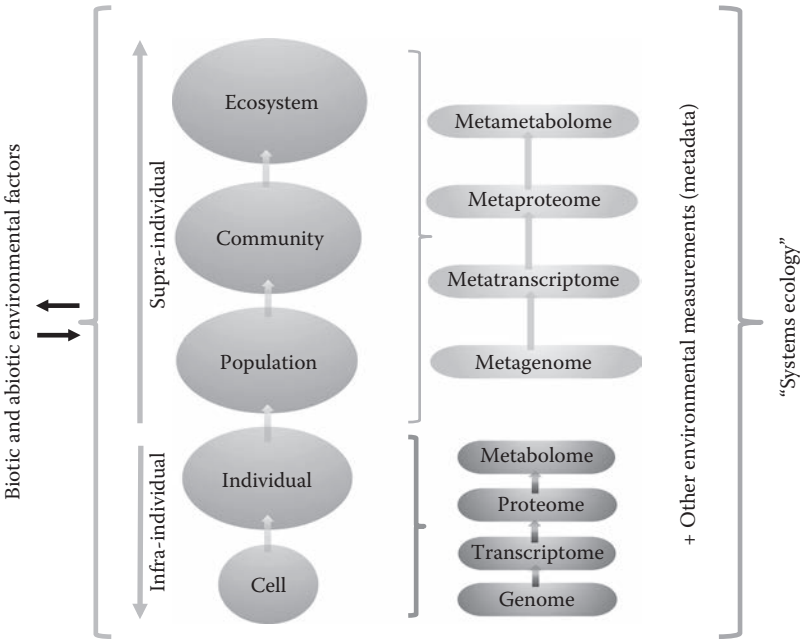

FIGURE 7.4
The ecological hierarchy and related data. Metadata and environmental factors to consider in a holistic perception of "systems ecology."

exemplified recently by analyzing the community and the population level of an iron-rich microbial mat (Quaiser et al. 2014). However, most studies published so far focus on the individual and/or the community level while the population level is omitted. This is certainly related to the *species concept** used for microorganisms but also to the *sequencing depth** and rate of sequencing errors.

7.4.3 Multi-Omics

It is currently possible to integrate heterogeneous data sources including (meta)genomic, (meta)transcriptomic, (meta)proteomic, and metabolomic to provide a *holistic* top-down overview of an organism or an environmental sample. This data cascade could allow deciphering the thin functioning of a tissue, an organism, or an environmental sample. This kind of strategy is still challenging but is important to model or validate in silico biological systems (Yoon et al. 2012) or "ecological system" (see Section 7.6).

The "multi-omics profiling" refers specifically to an expanding multi-omics resource that supports transcriptomics and proteomics information from publicly available studies on model organisms. This developing tool

aims to improve proteomic data analysis from RNA sequence annotation predictions for translated RNAs. We can anticipate that this "multi-omics profiling" will also be developed for non-model organisms to be useable for metaproteomic analyses especially with the development of the next-generation proteomic instruments.

Up to date, multi-omics approaches have intended to identify expressed genes (transcriptomic approach) to understand the genome functioning of a cell/tissue/organism under different life constraints in a community. This multi-omics approach has also been viewed from a combination of metagenomic–metatranscriptomic approaches in an aim to understand the metagenome functioning both (1) at the gene level and (2) at the individual level by differentiating present versus active cells within the analyzed environmental sample.

Compared to mRNA that have short half-time lives and a high variability in copy numbers, the use of (meta)proteomics is promising as proteins are more stable and more abundant than transcripts (Moran et al. 2013). Nevertheless, the identification of peptides revealed by mass spectrometry is difficult, since the variability of protein sequences is very high and the direct comparison of peptides to general protein databases inefficient, since the large majority of observed peptides do not match to the databases. A combined approach, metagenomics, metatranscriptomics, and metaproteomics, was applied on deep-sea hydrothermal vent sediments (Urich et al. 2014), which proved effective to detect the presence and estimated the abundance of *methanotrophs* affiliated to *Methylobacter* sp. at the level of the three molecules, rRNA, mRNA, and protein, thereby clearly linking the diversity to function (i.e., *Methylobacter* species perform aerobic *methanotrophy** in hydrothermal vents). A similar multi-omics approach was applied to the rhizosphere and phyllosphere of rice plants linking *methanogenesis** and *methanotrophy** to particular microbial genera (Knief et al. 2012). A more recent work questioning the carbon storage/emission from permafrost at different states of thaw elegantly combined metagenomics, metatranscriptomics, and metaproteomics approaches to address the phylogenetic composition of the microbial communities and their potential functions and activity (Hultman et al. 2015). The study showed clear linkages between metagenomics/metatranscriptomics data and functioning. In more details, the work provided a new understanding into adaptation strategies for life in frozen conditions, and actors and conditions for *methanogenesis** (Hultman et al. 2015) in relation to the question of permafrost carbon–climate feedback and consequences of global warming (Koven et al. 2011).

7.4.4 Single-Cell Genomics

Single-cell genomic (SCG) approaches proved to be a powerful tool to access genomes from uncultured microorganisms. A high number of complete or

draft genomes even from low-abundance microorganisms ("microbial dark matter") were obtained from all kind of habitats (for an overview, see Gasc et al. 2015). For example, the draft genome of the candidate phylum TM7 was reconstructed from the human mouth microbial community (Marcy et al. 2007) and from soil (Podar et al. 2007) revealing first insights into the metabolic potential and structure of representatives of this candidate phylum. The combination of SCG with other -omics tools applied on the same samples can have strong impacts on the understanding of the microbial community diversity and functioning. Combined with metagenomics and metatranscriptomics, draft genomes can be confirmed or completed as shown recently, where 35 draft genomes from 15 phyla were generated from a methanogenic bioreactor community (Nobu et al. 2015). Recently, more than 200 uncultivated archaeal and bacterial single-cell genomes were generated from nine diverse marine habitats. In this study, 29 major taxonomical lineages of the microbial dark matter could be enriched with genomic sequences (Rinke et al. 2013). In both cases, concrete community functions and interactions were deduced applying metagenome or metatranscriptome recruitment. For further details about the potential of SCG, see Gasc et al. (2015).

7.5 Linking Diversity to Functions

To determine the mechanistic links and combined theory between species diversity and community/ecosystem function is still an ecological grand challenge. This is particularly true for the most numerous fraction of life on the planet, the microorganisms, and is an active area of research. Metagenomic analyses can allow linking the predicted function to a given organism if sequence assembly into large scaffolds is successful (cf. Section 7.4). However, the within-sample microbial diversity is often too high to generate good-quality assemblages of sequences that allow deducing functions when applying bulk sequencing of microbial community DNA or RNA (*cDNA**). In other words, the signals hinting to community functions are too low or hidden by the overwhelming majority of other sequences. Alternative methods to focus on a specific fraction of the diversity involved in a specific process exist. Stable isotope probing (SIP) represents a powerful technique for targeting species involved in specific ecological functions. Typically, from a ^{13}C-enriched substrate, the microorganism able to metabolize the labeled molecule will integrate the ^{13}C leading to the enrichment in ^{13}C of its lipids, proteins, RNAs, and DNAs. The ^{13}C-enriched DNA or RNA of microorganisms active in the studied process can be separated from the others after isopycnic ultracentrifugation on the

basis of a density modification. SIP-RNA strategy has the advantage to be more direct and more sensitive than SIP-DNA, the latter being dependent on a cell division and the ^{13}C enrichment signal being diluted due to the semiconservative DNA replication (Dumont et al. 2011). To assess hypotheses about evolution of plant–microorganisms interactions, SIP-RNA was applied on plant roots after incubation of the plant using the stable isotope $^{13}C–CO_2$ at atmospheric concentration (Vandenkoornhuyse et al. 2007; Kiers et al. 2011). Interestingly, the results showed the ability of plants to sanction symbiotic cheaters (Kiers et al. 2011). SIP-RNA was also used to identify fractions of microbial diversity involved in the degradation of organic molecules (Monard et al. 2008) including xenobiotic pollutant in soil, the atrazine (Monard et al. 2011). Authors showed that atrazine degradation likely relied on keystone microbial species. Combined with high-throughput sequencing, SIP-DNA was applied on lake sediment microbial communities targeting single-carbon oxidizing compounds (Kalyuzhnaya et al. 2008). Samples of the same community were incubated with different labeled one-carbon substrates separately, and the different metagenomes containing only the labeled DNA were sequenced and compared. The results showed specific sequence enrichments in response to different carbon-one substrates, thereby identifying the ecological roles of individual phylotypes. In addition, due to the enrichment, via compositional binning, a nearly complete genome of a novel methylotroph, *Methylotenera mobilis*, could be reconstituted.

Alternative labeling can be performed with labeled bromodeoxyuridine, an analog of thymidine that incorporates in the DNA in actively replicating microorganisms. Combined with pyrosequencing, this was applied on coastal ocean plankton–targeting species that utilize dissolved organic carbon (Mou et al. 2008). This study revealed that the coastal microbial communities are populated by taxa capable of metabolizing a wide variety of organic carbon compounds.

Simpler approaches to link diversity to function can be applied if genes encoding key enzymes are characteristic for particular taxa. A good example is the gene amoA encoding for one of three subunits of the ammonium monooxygenase, the key enzyme of *nitrification*.* The diversity of nitrifying microorganisms is very reduced, and they are found within the bacteria affiliated to beta- and gamma-proteobacteria and recently identified within the Archaea from the phylum Thaumarchaeota. Both genes, the bacterial and archaeal amoA are sufficiently different that they can be discriminated by *primers** used in PCR. Therefore, simple PCR or qPCR methods can be applied to detect or quantify the two types of nitrifiers separately (Leininger et al. 2006; Agogué et al. 2008; and others). Combined with 16S rRNA applying qPCR, the distribution and repartition of bacterial and archaeal nitrifiers can be detected, representing a very efficient tool linking the diversity to function targeted on an essential step of the nitrogen cycle.

7.6 Future

7.6.1 Sequencing Technologies: Read Length Matters

The future is in part dependent on new instruments and technologies for sequencing. The throughput is no longer a problem for sequencing. The main expected innovation is the increase of sequencing read length, simplifying genome assembly, and in consequence data analyses and at the same time requiring lower sequence *coverage**. This innovation should facilitate the linkage between diversity and function in ecosystems processes to analyze the rules of *community assembly** and the evolution of host-associated microbiota. Different companies are working on such innovation. Nanopore technology is based on measurement of electric potential differences induced when a native DNA molecule pass through the nanopore (Oxford Nanopore Technologies, United Kingdom). The main advantage of this strategy is that there is no instrument required, the sequencing being done on a small device from which the data are collected directly. To date, the still relatively high error rate (i.e., >10%) can be combined with high-quality short reads (i.e., Illumina sequencing). The other most advanced technology is based on light emission optically detected within a small chamber each with a zero-mode waveguide (Pacific Bioscience, United States). The light emission generated while copying the target DNA is measured within the chamber. In addition to the low error rate, this technology allows the direct analysis of particular epigenetic modifications (i.e., bacterial DNA modifications and methyltransferase recognition motifs; eukaryotic hyper- and hypo-methylated CpG islands) (see Section 7.6.3).

7.6.2 Metaviromes

Viruses are the most abundant biological entities exceeding the number of microorganisms by a factor of 5–25 (Fuhrman 1999). Metaviromes, that is, application of metagenomics on viral communities, were generated already in 2002 applying Sanger sequencing (Breitbart et al. 2002) and later by pyrosequencing (Edwards and Rohwer 2005; Thurber et al. 2009). Since viruses do not possess a common conserved molecular marker as the SSU rRNA genes, the viral diversity is essentially assessed by metagenomic approaches searching primarily to reconstruct complete viral genomes leading to the discovery of novel viral lineages (i.e., Roux et al. 2012; Quaiser et al. 2015). Metaviromes from all kind of habitats are available, for example, from stromatolites (Desnues et al. 2008), Antarctic lakes (López-Bueno et al. 2009), freshwater lakes (Roux et al. 2012), and peatlands (Ballaud et al. 2016). One of the major outcomes of these studies is the enormous diversity of viral sequences making their genomes the largest resource of genetic information on earth. This is in accordance with the high rates of evolutionary changes

in viral genomes that are much less conserved than microbial genes (Duffy et al. 2008). Combined with the widespread horizontal gene transfers in the viral world, their abundance, and their diversity, viruses strongly impact the microbial world whose extent is still unclear.

7.6.3 Sequencing the Epigenome

In the Neo-Darwinian synthesis of evolution, phenotypes are seen as a consequence of genotypes. Nevertheless, we know that other drivers can control phenotypes. The epigenetic mechanisms (i.e., DNA methylation, histone modifications, histone variants, and small RNA mainly) can affect gene expression and phenotypes. If selection acts on phenotypes that are not necessarily genetically controlled, the new phenotypes, arising from adaptive plasticity, are not random variants and the trait frequency can be seen as a consequence of genetic accommodation or the transgenerational inheritance of epigenetic mark. Considering that both Eukarya and Prokaryotes have epigenetic marks in their DNAs, the understanding of genome functioning and evolution needs to consider more carefully these epigenetic mechanisms within a multi-omics approach.

7.6.4 Modeling: Toward "Systems Ecology"?

Research works published so far are generally focused and limited to a single level of ecological organization (Figure 7.4). They have allowed identifying a number of interactions but are limited to this level of ecological organization. In the same framework of system biology, one important prospect is to develop a "systems ecology" integrating the different levels of ecological hierarchy. The reductionist approach, still perfectly valid and justified, has however the disadvantage of not being able to provide convincing concepts or methods to understand the emergent properties of an ecological system.

It is now necessary to develop a *holistic** approach taking into account all the levels of the ecology and integrating the physical and chemical processes at different spatial and temporal scales. The pluralism of causes and effects on the organization and evolution of ecological systems should be integrated simultaneously by different approaches (i.e., multi-omics, spatio-temporal dynamics, interaction with environmental conditions) (Figure 7.4). Recently, a new theory based on the simple idea that flow of carbon through ecosystems, as a consequence of metabolisms of individual organisms, has been developed (Schramski et al. 2015). Other modeling strategies of multi-organisms are currently developed on the basis of metabolic networks with the idea of parameterization of pathways using stoichiometric coefficients (i.e., flux balance analysis) and by developing multi-objective approaches to integrate a (micro)organism network. This kind of modeling requires a very precise knowledge, and stoichiometric matrixes of reaction should be tuned by multi-omics data.

From this complex system, we can anticipate new understandings and new hypotheses from the description of emerging properties. We can also speculate the possibility that this "system ecology" will be a new tool to make up scaling inferences.

Glossary

Amplicon: The result of the in vitro synthesis of a targeted DNA using PCR.

cDNA: Double-stranded DNA with a sequence that is complementary to the RNA template from which it is synthesized in a reaction catalyzed by reverse transcriptase.

Community assemblage: A group of populations of different species co-occurring within a given habitat.

Coverage/sequence coverage: The average numbers of sequenced nucleotides that contribute to a portion of an assembly.

Diachronic information in DNA: Sequence variations due to mutation representing evolutionary information.

Evolutionary-related organisms: Organisms sharing a phylogenetic ancestor.

Fluorescence in situ hybridization: Method that uses labeled oligonucleotide probes binding to a specific DNA target to visualize particular cells or microorganisms.

Fosmid: Particular vector designed for the insertion of large-size DNA fragments.

Genome/sequence assembly: The aligning and merging of sequence fragments aiming to reconstruct the original sequence (i.e., full-length chromosome or genome).

Haploid: A single set of chromosome or a cell having a single copy genome.

Haplotype/haplotypic variation: A set of genetic variations, combination of allele on the same chromosome.

Holistic: Integrative perception; parts are interconnected to represent the whole.

Horizontal gene transfers: Transfers of DNA fragments from the environment into the genome of a given organism. This transfer differs from vertical gene transmission, that is, parent to offspring.

Large insert libraries: Collections of cloned DNA inserts of long size (>20,000 bp).

Methanogenesis: Metabolic pathway of methane production by microorganisms (methanogens).

Methanotrophy/methanotroph: The capacity to metabolize methane as a source of carbon and energy by microorganisms, that is, the methanotrophs.

Molecular clock: It is possible to detect and count mutations among homologous DNA regions (i.e., comparison). DNA sequences accumulate a certain number of mutations per unit of time. The more the organism is ancient in term of evolution, the more mutations accumulate.

Mesophile/mesophilic: Ability to grow at moderate temperature, that is, 20°C–40°C.

Metagenomics/metatranscriptomics: Sequencing and analysis of bulk DNA or cDNA (RNA) from the mixed genomes of microbial communities.

Monophyletic group: Group of organisms sharing the same common ancestor.

Nitrification: Key biological process of the nitrogen cycle by which a microorganism transforms ammonium to nitrite by oxidation.

Open reading frames (ORF): An ORF is a continuous stretch of codons that is not interrupted by stop codon.

Phylogeny: Sequence-based analyses of the evolutionary relationship between a set of organisms.

Plastid DNA: DNA contained in the chloroplast.

Pleiotropic/pleiotropy: Multiple effects from a single origin, for example, multiple phenotypes driven by a single gene.

Primer: Refers to a specific oligonucleotide designed by the user for in vitro synthesis of a targeted DNA for use in polymerase chain reaction.

Pyrosequencing: DNA sequencing method based on "sequencing by synthesis."

Sanger technology: First DNA sequencing method that incorporates chain-terminating dideoxynucleotides by copying the target DNA to sequence.

Scaffolds/sequence scaffolds: Portion of the genome sequence reconstructed from contigs. A contig is a shorter genomic sequence assembled from smaller sequences/reads.

Read length: The number of nucleotide sequenced from a sequenced DNA molecule.

Rhodopsin/proteorhodopsin: Molecule at the basis of photoheterotrophy capturing light energy.

Sequencing depth: See coverage.

Species concepts: Concept applied to group individual organisms within a species. In animals, the biological species concept defines a species as organisms that can sexually reproduce. However, most of the (micro)organisms have clonal reproduction and the biological species concept cannot be used. The phylogenetic species concepts defining a species as a *monophyletic group** of individuals is often adopted.

Vector: DNA molecule that carries foreign genetic material in cells.

References

Agogué, H, M Brink, J Dinasquet, and G J Herndl. 2008. Major gradients in putatively nitrifying and non-nitrifying archaea in the deep North Atlantic. *Nature* 456 (7223): 788–791.

Angly, F E, B Felts, M Breitbart, et al. 2006. The marine viromes of four oceanic regions. *PLoS Biology* 4 (11): e368.

Ballaud, F C, A Dufresne, A-J Francez et al. 2016. Dynamics of viral abundance and diversity in a sphagnum-dominated peatland: Temporal fluctuations prevail over habitat. *Frontiers in Microbiology* 6: 1494.

Béjà, O, L Aravind, E V Koonin et al. 2000. Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* 289 (5486): 1902–1906.

Béjà, O, E N Spudich, J L Spudich, M Leclerc, and E F DeLong. 2001. Proteorhodopsin phototrophy in the ocean. *Nature* 411 (6839): 786–789.

Ben Maamar, S, L Aquilina, A Quaiser et al. 2015. Groundwater isolation governs chemistry and microbial community structure along hydrologic flowpaths. *Frontiers in Microbiology* 6: 1457.

Biers, E J, S Sun, and E C Howard. 2009. Prokaryotic genomes and diversity in surface ocean waters: Interrogating the global ocean sampling metagenome. *Applied and Environmental Microbiology* 75 (7): 2221–2229.

Blaxter, M. 2003. Molecular systematics: Counting angels with DNA. *Nature* 421 (6919): 122–124.

Breitbart, M, P Salamon, B Andresen et al. 2002. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America* 99 (22): 14250–14255.

Bulgarelli, D, M Rott, K Schlaeppi et al. 2012. Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* 488 (7409): 91–95.

Carvalhais, L C, P G Dennis, G W Tyson, and P M Schenk. 2012. Application of metatranscriptomics to soil environments. *Journal of Microbiological Methods* 91 (2): 246–251.

Chase, M W and M F Fay. 2009. Barcoding of plants and fungi. *Science* 325 (5941): 682–683.

Collins, F S, E D Green, A E Guttmacher, M S Guyer, and US National Human Genome Research Institute. 2003. A vision for the future of genomics research. *Nature* 422 (6934): 835–847.

Curtis, T P and W T Sloan. 2005. Exploring microbial diversity—A vast below. *Science* 309 (5739): 1331–1333.

DeLong, E F, C M Preston, T Mincer et al. 2006. Community genomics among stratified microbial assemblages in the ocean's Interior. *Science* 311 (5760): 496–503.

Deschamps, P, Y Zivanovic, D Moreira, F Rodríguez-Valera, and P López-García. 2014. Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic thaumarchaeota and euryarchaeota. *Genome Biology and Evolution* 6 (7): 1549–1563.

Desnues, C, B Rodriguez-Brito, S Rayhawk et al. 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* 452 (7185): 340–343.

de Vargas, C, S Audic, N Henry et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348 (6237): 1261605.

Doré, J and H Blottière. 2015. The influence of diet on the gut microbiota and its consequences for health. *Current Opinion in Biotechnology* 32: 195–199.

Duffy, S, L A Shackelton, and E C Holmes. 2008. Rates of evolutionary change in viruses: Patterns and determinants. *Nature Reviews Genetics* 9 (4): 267–276.

Dumont, M G, B Pommerenke, P Casper, and R Conrad. 2011. DNA-, rRNA- and mRNA-based stable isotope probing of aerobic methanotrophs in lake sediment. *Environmental Microbiology* 13 (5): 1153–1167.

Edwards, R A and F Rohwer. 2005. Viral metagenomics. *Nature Reviews Microbiology* 3 (6): 504–510.

Edwards, R A, B Rodriguez-Brito, L Wegley et al. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7 (1): 57.

Eid, J, A Fehr, J Gray et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science (New York, NY)* 323 (5910): 133–138.

Feike, J, K Jürgens, J T Hollibaugh et al. 2011. Measuring unbiased metatranscriptomics in suboxic waters of the central Baltic Sea using a new in situ fixation system. *The ISME Journal* 6 (2): 461–470.

Fierer, N, J W Leff, B J Adams et al. 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences of the United States of America* 109 (52): 21390–21395.

Fieseler, L, A Quaiser, C Schleper, and U Hentschel. 2006. Analysis of the first genome fragment from the marine sponge-associated, novel candidate phylum poribacteria by environmental genomics. *Environmental Microbiology* 8 (4): 612–624.

Frias-Lopez, J, Y Shi, G W Tyson et al. 2008. Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences of the United States of America* 105 (10): 3805–3810.

Fuchs, B M, S Spring, and H Teeling. 2007. Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 104 (8): 2891–2896.

Fuhrman, J A. 1999. Marine viruses and their biogeochemical and ecological effects. *Nature* 399 (6736): 541–548.

Gasc, C, C Ribière, N Parisot et al. 2015. Capturing prokaryotic dark matter genomes. *Research in Microbiology* 166: 814–830.

Ghahramanzadeh, R, G Esselink, L P Kodde et al. 2013. Efficient distinction of invasive aquatic plant species from non-invasive related species using DNA barcoding. *Molecular Ecology Resources* 13 (1): 21–31.

Ghai, R, A-B Martin-Cuadrado, A G Molto et al. 2010. Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *The ISME Journal* 4 (9): 1154–1166.

Gifford, S M, S Sharma, M Booth, and M A Moran. 2013. Expression patterns reveal niche diversification in a marine microbial assemblage. *The ISME Journal* 7 (2): 281–298.

Gladyshev, E A, M Meselson, and I R Arkhipova. 2008. Massive horizontal gene transfer in bdelloid rotifers. *Science* 320 (5880): 1210–1213.

Goldenfeld, N and C R Woese. 2007. Biology's next revolution. *Nature* 445 (7126): 369.

Gómez-Consarnau, L, J M González, M Coll-Lladó et al. 2007. Light stimulates growth of proteorhodopsin-containing marine flavobacteria. *Nature* 445 (7124): 210–213.

Hallam, S J, T J Mincer, C Schleper et al. 2006. Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine crenarchaeota. *PLoS Biology* 4 (4): e95.

He, S, O Wurtzel, K Singh et al. 2010. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods* 7 (10): 807–812.

Hultman, J, M P Waldrop, R Mackelprang et al. 2015. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* 521 (7551): 208–212.

Kalyuzhnaya, M, A Lapidus, N Ivanova et al. 2008. High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology* 26: 1029–1034.

Keeling, P J and J Palmer. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9 (8): 605–618.

Kielak, A M, J A van Veen, and G A Kowalchuk. 2010. Comparative analysis of acidobacterial genomic fragments from terrestrial and aquatic metagenomic libraries, with emphasis on acidobacteria subdivision 6. *Applied and Environmental Microbiology* 76 (20): 6769–6777.

Kiers, E T, M Duhamel, Y Beesetty et al. 2011. Reciprocal rewards stabilize cooperation in the mycorrhizal symbiosis. *Science* 333 (6044): 880–882.

Knief, C, N Delmotte, S Chaffron et al. 2012. Metaproteogenomice. *The ISME Journal* 6 (7): 1378–1390.

Korem, T, D Zeevi, J Suez et al. 2015. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* 349 (6252): 1101–1106.

Koven, C D, B Ringeval, P Friedlingstein et al. 2011. Permafrost carbon-climate feedbacks accelerate global warming. *Proceedings of the National Academy of Sciences of the United States of America* 108 (36): 14769–14774.

Leininger, S, T Urich, M Schloter et al. 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 442 (7104): 806–809.

Liles, M R, B F Manske, S B Bintrim, J Handelsman, and R M Goodman. 2003. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Applied and Environmental Microbiology* 69 (5): 2684–2691.

López-Bueno, A, J Tamames, D Velázquez et al. 2009. High diversity of the viral community from an Antarctic lake. *Science* 858 (November): 1–25.

López-García, P and D Moreira. 2008. Tracking microbial biodiversity through molecular and genomic ecology. *Research in Microbiology* 159 (1): 67–73.

Lundberg, D S, S L Lebeis, S H Paredes et al. 2012. Defining the core *Arabidopsis thaliana* root microbiome. *Nature* 488 (7409): 86–90.

Mackelprang, R, M P Waldrop, K M DeAngelis et al. 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480 (7377): 368–371.

Marcy, Y, C Ouverney, E M Bik et al. 2007. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the National Academy of Sciences of the United States of America* 104 (29): 11889–11894.

Margulies, M, M Egholm, W E Altman et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437 (7057): 376–380.

Martín-Cuadrado, A-B, F Rodríguez-Valera, D Moreira et al. 2008. Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *The ISME Journal* 2 (8): 865–886.

Maxam, A M and W Gilbert. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* 74 (2): 560–564.

Microbiology by Numbers. 2011. *Nat Rev Micro* 9 (9). Editorial: 628.

Monard, C, F Binet, and P Vandenkoornhuyse. 2008. Short-term response of soil bacteria to carbon enrichment in different soil microsites. *Applied and Environmental Microbiology* 74 (17): 5589–5592.

Monard, C, P Vandenkoornhuyse, B Le Bot, and F Binet. 2011. Relationship between bacterial diversity and function under biotic control: The soil pesticide degraders as a case study. *The ISME Journal* 5 (6): 1048–1056.

Moran, M A, B Satinsky, S Gifford et al. 2013. Sizing up metatranscriptomics. *The ISME Journal* 7 (2): 237–243.

Mou, X, S Sun, R A Edwards, R E Hodson, and M A Moran. 2008. Bacterial carbon processing by generalist species in the coastal ocean. *Nature* 451 (7179): 708–711.

Nelson, K E, R A Clayton, S R Gill et al. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature* 399 (6734): 323–329.

Nobu, M K, T Narihiro, C Rinke et al. 2015. Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *The ISME Journal* 9 (8): 1710–1722.

Nunes, F L D, L Aquilina, J de Ridder et al. 2015. Time-scales of hydrological forcing on the geochemistry and bacterial community structure of temperate peat soils. *Scientific Reports* 5: 14612.

Pace, N R, D A Stahl, D J Lane, and G J Olsen. 1986. The analysis of natural microbial populations by ribosomal RNA sequences. *Advances in Microbial Ecology* 9: 1–55.

Podar, M, C B Abulencia, M Walcher et al. 2007. Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Proceedings of the National Academy of Sciences of the United States of America* 73 (10): 3205–3214.

Poretsky, R S, N Bano, A Buchan et al. 2005. Analysis of microbial gene transcripts in environmental samples. *Proceedings of the National Academy of Sciences of the United States of America* 71 (7): 4121–4126.

Qin, J, R Li, J Raes et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464 (7285): 59–65.

Quaiser, A, X Bodi, A Dufresne et al. 2014. Unraveling the stratification of an iron-oxidizing microbial mat by metatranscriptomics. *PLoS One* 9 (7): 1–9.

Quaiser, A, A Dufresne, F Ballaud et al. 2015. Diversity and comparative genomics of microviridae in sphagnum-dominated peatlands. *Frontiers in Microbiology* 6: 375.

Quaiser, A, P López-García, Y Zivanovic et al. 2008. Comparative analysis of genome fragments of acidobacteria from deep Mediterranean plankton. *Environmental Microbiology* 10 (10): 2704–2717.

Quaiser, A, T Ochsenreiter, H-P Klenk et al. 2002. First insight into the genome of an uncultivated crenarchaeote from soil. *Environmental Microbiology* 4 (10): 603–611.

Quaiser, A, T Ochsenreiter, C Lanz et al. 2003. Acidobacteria form a coherent but highly diverse group within the bacterial domain: Evidence from environmental genomics. *Molecular Biology* 50 (2): 563–575.

Quaiser, A, Y Zivanovic, D Moreira, and P López-García. 2011. Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *The ISME Journal* 5 (2): 285–304.

Radax, R, T Rattei, A Lanzen et al. 2012. Metatranscriptomics of the marine sponge *Geodia barretti*: Tackling phylogeny and function of its microbial community. *Environmental Microbiology* 14 (5): 1308–1324.

Rinke, C, P Schwientek, A Sczyrba et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499 (7459): 431–437.

Rodríguez-Martínez, R, G Rocap, G Salazar, and R Massana. 2013. Biogeography of the uncultured marine picoeukaryote MAST-4: Temperature-driven distribution patterns. *The ISME Journal* 7 (8): 1531–1543.

Roman, J and B W Bowen. 2000. The mock turtle syndrome: Genetic identification of turtle meat purchased in the south-eastern United States of America. *Animal Conservation* 3 (01): 61–65.

Rondon, M R, P R August, A D Bettermann et al. 2000. Cloning the soil metagenome: A strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and Environmental Microbiology* 66 (6): 2541–2547.

Roux, S, F Enault, A Robin et al. 2012. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* 7 (3): e33641.

Rusch, D B, A L Halpern, G Sutton et al. 2007. The sorcerer II global ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology* 5 (3): e77.

Sanger, F and A R Coulson. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94 (3): 441–448.

Schlaeppi, K, N Dombrowski, R G Oter, E V L van Themaat, and P Schulze-Lefert. 2014. Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives. *Proceedings of the National Academy of Sciences of the United States of America* 111 (2): 585–592.

Schmidt, T M, E F DeLong, and N R Pace. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology* 173 (14): 4371–4378.

Schramski, J R, A I Dell, J M Grady, R M Sibly, and J H Brown. 2015. Metabolic theory predicts whole-ecosystem properties. *Proceedings of the National Academy of Sciences of the United Staes of America* 112 (8): 2617–2622.

Schuster, S C. 2008. Next-generation sequencing transforms today's biology. *Nature Methods* 5 (1): 16–18.

Shi, Y, G W Tyson, J M Eppley, and E F DeLong. 2011. Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *The ISME Journal* 5 (6): 999–1013.

Sogin, M L, H G Morrison, J A Huber et al. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* 103 (32): 12115–12120.

Stein, J L, T L Marsh, K Y Wu, H Shizuya, and E F DeLong. 1996. Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology* 178 (3): 591–599.

Stern, A and R Sorek. 2011. The phage-host arms race: Shaping the evolution of microbes. *BioEssays* 33 (1): 43–51.

Thurber, R V, M Haynes, M Breitbart, L Wegley, and F Rohwer. 2009. Laboratory procedures to generate viral metagenomes. *Nature Protocols* 4 (4): 470–483.

Treusch, A H, A Kletzin, G Raddatz et al. 2004. Characterization of large-insert DNA libraries from soil for environmental genomic studies of archaea. *Environmental Microbiology* 6 (9): 970–980.

Treusch, A H, S Leininger, A Kletzin et al. 2005. Novel genes for nitrite reductase and amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environmental Microbiology* 7 (12): 1985–1995.

Tringe, S G, C von Mering, A Kobayashi et al. 2005. Comparative metagenomics of microbial communities. *Science* 308 (5721): 554–557.

Turnbaugh, P J, M Hamady, T Yatsunenko et al. 2009. A core gut microbiome in obese and lean twins. *Nature* 457 (7228): 480–484.

Tveit, A, R Schwacke, M M Svenning, and T Urich. 2012. Organic carbon transformations in high-Arctic peat soils: Key functions and microorganisms. *The ISME Journal* 7 (2): 299–311.

Tveit, A T, T Urich, and M M Svenning. 2014. Metatranscriptomic analysis of Arctic peat soil microbiota. *Applied and Environmental Microbiology* 80 (18): 5761–5772.

Tyson, G W, J Chapman, P Hugenholtz et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428 (6978): 37–43.

Urich, T, A Lanzén, J Qi et al. 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* 3 (6): e2527.

Urich, T, A Lanzén, R Stokke et al. 2014. Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environmental Microbiology* 16 (9): 2699–2710.

Vandenkoornhuyse, P, A Dufresne, A Quaiser et al. 2010. Integration of molecular functions at the ecosystemic level: Breakthroughs and future goals of environmental genomics and post-genomics. *Ecology Letters* 13 (6): 776–791.

Vandenkoornhuyse, P, S Mahé, P Ineson et al. 2007. Active root-inhabiting microbes identified by rapid incorporation of plant-derived carbon into RNA. *Proceedings of the National Academy of Sciences of the United States of America* 104 (43): 16970–16975.

Vandenkoornhuyse, P, A Quaiser, M Duhamel, A Le Van, and A Dufresne. 2015. The importance of the microbiome of the plant holobiont. *The New Phytologist* 206 (4): 1196–1206.

Venter, J C, K Remington, J F Heidelberg et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304 (5667): 66–74.

Whitman, W B, D C Coleman, and W J Wiebe. 1998. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* 95 (12): 6578–6583.

Wilson, E O. 1989. Threats to biodiversity. *Scientific American* 261: 108–116.

Woese, C R. 1987. Bacterial evolution. *Microbiological Reviews* 51 (2): 221–271.

Woese, C R. 2004. A new biology for a new century. *Microbiology and Molecular Biology Reviews* 68 (2): 173–186.

Yoon, S H, M-J Han, H Jeong et al. 2012. Comparative multi-omics systems analysis of *Escherichia coli* strains B and K-12. *Genome Biology* 13 (5): R37.

Yoon, S S, E-K Kim, and W-J Lee. 2015. Functional genomic and metagenomic approaches to understanding gut microbiota-animal mutualism. *Current Opinion in Microbiology* 24: 38–46.

Yooseph, S, G Sutton, D B Rusch et al. 2007. The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biology* 5 (3): 432–466.