Cell. Mol. Life Sci. (2015) 72:4287–4308
DOI 10.1007/s00018-015-2004-1

Cellular and Molecular Life Sciences

REVIEW

From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems

Daniel R. Garza[1] · **Bas E. Dutilh**[1,2,3]

Received: 17 May 2015 / Revised: 23 July 2015 / Accepted: 28 July 2015 / Published online: 9 August 2015

Abstract Microorganisms and the viruses that infect them are the most numerous biological entities on Earth and enclose its greatest biodiversity and genetic reservoir. With strength in their numbers, these microscopic organisms are major players in the cycles of energy and matter that sustain all life. Scientists have only scratched the surface of this vast microbial world through culture-dependent methods. Recent developments in generating metagenomes, large random samples of nucleic acid sequences isolated directly from the environment, are providing comprehensive portraits of the composition, structure, and functioning of microbial communities. Moreover, advances in metagenomic analysis have created the possibility of obtaining complete or nearly complete genome sequences from uncultured microorganisms, providing important means to study their biology, ecology, and evolution. Here we review some of the recent developments in the field of metagenomics, focusing on the discovery of genetic novelty and on methods for obtaining uncultured genome sequences, including through the recycling of previously published datasets. Moreover we discuss how metagenomics has become a core scientific tool to characterize eco-evolutionary patterns of microbial ecosystems, thus allowing us to simultaneously discover new microbes and study their natural communities. We conclude by discussing general guidelines and challenges for modeling the interactions between uncultured microorganisms and viruses based on the information contained in their genome sequences. These models will significantly advance our understanding of the functioning of microbial ecosystems and the roles of microbes in the environment.

Keywords Biological dark matter · Pan-genomics, data recycling · Eco-systems biology · Metagenome-wide modeling · Virus-host association

✉ Bas E. Dutilh
bedutilh@gmail.com

[1] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands

[2] Theoretical Biology and Bioinformatics, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

[3] Department of Marine Biology, Institute of Biology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Introduction

Metagenomics is the study of genetic material recovered directly from environmental samples in an untargeted (shotgun) way. Current developments increasing the depth and breadth of metagenomic shotgun sequencing have facilitated the identification of complete or nearly complete microbial and viral genome sequences from environmental samples without the need to first cultivate these organisms. Here we name these sequences the "uncultured genome sequences" that can either be obtained from metagenomic datasets or from single-cell sequencing. While they frequently have a draft status, and depending on the approach may represent a locally occurring metapopulation rather than a single clone, uncultured genome sequences can supplement the genome sequences obtained by sequencing pure or nearly pure cultures of microbial isolates (Fig. 1), therewith greatly increasing the amount of data that is available for comparative genome analyses. Moreover, by providing reference sequences for the alignment of both

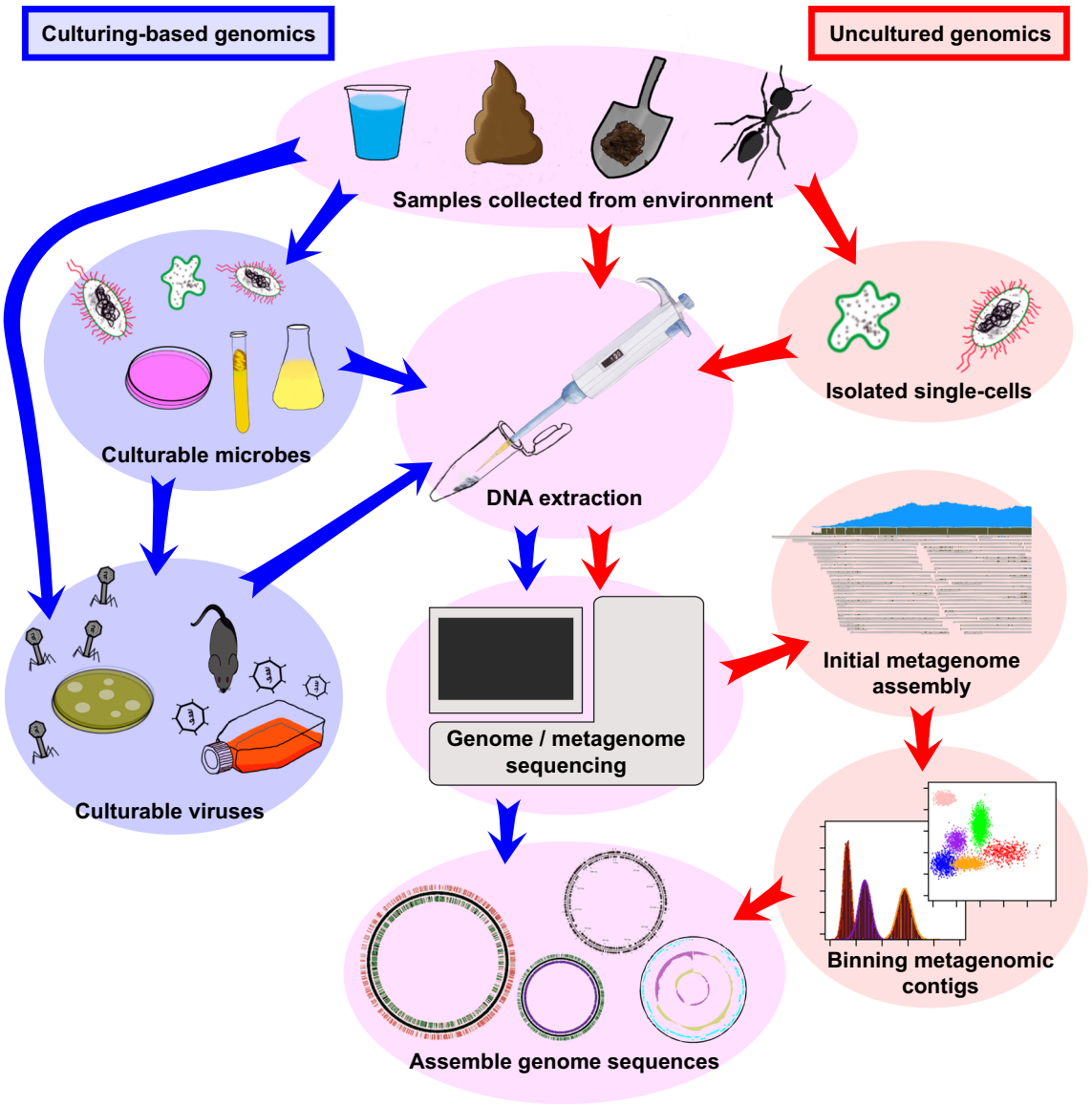

Fig. 1 Illustration of simplified pipelines to obtain genome sequences from cultured and uncultured microbes and viruses. There are many variations of each protocol and additional steps, such as filtering samples according to molecular size cutoffs and normalization of data which are not illustrated in this diagram. The purpose is to illustrate simplified general steps to obtain uncultured genomes, which are common in most of the studies discussed in this review

known and unknown metagenomic shotgun sequencing reads [1], they greatly enhance the breadth of our understanding of microbial ecosystems. Uncultured organisms may, or may not have close cultured relatives, but isolating complete or nearly complete genome sequences from metagenomes invariably identifies genetic novelty, revealing flexible pan-genomes, genetic variants, and new subpopulations of microbes.

The goal of this review is to introduce some of the recent landmarks of metagenomics in providing new insights into the uncultured microbial biosphere, and highlight the promises and challenges these new genome sequences bring for modeling natural microbial ecosystems. A historic perspective of the discovery of new microbes and viruses before and after metagenomics is given, followed by a discussion of the innovative tools that have been recently used by several research groups to obtain uncultured genomes from metagenomic datasets. Metagenomics is primarily a science of microbial communities, and a key interest is to describe and predict the interactions between different populations of microbes and viruses [2]. Thus, in the further sections of this review we focus on the use of metagenomics and uncultured genome sequences to understand the ecological and evolutionary

dynamics of microbial populations within the context of their natural environments. We conclude by discussing the recent developments and perspectives of genome-guided systems biology modeling frameworks to functionally couple the biological knowledge obtained from uncultured genome sequences with systems-level predictions of the dynamics of microbial communities.

Before metagenomics: culturing-dependent discovery of new microbes and viruses

The first accounts of the microscopic world beyond the resolution of the human eye were made by direct observations of microbes in environments such as water, soil, or diseased tissues. Antonj van Leeuwenhoek, a Dutch tradesman, was the first to build microscopes capable of viewing single-celled organisms. In the late seventeenth century when he reported his observations of "little animals" in water, he was ridiculed by the scientific establishment. Only after his observations were validated by an independent committee did scientists begin to believe that invisible single-celled organisms could be hidden in many habitats in our planet [3]. Before long, microorganisms were recognized as the causative agents of many poorly understood phenomena, particularly in human disease. More powerful microscopes and staining methods were further developed, including the Gram stain in the 19th century, which is still used widely as a first classification scheme for bacteria [4].

Despite the dominance of direct observation and culture-independent methods in early years [5–7], microbiology soon became a science of microbial isolates. After Robert Koch pioneered methods for the isolation of microbial colonies and established postulates to link diseases with causative microbial agents, isolation and cultivation became the most common approach for microbial characterization [8]. Today, many taxonomic and strain typing schemes depend on culturing, as do most laboratory methods for determining the identity and biological characteristics of microbial species.

Virology has followed a path that is very similar to bacterial microbiology. Much of the known viral biodiversity encompasses medically relevant viruses. Before the advances of PCR and DNA sequencing methods, sampling from diseased phenotypes and inoculating into tissue cultures or susceptible animals was the main source of isolation and discovery of new viruses [9]. Additionally, many bacterial viruses (known as bacteriophages) were discovered in rapidly growing, cultivable bacteria, thereby attributing the majority of the recognized bacteriophage biodiversity to fast growing hosts [10]. Thus, by the use of cultivation as a dominant technique in both bacterial and viral microbiology, much of the scientific knowledge has been based on cultivable species, biasing our understanding of microbial biodiversity towards the biology and ecology of the "easy growers" [11].

Caveats in studying cultured isolates

The study of cultured isolates has propelled microbiological research. The success of culturing microbial species and studying them in isolation is a consequence of the difficulties that would be involved in analyzing them within their natural environment, which is complex and contains many unknown variables. Reproducibility of results, control of external variables, and simple design of laboratory experiments are all advantageous properties that are greatly facilitated in pure culture studies. Nevertheless, studies of environmental microbes and viruses repeatedly confirm that the large majority has not been cultured and is thus poorly understood. The early studies that pointed to an abundance of unculturable microorganisms in the environment were largely forgotten by the scientific community [12–14]. As a result, the development of modern culture-free methods including metagenomics, have sometimes led to surprises in the past 20 years. For example, by visualizing and counting the microscopic biological particles in the environment and comparing these counts to the number of archaeal and bacterial isolates, or to the number of phage plaques that grew on a bacterial lawn, a great numeric discrepancy was observed between what was counted in the wild, and what could be cultured on a plate [10, 11]. Different environments, such as seawater, soil, or marine sediments, showed that only about 0.01–1 % of the microorganisms seen in the microscope could be isolated on artificial media, while the vast majority remained intractable to culture-dependent techniques. These discrepancies have been named the "great plate count anomaly" [11] and the "great plaque count anomaly" [10], respectively. Clearly, we do not yet truly understand microbial biodiversity, which begs basic questions such as, which bacteria or viruses are out there? What is a microbial species? How do microbes and viruses interact with each other? And how do they interact with their environment?

It is particularly relevant to broaden the phylogenetic breadth of cultured isolates in order to have more diversity available for experimental testing [15]. Moreover, since the majority of viruses in natural environments consist of bacteriophages, having a greater diversity of cultured bacterial isolates will also allow for a higher throughput in virus isolation strategies [16]. Given the observations of a vast, uncultured majority of microbes and viruses as outlined above (the great plate/plaque count anomaly), a natural question to ask is "Why do most bacteria, archaea, and viruses not grow in synthetic media?" [17]. Another

related question is "How can we increase the recovery of environmental microbes in pure culture?" Many authors who discuss these and similar questions suggest that there are no single answers, and that many answers are applicable only to specific taxonomic groups or hold only in particular environments [17]. Among the commonly suggested causes for the plate count anomaly, we can list (1) lack of essential nutrients in the isolation media [18–20]; (2) lack of an essential biological interdependency with other species, such as auxotrophs or obligate mutualists [21–23]; (3) poor correlation between the in vitro growth condition and the environment: e.g. the media are too rich or too poor in nutrients, or they have inappropriate pH, salinity, or temperature [19, 24, 25]; (4) microbe-specific features, such as small non-cultivable cells, or extremely slow growers [26]. Some of these causes are interrelated and may be addressed together (see below).

Methods to increase the plate count

Early approaches to increase the plate count were based on extensive testing of different media, such as the R2A media for drinking water biofilms [24], and low-throughput screening for compounds and co-factors that could increase the plate count for different environments [27]. Promising technologies are being developed, some of which can be extended to high-throughput approaches [28, 29]. These technologies allow for many different conditions and samples to be screened in parallel. Simultaneously screening bacterial phenotypes in different conditions is one example of a high-throughput approach that can be used to identify optimal culturing conditions [30]. Other approaches involve the cultivation of bacteria in their natural environment or the use of supplements and specific growth factors such as iron-chelating siderophores [19, 20]. Fe(II) is severely limited in most aerobic environments and some bacteria release siderophores to scavenge for Fe(II), which is then transported back into the cells. Siderophores from neighboring species induce growth of uncultured marine bacteria. By inoculating marine bacteria with high concentrations of Fe(II) as a surrogate for siderophores, D'Onofrio et al. [20] reported the isolation of many colonies of previously uncultured bacteria, including three with 16S rRNA gene sequences that were highly divergent from any known species [20].

Allowing small metabolites or signaling molecules from the natural sites of isolates to diffuse into inoculated surfaces was shown to recover up to 50 % of bacteria from some environmental samples, where traditional methods would only recover 0.01–0.05 % [18, 19]. In order to achieve these expressively higher colony yields, diffusion chambers built with washers, sandwiched between 0.03 μm pore membranes were used, and incubated together with the sediment collected from marine environments in a marine aquarium. Some bacteria grow in diffusion chambers only when paired with so-called "helper" species [22]. One of these bacteria, *Psychrobacter* sp. strain MSC33, started growing in isolation after successive co-cultures with its helper strain, *Cellulophaga lytica*. After acquiring the capacity to grow in isolation, *Psychrobacter* MSC33 in turn could be used as a helper strain for other bacteria. This phenomenon was reproduced with other strains that could only grow in co-culture and, importantly, it was also observed in rich media, suggesting that nutrient limitation was not the underlying mechanism for the initial inability of these strains to grow in isolation. Indeed, the authors identified a five-amino-acid signaling peptide, LQPEV, as responsible for inducing the growth of the otherwise unculturable *Psychrobacter* [22].

One example of nutrient interdependency as the limiting factor for obtaining pure bacterial cultures was found with *Treponema primitia*. This bacterium is a hydrogen consuming, carbon dioxide-reducing homoacetogenic spirochete from the termite hindgut, and relevant for the host due to nitrogen-fixing and acetate production functions. Graber and Breznak [21] showed that *T. primitia* only grows when folate is available and they suggest that this nutrient is provided by other microbial members in the termite hindgut [21].

A promising device for high-throughput isolation of microbes from natural environments is the iChip, which consists of hundreds of miniaturized diffusion chambers [29]. Recently a previously uncultured proteobacterium, *Eleftheria terrae*, was discovered by using this technology [25]. This bacterium produces a potent antibiotic named Teixobactin, which has was found to be active against Gram-positive bacteria not amenable to treatment, and is being suggested as an effective drug against methicillin-resistant *Staphylococcus aureus* MRSA [25].

Genome-guided culturing efforts

Finding the right culturing conditions or hosts to isolate novel microbes and viruses can be guided by mining uncultured genome sequences for clues of potential nutrient requirements. An example is provided by the SAR11 clade, which is the most abundant clade of heterotrophic bacteria in the ocean. As of 2002, these bacteria were known solely from evidence based on environmental sequencing data [31]. Although SAR11 isolates were obtained by using sterile seawater with several supplements [32], genome mining showed that these bacteria lacked assimilatory sulfate reduction genes, thus requiring exogenous sources of reduced sulfur, such as methionine or 3-dimethylsulphoniopropionate (DMSP) for growth. DMSP is provided by other plankton members and its

addition to the culture media significantly increased the biomass yield of SAR11 bacteria [23]. These results suggest that the availability of complete or nearly genome sequences for different representatives of the uncultured groups could guide isolation strategies for these different microbes.

Besides providing access to uncultured genome sequences, metagenomics can also be used to study microbes and viruses in the context of their interactions with other members of the biological community. This makes metagenomics a fundamental tool to be integrated with environmental microbiology and the study and discovery of novel microbial biodiversity. Ideally, there is a feedback loop between bioinformatic approaches that obtain uncultured genome sequences from shotgun metagenomic datasets, and the laboratory where these genome sequences are exploited to guide the cultivation efforts of new microbial species (Fig. 2). First, the phenotypic and genetic characterization of cultured microbial isolates can populate databases with data that help to increase the accuracy of the information that can be obtained from their genome sequences. Second, obtaining uncultured genome sequences from metagenomes can uncover the gene composition of a species and its putative phenotype space, providing meaningful information for attempts to isolate microbial species. Moreover, the distribution across environments can also be retrieved from metagenomic analyses, which can be used to predict ecological interactions and lifestyles.

Genome-guided culturing is a vastly underexplored area in the field of metagenomics. Examples of uncultured genomes that could be amenable to these approaches include the candidate phyla OD1, OP11, and BD1-5 [33]. These three candidate phyla are part of a monophyletic group of widespread uncultured bacteria that have only recently been recognized by metagenomic sequencing, and were shown to comprise a super-phylum that encompasses an estimated 15 % of the bacterial domain [34]. Genomic evidence suggests that these bacteria have small genomes and may depend on other community members for

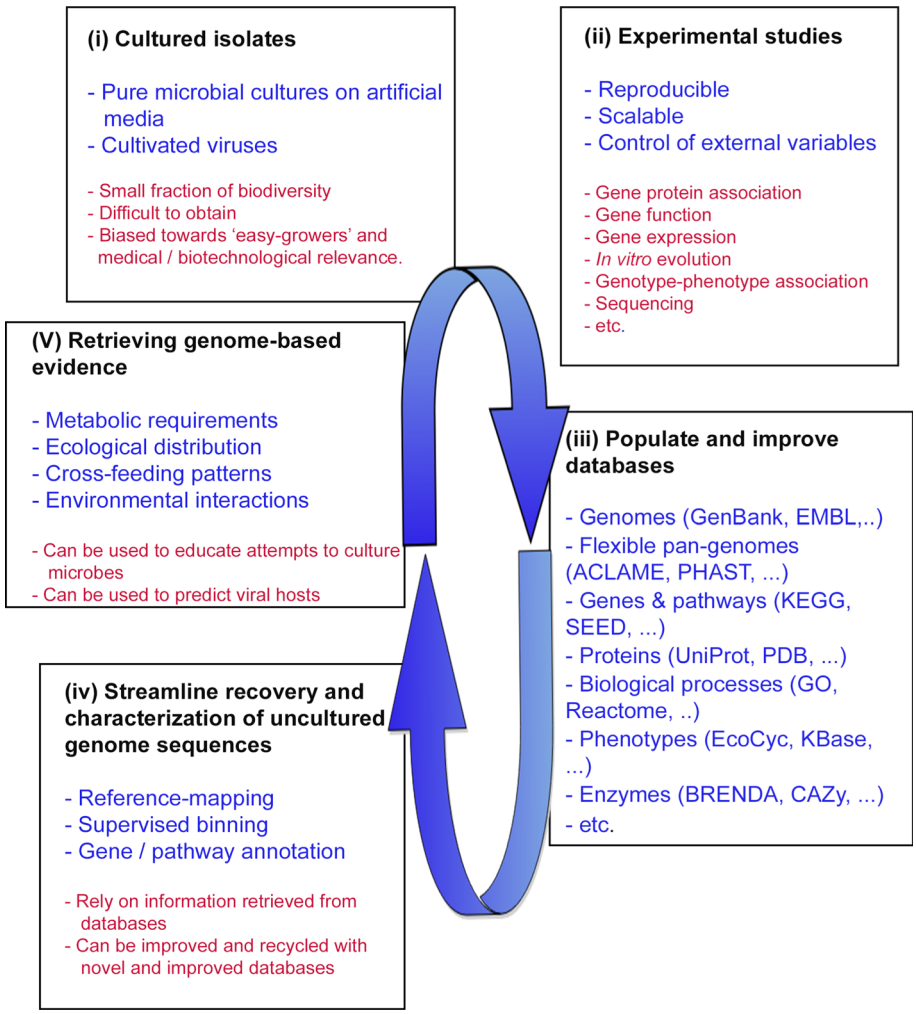

Fig. 2 Diagram of the feedback loop between experimental studies on cultured isolates and genome-based evidence retrieved from sequenced genomes. Uncultured genomes can educate genome-guided culturing attempts, which are suggested in the main text

essential nutrients [34, 35]. Deep sequencing revealed that besides remarkably small genomes, they lack many known biosynthetic pathways [36] and analysis of their ultrastructure suggests that they are indeed naturally ultra-small cells with median volumes of 0.009 μm^3, but are biologically active [35]. Enrichment for a member of the BD1-5 bacteria in a chemostat containing a mixed culture [37] suggests that these bacteria could be amenable to cultivation under laboratory conditions. Even before uncultured genome sequences were available, Harris et al. [33] suggested using the environmental distribution patterns inferred from 16S rRNA amplicon sequencing to develop isolation strategies for these groups.

To conclude, cultured isolates are critical for reproducible experimental studies. Isolates are useful for many biotechnology and health applications, such as genotype-phenotype screening, gene knockouts, screening for secondary metabolites, and phage-host assays. Nevertheless, there are many difficulties in the process of obtaining cultured representatives for the vast diversity of microorganisms and viruses, which can thus only be studied by using culture-independent methods.

Metagenomics approaches to study new microbes and viruses

Marker genes and the phylogenetic identity of uncultured bacteria and archaea

Estimates of the size of the environmental microbial and viral biodiversity that remains to be discovered are vast. In bacteriophages, for instance, it has been estimated that there are on the order of 100 million undiscovered types with possibly billions of new genes [38]. Knowledge of the microbial world is dependent on tools that increase the signal-to-noise ratio of the uncultured genome sequences in metagenomes that represent the hidden members of microbial communities. While the first studies that addressed uncultured microorganisms could only infer their presence by the shapes and stains under the microscope, in the past 50 years, developments in molecular biology have provided advanced tools to survey and quantify this hidden majority. The developments of the polymerase chain reaction (PCR), fluorescence in situ hybridization (FISH) [39], advances in DNA sequencing technology, and use of the 16S rRNA gene as a taxonomic marker [40, 41], have enabled the genetic identification of bacteria and archaea that are found in different environmental samples. By isolating DNA samples from whole communities of microorganisms and further amplifying and sequencing fragments of the 16S rRNA gene selected with degenerate primers, the genetic identity of a representative portion of the microbial community can now be known (for a review see ref [42]).

The 16S rRNA gene and other taxonomic marker genes have provided the means both to identify microbes by sequence similarity, and to cluster them into taxonomic groups in a phylogenetic context. Moreover, these marker genes have enabled estimates of the proportion of biodiversity that remain uncultured, revealing whole phyla that lack cultured representatives [34, 43]. Importantly, these phyla cannot be classified with conventional taxonomic approaches, which rely on polyphasic phenotypic and genetic typing schemes that are currently inaccessible for uncultured microbes [44, 45]. Uncultured groups suggested by this method are thus termed candidate phyla. Currently, more than half of the known bacterial and archaeal phyla lack cultured representatives.

Uncultured genome sequences come into play

A metagenome consists of the genomic sequences of all the organisms present in a given environment. Metagenomics can be defined as the application of high-throughput sequencing and analysis pipelines to elucidate a representative, random fraction of the genome sequences in a biological sample [46].

Before shotgun metagenomics, environmental sequencing efforts focused on the processing of amplified phylogenetic marker gene sequences. Since then, metagenomics has evolved into the application of shotgun sequencing aimed at obtaining sequencing reads from a comprehensive fraction of the nucleic acids in a sample (for general reviews about metagenomics see refs. [47–49]). Some of the first metagenomic studies consisted of shearing environmental DNA from soil samples into large fragments, cloning these fragments into BAC vectors and screening for functional traits [50, 51]. This approach of enriching and screening for functional genes is now named functional metagenomics to differentiate it from approaches that were aimed primarily at discovering the global sequence content of environmental samples [52–54].

One of the first comprehensive shotgun metagenomics studies was conducted on eight large water samples from different sites of the Sargasso Sea [54]. Fosmid libraries were generated from isolated and fragmented DNA from this community and sequenced by the dideoxy chain-termination method (Sanger sequencing). More than 1.5 Gbp of sequences were generated, many of which could be assembled into scaffolds, suggesting the presence of countable, discrete species rather than a genomic continuum [54]. These were among the first sequences of uncultured microorganisms and contained partial genome sequences from phyla that had no cultured representatives, such as the SAR86 clade. Using the term "genomic

species", the authors clustered genome fragments by using a similarity cutoff and found direct evidence that at least 451 different uncultured genome sequences were sampled. Additionally, many new genes were discovered and assigned to functional categories.

Since these first endeavors, DNA sequencing of microbial communities has evolved from the Sanger sequencing methods, which rely on a labor-intensive cloning process, to Next Generation Sequencing (NGS) technologies such as the 454/Roche, Illumina/Solexa, and Ion Torrent/Ion Proton platforms [55]. These short read approaches are particularly suited for taxonomic and functional profiling of metagenomic samples, as they provide a random sample of the sequences therein [56, 57]. Thus, and as a result of the rapidly decreasing cost of short read sequencing, such profiling analyses have been the driver of the field of metagenomics in the past decade. With the further decrease in cost and increase in sequencing volumes and read lengths, for example by PacBio and Oxford Nanopore sequencing technologies, assembly of (draft) uncultured genome sequences is now becoming increasingly accessible. We will discuss new promising methods for identifying and characterizing these uncultured genome sequences in the paragraphs below.

Bioinformatic approaches to obtain uncultured genome sequences

Assembly of uncultured genome sequences from complex shotgun metagenomes is progressing with the rapid development of new sequencing methods and bioinformatics pipelines [58]. Below we will review approaches that have been developed and used by several research groups to build uncultured genome sequences de novo. A metagenomic sample consists of random fragments of multiple genomes from different organisms. These genomes contain signals such as phylogenetic or sequence based signals that have been acquired in the course of evolution [59–61], signals that are the result of the ecological process [62, 63], or signals resulting from the sampling strategy [64]. These signals may be exploited to group metagenomic sequence fragments belonging to the same organism together, in order to bin and assemble the original uncultured genome sequences.

The naturally occurring sequence diversity of microbial genomes, whether derived from co-existing strains or from a (viral) quasispecies, often prohibits the assembly of longer contigs [65]. In bioinformatics, the process of grouping genomic fragments such as reads or assembled contigs putatively derived from the same organism based on sequence signals, is called binning, and many bioinformatic tools are available to do this [64, 66–69]. From a bioinformatics point of view, the most important signals available for contig binning are: homology to a reference sequence, paired sequencing read information, oligonucleotide composition, and differential abundance patterns across metagenomic samples. Moreover, an experimental approach that was recently developed exploits Hi-C, a technology that was developed to detect chromosomal organization in eukaryotic cells, to identify DNA sequences that are co-localized within microbial cells of an environmental sample [70–72]. We expect that additional experimental and bioinformatic approaches will be developed for binning uncultured genome sequences from metagenomes, as the opportunities for interpreting and analyzing uncultured genome sequences improve (see below).

Binning approaches can be classified into supervised and unsupervised methods. Supervised methods generally use a reference database of known genomes as a training set, and apply statistical classification methods, such as hidden Markov models [73, 74] or similarity/distance matrix models [75], to classify reads. These classification approaches can be used to remove or isolate clusters of sequence fragments according to a specific signal and thus reduce the complexity and size of the assembly challenge. Supervised methods can also be used to classify reads and assemble genomes from the resulting bins [73, 76, 77].

Homology-based signals consist of aligning sequencing reads or contigs to a reference sequence that can consist of the genome of a known species, or of contigs assembled from the same or a similar metagenome [78–80]. An obvious limitation of the supervised approaches is that they are restricted to discovering genomes that are similar to the genomes which were used as training sets, making them unsuitable to discover completely novel genome sequences. Nevertheless, these algorithms tend to improve as an increasing number of reference sequences become available, particularly of uncultured organisms, and they can be continuously calibrated towards more adequate training sets.

Unsupervised methods do not depend on a database of known reference genomes [81]. These methods are generally dependent on sequencing strategy, or on sequence content and sample composition. For example, as in cultured genome assembly, paired sequencing reads are commonly exploited for scaffolding assembled contigs, by mapping read pair sequences to the assembled contigs, and linking the contigs that share many paired sequencing reads [82]. The computational performance of alignment-based approaches that depend on alignment of many sequences is also rapidly improving thanks to innovative bioinformatic tools [83, 84].

Binning signals based on sequence content include the percentage of G and C nucleotides in the contig, as well as the oligonucleotide usage profile that are both relatively

consistent along the length of the genome. For these approaches, larger fragments or contigs result in better approximations of the genomic oligonucleotide usage profile, and better binning. These alignment-free methods can be very fast and memory-efficient, because binning is achieved by simple binary vector operations, which computers perform extremely fast. An example of an unsupervised approach that exploits oligonucleotide usage profiles is emergent self-organizing maps (ESOMs) [85–88].

Signals that are based on differential abundance patterns of a genome within or across metagenomic samples exploit the consistency in the expected depth of coverage of contigs that are derived from the same genome. Because different genomes are present in different frequencies in a sample, fragments from one genome are expected to have the same depth of coverage in the metagenomic dataset, thus reflecting the abundance of that genome in the original sample. If multiple metagenomic samples are obtained from a similar environment, each with variations in the abundances of the different members of the microbial community, the depth of coverage of contigs derived from one genome is expected to vary consistently across samples. This allows for fragment binning based on co-abundance across multiple metagenomic samples [64]. This differential abundance signal, in combination with oligonucleotide usage profiles were used to identify 49 nearly complete uncultured bacterial genome sequences from an acetate-amended aquifer [62].

The assembly of high quality uncultured genome sequences from metagenomic datasets is still a relatively low throughput process that usually yields only a few nearly complete genomes. In part, this depends on the sequencing volume and the species richness of the sampled community, which together determine the expected assembly depth of the uncultured genome sequences. The greatest bottleneck, however, is the effort that goes into finishing a genome sequence. For bacteria and archaea, the completeness and redundancy of an assembled uncultured genome sequence consisting of a cluster of binned contigs, can be assessed by identifying universal single copy marker genes [34, 89, 90]. The percentage of these universal genes identified in the assembled genome corresponds to the expected genome completeness, while duplicates among these single copy genes indicate redundancy. For viruses, such universal marker genes are not available, and currently the most reliable way to establish completeness of an assembled genome sequence is by validating that the assembled contig represents a circular genome [63, 91, 92]. However, with new bioinformatic advances [64, 66–69], the recovery of uncultured genome sequences, whether in draft or complete form, is increasingly yielding new knowledge about natural microbes and viruses, as outlined below.

Examples of landmark uncultured genome sequence assemblies

Tyson et al. [53] were the first to assemble nearly complete uncultured genome sequences from a metagenome library of small-insert plasmid clones. The isolation and reconstruction of genomes was possible because the sampled community consisted of low-complexity biofilms containing few different species. After an initial assembly of shotgun reads, the larger contigs were binned based on the GC content and read coverage, allowing the recovery of nearly complete genomes of *Ferroplasma* type II and *Leptospirillum* group III. These organisms had never been cultivated. With these genome sequences, the phylogenetic origin of these organisms could be inferred, as well as their relative dominance across similar samples. Based on gene annotation, the authors suggested metabolic functions across genomes and inferred ecological cross-feeding interactions between organisms involved in the community's carbon and nitrogen cycles [53].

Narasingarao et al. [93] obtained scaffolds from Sanger sequencing of size-fractionated samples from a hypersaline lake in Victoria, Australia. By binning these scaffolds based on the GC content, they found two distinctive GC profiles from very small cells that passed a 0.8 μm filter but were retained at 0.1 μm pore sizes. Using a phylogenetic binning approach, they recovered two draft uncultured genome sequences, which were representatives of a totally new branch of uncultured *Halobacteria*. Nearly 60 % of the predicted genes in these archaea had no homology with proteins in Genbank and they exhibited a very distinctive codon usage profile when compared to other archaea [93]. Although most genes in these microorganisms were unknown, the fraction of annotated genes suggested a predominantly aerobic heterotrophic lifestyle and also the presence of a complete pentose phosphate pathway, which had not previously been found in archaea [94]. These genomes were compared with other databases, suggesting that these archaea belong to a new, widespread class for which the authors coined the name "*Nanohaloarchaea*". At least eight distinct clades of this class have been found in hypersaline environments across different continents [93].

In a recent article, Spang et al. [89] reconstructed three partial uncultured archaeal genome sequences from marine sediment metagenomes, which together comprise the *Lokiarchaeota*, a candidate archaeal phylum that, based on phylogenomic analyses, encompasses the base of all eukaroytes. Comparative genomic analyses of the

uncultured genome sequences identified eukaryotic signature genes, including genes that are involved in membrane remodeling and vesicular trafficking. Based on these genomic observations, the authors proposed that the uncultured *Lokiarchaeota* contain a complex cellular machinery that may have facilitated the acquisition of the proto-mitochondrial endosymbiont into the ancestor of all eukaryotes. This example from the field of evolutionary biology highlights that metagenomic discovery of uncultured genome sequences can impact all areas of biology and is not limited to microbial ecology.

The new taxonomic groups identified by metagenomics can be vast. In a recent study, Brown et al. [34] assembled 8 complete and 789 draft genome sequences from tiny uncultured bacteria in size fractioned samples from an aquifer adjacent to the Colorado River. These genomes are members of a new super-phylum of at least 35 different bacterial phyla that was estimated to encompass 15 % of the bacterial domain [34]. Phylogenetic evidence suggests that this phylum forms a monophyletic group, which the authors named the candidate phyla radiation (CPR). Analysis of these uncultured genomes revealed many unusual features. For example, several nearly universal ribosomal genes [95] were absent from many draft genomes, such as rpL9 that was not detected in any of the 16 uncultured genome sequences from the WS6 candidate phylum [34]. Although the uncultured genome sequences were estimated to be only $\geq$50 % complete (median completeness 91 % for WS6), the authors suggest that it is highly unlikely that all draft genome sequences lack the same gene by chance. Moreover, analysis of the 16S sequences of these uncultured genomes revealed the widespread presence of large introns within the 16S rRNA genes. It was suggested that the commonly used primers for 16S amplicon sequencing would miss a large fraction of these bacteria due to primer mismatching and the presence of these introns [34].

It may be expected that similar metagenomic investigations into the vast, uncultured microbial biosphere, including archaea and viruses that remain poorly represented in current databases, will yield many new and exciting discoveries in the near future.

Minority groups

It is important to realize that the uncultured genome sequences obtained from metagenomes represent consensus sequences of closely related genomes [65]. If there are multiple highly similar strains within a sample, metagenome assembly approaches tend to collapse these genotypes into a single consensus sequence. Indeed, most genome sequences that are available today represent consensus genome sequences, including the reference genomes of many bacteria and animals. For most applications this is sufficient and allows firm conclusions to be drawn. However, some applications may require genotypes of individuals, for example in population genomics, and an alternative to obtain uncultured genome sequences of such individual genotypes is to perform single-cell sequencing [15, 96, 97]. In this approach, single-cells are separated by cell sorting, their genomic content is randomly amplified by multiple displacement amplification (MDA) that exploits the phage-derived Φ29 DNA polymerase and random short primers, and subsequently sequenced. Several groups have used single-cell sequencing combined with metagenomics to simultaneously obtain consensus sequences and individual genotypes [98–100]. Like with genome assembly from metagenomes, the completeness of single cell genomes can vary widely from <10 to 98 % [15, 96–98, 100].

The identification of minority groups that are underrepresented in the community, and thus in the bulk of shotgun metagenomic sequencing reads is a challenge when identifying uncultured genome sequences in metagenomes. Single cell sequencing may not be a good approach in these cases because isolation strategies tend to favor the majority, although the identity of cells can be determined by using probes before sequencing [101–103]. The genome sequences of some minority members from a marine community were recovered by using mate-paired reads sequenced on a SOLiD platform [82]. In this study, 58.5 Gb of mate-paired reads were generated and assembled into contigs. The mate-pairing information was used to link the contigs into interconnected graphs, and oligonucleotide usage profiles and read-coverage statistics were used to bin the contigs into larger linear scaffolds. Several candidate genomes were assembled with this approach, including a member of uncultured group II *Euryarchaeota*, whose genome indicated that this microbe is photoheterotrophic with aerobic metabolism and the ability to degrade lipids and proteins [82]. Other uncultured genomes of this group were later sequenced and assembled from metagenomic fosmid clones, confirming similar features [104]. This approach of deep mate-paired sequencing combined with partial-assembly and binning based on compositional features was also used to assemble 15 draft genomes from samples enriched for biomass-degrading microbes from cow rumen [99]. The completeness of one of these genome sequences was assessed by single cell sequencing, showing that a significant part of the genome was present in the original draft assembly and that no spurious reads had been incorporated. These results demonstrate the validity of this assembly pipeline to produce draft genomes of minority groups within the microbial community.

A similar approach for obtaining the uncultured genome sequences of rare minority groups uses binning based on the relative depth of coverage of fragments from two different DNA extractions of the same sample [64]. This approach was followed by principal component analysis of tetranucleotide usage profiles, and information from paired-end reads were used to isolate 13 nearly complete genomes, including four rare genomes (0.06–1.58 % relative abundance) of uncultured representatives of the TM7 phylum [64].

Data recycling

In the examples above, metagenomic datasets were newly sequenced and analyzed to discover species in environments that were of particular interest to the researchers. Due to the invaluable efforts of these and other research groups, many metagenomes are now becoming available in the public databases that can be used in secondary analyses. Public databases [105, 106] now contain thousands of metagenomic datasets that can be mined for novel microbial and viral genome sequences. The opportunity for data recycling is strongly driven by the development of new bioinformatic tools and methods for metagenomic analysis. We turn now to some significant examples of uncultured genome sequences that were obtained from recycled datasets.

Cross-assembly (crAss) of different samples from similar environments is one example of a strategy that can point to co-occurring sequences that are shared between environments and may not be identified with other methods such as reference mapping [67]. Our group cross-assembled previously published viral metagenomes of human fecal samples from four homozygotic female twin pairs and their mothers, and found a previously unknown viral sequence that was highly prevalent in human gut microbiomes from different continents, named crAssphage [63]. Up to 24 % of the viral shotgun metagenomic sequencing reads in samples from Korea, and up to 22 % of the reads in unrelated total fecal community metagenomes from USA aligned to the crAssphage genome sequence. The complete genome assembly and the metagenomic context in which it was isolated allowed the prediction of candidate host species, suggesting that it may infect *Bacteroides* hosts.

An alternative approach to analyze multiple metagenomic datasets was used to extract co-abundance gene groups (CAGs) from 396 gut metagenomes [107]. In this approach, metagenomes were first assembled and genes extracted to create a comprehensive non-redundant gene catalog of almost four million gut microbial genes. Genes were then picked randomly, and the abundance profiles across the 396 gut metagenomes of all other genes was compared to the query gene by using Pearson correlation. Highly correlating genes ($r > 0.9$) were iteratively grouped into CAGs, and their abundance profiles averaged until the CAG stabilized. The size distribution of CAGs showed a bimodal distribution with peaks at approximately 50 and 1700 genes, respectively. The CAGs that contained more than 700 genes were re-assembled, and 238 of those yielded genome sequences that met the criteria for high-quality draft genome sequences as defined by the Human Microbiome Project. A total of 181 of these uncultured genome sequences were derived from species that had no previously sequenced representative. Many of the smaller CAGs, potentially representing bacteriophages and mobile genomic elements such as plasmids or integrons, were observed to be dependent on the large CAGs, i.e. they were only present in the samples if the larger CAG was also present [107].

Metagenomics and omics-related approaches are increasingly advancing fields ranging from human and veterinary medicine, to microbial ecology and evolutionary biology. The availability of data and new analytic approaches not only provides new uncultured genome sequences as discussed above, but also enables the characterization of novel clades of archaea, bacteria, and viruses. Identifying the genome sequence of an uncultured organism allows us to ask questions about its diversity, genomic evolution, preferred environments, relative abundances, and co-occurrence with other species. For example, a recently published web tool, Phage Ecol-Locator, allows the investigation of bacteriophage genes across environments in order to answer questions about phage biology, lifestyle, and ecology [108]. These and other questions can be addressed by leveraging publicly available metagenomic datasets. We expect that new tools for metagenomic data recycling will increasingly become available to exploit the knowledge contained in large public databases, with the potential to describe the identity, evolution, and ecological interactions of cultured, as well as uncultured microbes and viruses.

Top-down approaches to study uncultured genome sequences

Metagenomes can be studied by using top-down and bottom-up approaches. Top-down approaches are based on metagenome-wide statistical patterns that are obtained from the sequence fragments of metagenomic reads, and can, for example, be used to study the structure of the ecosystem, as well as the identity and relative abundances of microorganisms [109]. Bottom-up approaches begin from flexible pre-defined structures of the system, such as genome-scale metabolic models and aim to mechanistically

reconstruct patterns and signals that can be measured from the system as a whole by integrating its constitutive parts into a model [110]. Bottom-up approaches will be discussed in a further section.

Obtaining a metagenomic sample, i.e. a random, minimally biased sample of the genomic sequence content of a microbial community, allows for direct and statistical estimates of ecological and evolutionary variables that help explain the structure and function of the microbial ecosystems [78, 111]. With more and better metagenomic data becoming available from sites across the planet, there is an unparalleled wealth of data available in the digital space for scientists to generate, test, and evaluate new hypotheses about microbial ecosystems [112]. Examples of ecological and evolutionary parameters that can be studied in metagenomic datasets include microbial species abundances, richness, evenness, and diversity [113, 114]. Moreover, eco-evolutionary processes can be studied, including competition, cooperation [115, 116], Red Queen dynamics [117, 118], structure and function of communities, as well as patterns of assembly, colonization, and composition of the microbiota [119–121]. Below we outline some of these patterns and emphasize that metagenomics provides not only a comprehensive window to discover and isolate new uncultured genome sequences as outlined above, but also provides the principal data to characterize the ecological context in which these genomes are found.

Global abundance and distribution patterns

The ecological context of uncultured organisms can be studied by exploiting metagenomic datasets. Many discoveries in this young field have changed established textbook frameworks of microbial relationships with the earth's physics and chemistry, revealing a less biased view of the structure and function of microbial ecosystems. Light harvesting in the ocean is one example where non-chlorophyll pathways based on bacteriorhodopsin were shown by metagenomics to be a widespread mechanism in the ocean, not only limited to *Proteobacteria* or *Archaea* [54, 122]. Another example is the elucidation of the biogeography and ecology of specific uncultured microbial groups. For example, a group of archaea, (previously called *Crenarchaeota* because of a somewhat close relationship with this phylum [123, 124] but now known as *Thaumarchaeota*, see below) was found by metagenomics to be present in many different environments, such as freshwater [125], sediments [126], ocean water [54], and the digestive tract of aquatic and terrestrial animals [127, 128]. One representative was cultivable in a marine aquarium when grown as a symbiont to the sponge *Axinella mexicana* [127]. Several genomic surveys and later the cultivation of one marine representative of this phylum showed that many of these species encoded ammonia-oxidizing genes [129, 130]. Given the abundance of this phylum in several environments, they have recently been suggested to be major players in the global cycling of nitrogen through ammonia oxidation [131]. Before this group was discovered, ammonia oxidation was thought to be performed almost exclusively by autotrophic ammonia oxidizing bacteria [132]. Later, the assembly of several uncultured genomes and genomic evidence from different sequencing projects rooting this group further apart from the *Crenarchaeota*, led to the recognition of a new archaeal phylum, the *Thaumarchaeota* [133].

Niche-driven and neutral community assembly

Metagenomic data can be used to determine the mode of assembly of a microbial community. Processes of assembly are relevant to the study of community ecology because they indicate which forces have shaped biological communities and likely influence their structure and function [134]. Two different types of processes are commonly distinguished that shape the composition of microbial ecosystems: deterministic niche-driven, and stochastic neutral processes [135]. Both processes, and combinations thereof, can predict the distribution curve of the relative abundances of species. If a neutral stochastic process has shaped the community, the relative abundances of species are expected to fit a zero-sum multinomial (ZSM) distribution [136, 137]. In the niche-driven process, species are related to environmental changes and the relative abundances are expected to fit a log-normal or a zipf distribution [138]. In the healthy lung, for example, the composition of the microbiota was shown to fit a neutral model with species derived mainly from the oral cavity, while samples from the lungs of patients with cystic fibrosis and idiopathic interstitial pneumonia could not be explained by the neutral model [139]. Mendes et al. [140] compared soil and soybean rhizosphere microbiomes and found a log-normal distribution in the rhizosphere community, while the bulk soil community fit the ZSM distribution. Metagenomics has provided evidence of niche-driven or neutral-processes in several other environments [141–143].

Biodiversity and ecosystem stability

Biodiversity is another important ecological parameter that can be measured by top-down metagenomics. Biodiversity can be defined as the species richness, i.e. the number of different species that are present in an environment; as the relative abundances of the different species; or as the evenness, a measure that incorporates the phylogenetic breadth of the species [144]. Biodiversity is often related to

the stability of an ecosystem [145]. This is the basis of the insurance hypothesis, in which greater diversity insures ecosystems against losses of functionality due to environmental fluctuations and perturbations [146, 147]. Uncultured bacterial and archaeal genomes can be readily inserted into a biogeographic and evolutionary context by comparing their marker genes across these datasets. Data for species richness in microbial ecosystems based on marker genes provides a wide spectrum of information about their distribution patterns, as well as the alpha and beta diversity, and can shed light on migration and colonization patterns [148].

The relative abundance of functional categories of genes in a microbial ecosystem is an alternative parameter of biodiversity, which can be related to the concept of evenness if one assumes that phylogenetic distance is correlated with functional distance [149]. Note that it is not necessary to make this assumption when analyzing shotgun metagenomes because the relative abundance of different categories of genes can be directly measured. When the phylogenetic and functional measures of biodiversity are compared, very complex interplays between stability and environmental functioning can be revealed, providing the starting material to evaluate and test hypotheses about the ecological role of uncultured genomes obtained from metagenomes. An interesting example of the potential of metagenomics to simultaneously discover new species and provide a broad description of their ecology and natural history is provided in a recent study by Lynch et al. [150]. The authors characterized an uncultured genome sequence obtained from metagenomic data of a volcanic deposit collected 6 km above the sea level in the Atacama Desert. Their study suggested that this uncultured bacterium was indigenous to this harsh environment, with a chemoautotrophic metabolism dependent on trace atmospheric gases [150].

The ecological concept of ecosystem stability is related to biodiversity, and it can be interpreted and measured in different ways [151]. For example, Wittebolle et al. [152] measured the relationship between evenness and stability in different microcosm experiments with denitrifying bacteria. In their study the microcosms were subject to temperature and salt stress, and the stability of the microbial ecosystem was measured as the maintenance of the nitrifying function under stress. The authors showed that the effect of stress on functional stability differed depending on the kind of stress, and that microbial communities with an even functional profile tended to be more resilient to salt induced stress than functionally uneven communities [152]. In the human microbiome, which has become one of the best studied microbial ecosystems, widely different taxonomic compositions have been observed to lead to very similar functional profiles across individuals [153]. This observation of a functional stability supports the insurance hypothesis, being driven by the potential of phylogenetically divergent gut bacteria to acquire similar functions [154, 155]. The relationship between stability and biodiversity is an open research field in microbial community ecology. Top-down metagenomics is providing the means to study this relationship across many different microbial ecosystems, particularly through studies that analyze fluctuations of the taxonomic and functional profiles of communities in space and time [156–159].

Integrating uncultured genome sequences into a systems biology modeling platform

While the top-down statistical approaches described above provide fundamental information to understand the distribution and ecology of uncultured microorganisms and viruses, they are limited to providing broad-scale predictions that are not always mechanistic. The predictive power of such statistical models can be improved by including more omics data from an environment, such as gene expression, proteomics, and metabolite concentrations [156, 160, 161]. Furthermore, incorporating time series datasets or environmental data such as physicochemical parameters can also contribute to more mechanistic and predictive models [162]. However, a deeper understanding of the biology of new uncultured genomes would come from mechanistic descriptions of the dynamics and biochemical interactions of each subpopulation [163, 164]. Such bottom-up approaches employ computational models to identify robustly predicted patterns in an ecosystem that can subsequently be studied ex silico, for example by exploiting metagenomic datasets. Progress in building genome-scale models for small microbial consortia is beginning to provide a roadmap for describing microbial communities in terms of their individual sub-populations. Below we will discuss several approaches for integrating uncultured genome sequences into computational models, towards describing and understanding the interactions that shape a microbial ecosystem.

Computational models of microbial cells

The most complete computational model of a cell that integrates several components of the cellular dynamics, such as protein synthesis, and gene expression, was built by Karr et al. [165] for *Mycoplasma genitalium*. This model describes a single organism and reconstructs several patterns of the bacterial cell cycle that are consistent with measurements in vitro [165]. Whole cell models with such level of detail are not currently feasible for most microbes

because the roles of novel genes, poorly characterized proteins, and kinetic enzyme parameters remain unknown. Nevertheless, draft biochemical models that propagate and integrate knowledge from known genes that are characterized in other organisms already show significant potential to predict and explain patterns observed in experimental systems [166, 167].

Several different modeling approaches exist that build mechanistic metabolic models of a microbial cell by starting from the genomic sequences, but are beyond the scope of this review [168]. Here we will only point to some of the general principles and possible directions to build predictive models of uncultured genome sequences, and address their role in the community. Our goal is to highlight directions that will position these newly discovered genomes on in silico experimentation platforms. This will accelerate the characterization of these organisms by providing the means to quantitatively describe their interactions with other microbes and the environment, and guide experimental follow-up by providing testable hypotheses about species interactions and their responses to environmental changes.

Models based on individual genome sequences

When uncultured genome sequences are recovered from an environment by using e.g. metagenomics or single-cell sequencing, the component of their genes that can be annotated can be integrated into a basic biochemical model of directional interactions between proteins and metabolites (for a review of these steps see Refs. [167, 169]). If we assume that several of these models can be inferred for microbes that co-occur within an environment, an important feature that describes their interaction are the exchange reactions that reflect the flow of metabolites in and out of cells. Moreover, the rate by which the cells synthesize biomass components, and the flow of byproducts and secondary metabolites that leave the cell can also be captured. Such metabolic flow models might be used to make predictions about which species grows faster in a given environment [170], the secretion of a products of interest under given conditions [171], the expected biochemical effect of adding or removing a species or metabolite [172], as well as the conditions of the external environment that are required for (mutual) growth [173].

The flow of metabolites

While some of the information about the metabolic flows can be assessed from the biochemical networks, these networks do not contain information about the kinetic rates of uptake, secretion, and the flow of the metabolites, nor do they contain information about the rates of biomass conversion. In practice, and especially for novel species that contain many unknown genes, we can only reconstruct partial blueprints of the biochemical networks [174]. This suggests that the real flow of metabolites between the organisms consists of complex functions that integrate protein concentrations and affinities, resulting in different reaction rates [175]. Another challenge is capturing the simultaneous reactions from many different biochemical networks within a single model that could contain multiple solutions. Thus, comprehensive models of microbial communities based on individual metabolic networks are not yet available.

Tackling the complexity of microbial communities

Small scale models of interacting consortia of few microbes are paving the way for applications to larger communities [171–173, 176–178]. Three important general principles may be extracted from these studies and applied to larger-scale models (Fig. 3). First, the multi-dimensional attractor landscape should be constrained to reduce the degrees of freedom of the solution-space. Second, optimization approaches should be applied to deal with multiple solutions. Third, computational simulations should be used instead of analytical approaches to sample from the possible solution space of multi-level models.

As explained above, the functional insurance hypothesis suggests that there are different possible solutions to how microbial communities may fulfill an environmental niche. In terms of modeling the microbial ecosystem, this can be thought of as different domains of attraction of a highly-dimensional system. This system is subject to important constraints that need to be incorporated in the model. For example, there are hard constraints like the stoichiometric balance of chemical reactions between the metabolites and the second law of thermodynamics, but there are also softer constraints like the spatial boundaries of the system and the diffusion of metabolites that may be captured by stochastic models. Integrating these constraints into systems biological models of the microbial ecosystem can significantly reduce the degrees of freedom of the system, therewith constraining the landscape of its domains of attraction (Fig. 3c). A further way to constrain these models would be to use additional omics datasets to assess gene expression and/or metabolite concentrations [179, 180]. However, even with a constrained landscape of solutions, models of interacting microorganisms could potentially hold an infinite number of solutions. To deal with this degeneracy of solutions, a heuristic approach can be applied that identifies local optima within the attractor landscape that represent biologically meaningful solutions [181, 182].

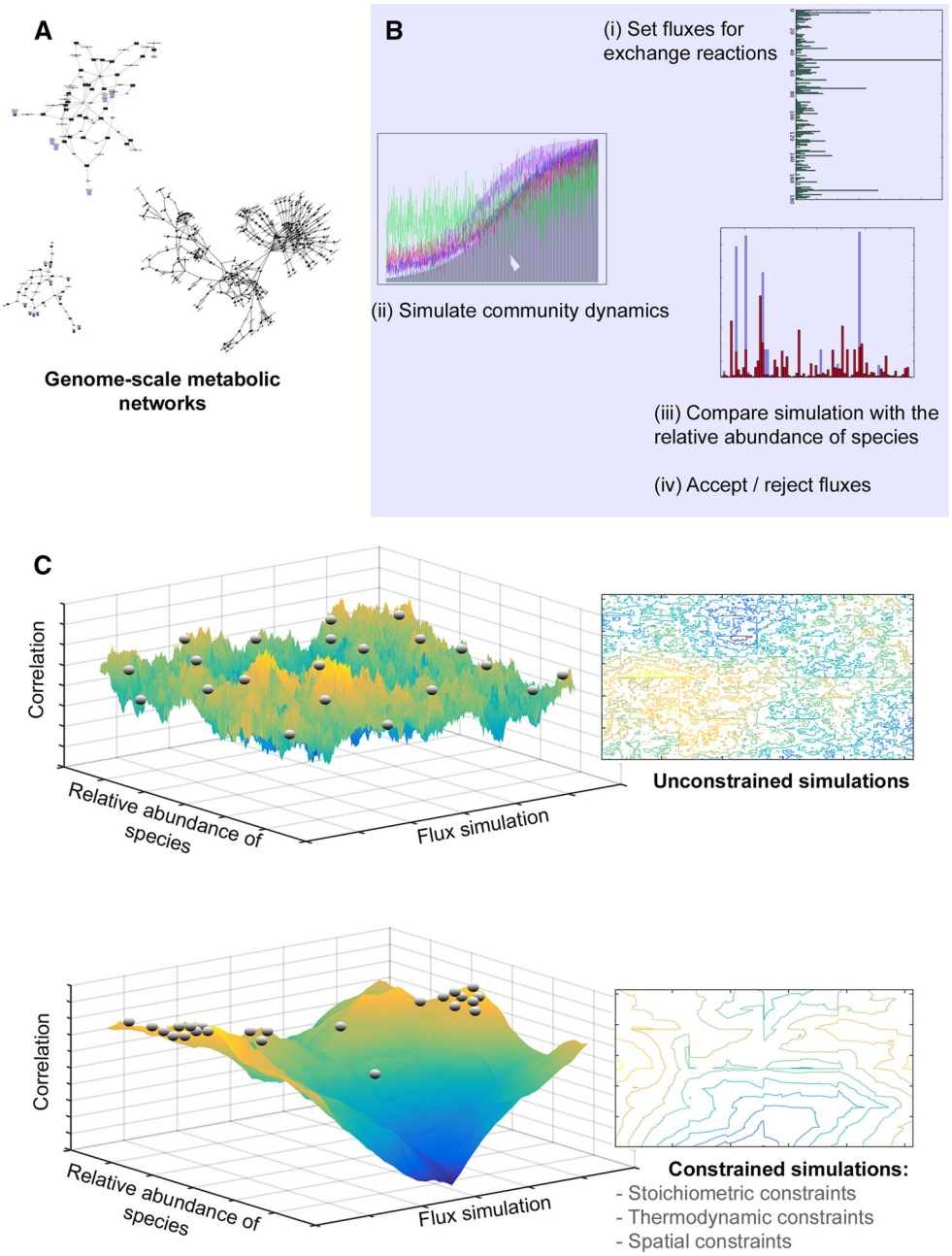

Fig. 3 Theoretical representation of the guidelines to build genome-guided simulation-based models for microbial communities applied to a simple model. **a** The model was built for a hypothetical community of biochemical networks corresponding to uncultured genomes. **b** In this model, the variable of interest is the flow-rate of metabolites through exchange reactions in steady-state conformations. Random initial flow-rates were chosen and the growth of the community in a media containing this concentration of metabolites is simulated as in [178]. After equilibrium is reached, the relative abundance of each species is compared to the actual relative abundance from the metagenomic data-set. New values for exchange flow-rates are chosen and simulated, and accepted or rejected according to a stochastic rule or if the predicted relative distribution of species is closer to its actual value. **c** Simulations with or without constraints significantly reduce the solution landscapes indicated by the contour *plots*. The correlations are also significantly higher and have a small number of high-correlation solutions, which can be further studied individually

Objective functions

Different biological objectives can be defined and expressed as functions in a system of equations with a goal to maximize or minimize this objective, including the objectives that are used in single-species systems [183]. Moreover, approaches to model multiple objectives within a single model have also been explored [184]. The mathematical formulation of a reasonable objective function allows for the optimization of the system for this objective, and depending of the relation of this objective with other variables, the optimization may limit the values and states that may be assumed by the other variables in the system [183]. For example, in a genome-scale model of three gut bacteria, Shoaie et al. [185] used as an objective function the minimization of the uptake of nutrients while maintaining fixed concentrations of biomass. By setting up this configuration, they accurately predicted the concentration of butyrate, CO_2, and H_2 obtained from experimental data of germ-free mice colonized with these bacteria [185].

Optimization of objectives in simple systems, such as single-species models, is a straightforward process that usually involves minimizing or maximizing an objective function, while constrained by systems of linear, mixed

integer-linear, or simple nonlinear equations. However, optimizing multiple and potentially different objectives from many interacting species that grow at different rates and consume and secrete metabolites at the same time is a significantly more challenging problem. Some of the studies yielding the most promising results have applied approaches that were based on simulating the system, rather than solving it [178, 186]. In simulation-based approaches, the current state of the system is sampled and transition rules are applied that determine its state in the next time point. The system is updated based on these rules and sampled again; this goes on until the system stabilizes in a pattern or distribution.

Models of microbial consortia: linking to experiments

Using a simulation-based approach for pairs of species, Chiu et al. [178] coupled metabolic networks to Michaelis–Menten dynamics for exchange reactions of the metabolites across the cell membrane. In small time steps, each species would take up, and secrete metabolites proportionally to its biomass and the concentration of the metabolite in the medium. The medium and the biomass of each species were then updated and simulated again, until metabolites were depleted and the growth-rates became zero. This approach predicted the relative abundances of the two bacteria, their temporal growth-rates, and the dynamics of metabolites inside and outside of the cells [178]. A similar approach was used by Harcombe et al. [186], with the addition that they incorporated a spatial lattice into the model where all species could diffuse stochastically. This framework consistently predicted the rate of colony diameter increase in various carbon sources for *E. coli*, as well as the outcome of co-culture experiments of two and three species. Interestingly, an unexpected emergent behavior of the in silico model was confirmed experimentally, showing that the species with the lower growth-rate dominates the co-culture in the long run.

Linking uncultured viruses to their cellular hosts

Viruses necessarily depend on a cellular host organism for replication, and these virus-host associations can be very specific. Until recently, virus discovery involved isolation of the virus, e.g. by using cell culture or plaque assays, leading to a clear link between a virus and its host. However, with the advent of metagenomic approaches to identify the uncultured viral genome sequences, as described above, virus discovery is no longer dependent on culturing. New bioinformatic approaches are being explored to link viruses to their hosts, based on the information contained in their uncultured genome sequences (Edwards et al., submitted). Signals for virus-host association that have been used in recent studies include the co-occurrence profiles across samples, as described above [63, 107, 187]. Moreover, homology between virus and host genes can indicate a recent gene exchange between their genome sequences, possibly during a recent infection event, and thus homology has also been used to identify virus hosts [63, 188]. For bacteria and archaea, CRISPR spacers that are identified within their genomes can be used to identify the phages that infect them [187, 189], because short fragments from the phage genome sequence are incorporated into CRISPR arrays of the host. Finally, oligonucleotide usage profiles also contain a signal that can be exploited to link an uncultured virus to its cellular host. This depends on viruses ameliorating their genomic oligonucleotide usage to that of the host they infect, for example to avoid recognition by host restriction enzymes, or to adjust their codon usage to match the availability of host tRNAs [190, 191].

Linking uncultured viral genome sequences to a cellular host organism, cultured or uncultured, is an important step towards understanding the microbial ecosystem. Phage-bacterial infection networks (PBIN) describe which phages infect which bacterial hosts [192]. A recent meta-analysis of PBIN showed a characteristic structuring that is globally modular and locally nested [193, 194]. This means that bacteria and phages from different locations are mostly incompatible (global modularity). Within one location, phages co-exist with varying host specificity (local nestedness), e.g. generalist phages that infect many bacteria, and specialist phages that infect only one bacterium. Phage predation can have a huge impact on microbial ecology, maintaining biodiversity through Kill-the-Winner dynamics [195], and releasing nutrients through the viral shunt [196]. Incorporating phage predation into ecosystem models will allow the effects of this important parameter in microbial ecology to be studied [196, 197].

Conclusions

Obtaining the genome sequences of uncultured microbes and viruses in metagenomes is one of the most promising areas of research in microbiology. Novel strategies to sample and sequence environmental metagenomes as well as significant advances in bioinformatics and data recycling are increasing our knowledge of uncultured microorganisms. With metagenomic approaches, we can discover the identity, evolution, gene composition, distribution, and ecological patterns of uncultured microbes and viruses. Our challenge now is to integrate this knowledge into predictive analytical models of microbial ecosystems that incorporate the knowledge that can be mined from both uncultured and

cultured genome sequences [163]. It is still difficult to realistically capture important properties of microbial ecosystems in analytical models, such as spatial structuring, diffusion of nutrients, energy barriers, selective sweeps by bacteriophages, and the immune system in case of host-associated microbiota. Recent progress has shown that the way forward is to apply modeling through multi-step simulation-based approaches. Although there are still many caveats to these approaches, we believe that future development in this area will provide outstanding tools to mechanistically understand the biology of uncultured microbes. Some of the variables that could be predicted by these models and experimentally validated are energy flux patterns, cross feeding patterns, and the dynamics of diversity within the community of study. If a community is described in terms of energy and matter flow, it can also be compared in these terms, providing not only a unique insight into the evolutionary processes that have shaped microbial communities, but also informing in a precise and mechanistic manner how these balances could be changed, or how changes in these balances impact biodiversity. Systems biology platforms with these potentials are the immediate goals for further advances in discovering and understanding the microscopic and submicroscopic biosphere. The major remaining challenges include providing the expanding number of sequences available with reliable annotations, and incorporating these into consistent models of interacting microbes and viruses in the natural ecosystem. To conclude, the exciting field of uncultured microbe and virus discovery, and the study of interactions in natural microbial ecosystems has grown with metagenomics throughout the past decade, and recent developments hold promise of many more discoveries in the near future.

Acknowledgments This work was supported by CAPES/BRASIL. DRG is supported by CNPQ/BRASIL. The authors thank Noriko Cassman for assistance with the proof-reading and editing of the manuscript.

References

1. Dutilh BE (2014) Metagenomic ventures into outer sequence space. Bacteriophage 4:e979664. doi:10.4161/21597081.2014.979664
2. Marx CJ (2013) Can you sequence ecology? Metagenomics of adaptive diversification. PLoS Biol 11:e1001487. doi:10.1371/journal.pbio.1001487
3. Antony van Leeuwenhoek and his "Little animals"; being some account of the father of protozoology and bacteriology and his multifarious discoveries in these disciplines: Dobell, Clifford, 1886–1949: Free Download & Streaming. In: Internet Arch. https://archive.org/details/antonyvanleeuwen00dobe. Accessed 15 Jul 2015
4. Madigan MT, Martinko JM, Bender KS, Buckley DH, Stahl DA, Brock T (2014) Brock biology of microorganisms, 14 edn. Benjamin Cummings, Boston
5. Paccini F (1854) Gazzetta medica italiana: federativa toscana
6. Gram H (1884) Über die isolierte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten. Fortschr Med 2:185–189
7. Fite GL, Wade HW (1955) The contribution of Neisser to the establishment of the Hansen bacillus as the etiologic agent of leprosy and the so-called Hansen-Neisser controversy. Int J Lepr 23:418–428
8. Ben-David A, Davidson CE (2014) Estimation method for serial dilution experiments. J Microbiol Methods 107:214–221. doi:10.1016/j.mimet.2014.08.023
9. Mokili JL, Rohwer F, Dutilh BE (2012) Metagenomics and future perspectives in virus discovery. Curr Opin Virol 2:63–77. doi:10.1016/j.coviro.2011.12.004
10. Weinbauer MG (2004) Ecology of prokaryotic viruses. FEMS Microbiol Rev 28:127–181. doi:10.1016/j.femsre.2003.08.001
11. Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39:321–346. doi:10.1146/annurev.mi.39.100185.001541
12. Razumov AS (1932) The direct method of calculation of bacteria in water. Comparison with the Koch method. Mikrobiol 1:131–146
13. Jannasch HW, Jones GE (1959) Bacterial populations in sea water as determined by different methods of enumeration1. Limnol Oceanogr 4:128–139. doi:10.4319/lo.1959.4.2.0128
14. Jones JG (1970) Studies on freshwater bacteria: effect of medium composition and method on estimates of bacterial population. J Appl Bacteriol 33:679–686. doi:10.1111/j.1365-2672.1970.tb02250.x
15. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437. doi:10.1038/nature12352
16. Breitbart M (2012) Marine viruses: truth or dare. Annu Rev Mar Sci 4:425–448. doi:10.1146/annurev-marine-120709-142805
17. Stewart EJ (2012) Growing unculturable bacteria. J Bacteriol 194:4151–4160. doi:10.1128/JB.00345-12
18. Kaeberlein T, Lewis K, Epstein SS (2002) Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. Science 296:1127–1129. doi:10.1126/science.1070633
19. Bollmann A, Lewis K, Epstein SS (2007) Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. Appl Environ Microbiol 73:6386–6390. doi:10.1128/AEM.01309-07
20. Onofrio AD, Crawford JM, Stewart EJ, Witt K, Gavrish E, Epstein S, Clardy J, Lewis K (2010) Siderophores from neighboring organisms promote the growth of uncultured bacteria. Chem Biol 17:254–264. doi:10.1016/j.chembiol.2010.02.010
21. Graber JR, Breznak JA (2005) Folate cross-feeding supports symbiotic homoacetogenic spirochetes. Appl Environ Microbiol 71:1883–1889. doi:10.1128/AEM.71.4.1883-1889.2005
22. Nichols D, Lewis K, Orjala J, Mo S, Ortenberg R, O'Connor P, Zhao C, Vouros P, Kaeberlein T, Epstein SS (2008) Short

peptide induces an "uncultivable" microorganism to grow in vitro. Appl Environ Microbiol 74:4889–4897. doi:10.1128/AEM.00393-08
23. Tripp HJ, Kitner JB, Schwalbach MS, Dacey JWH, Wilhelm LJ, Giovannoni SJ (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. Nature 452:741–744. doi:10.1038/nature06776
24. Kalmbach S, Manz W, Szewzyk U (1997) Isolation of new bacterial species from drinking water biofilms and proof of their in situ dominance with highly specific 16S rRNA probes. Appl Environ Microbiol 63:4164–4170
25. Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, Mueller A, Schäberle TF, DE Hughes, Epstein S, Jones M, Lazarides L, Steadman VA, Cohen, Felix CR, Fetterman KA, Millett WP, Nitti AG, Zullo AM, Chen C, Lewis K (2015) A new antibiotic kills pathogens without detectable resistance. Nature 517:455–459. doi:10.1038/nature14098
26. Stokell JR, Steck TR, Todd R (2012) Viable but nonculturable bacteria. In: eLS. Wiley, Chichester. doi:10.1002/9780470015902.a0000407.pub2
27. Davis IJ, Bull C, Horsfall A, Morley I, Harris S (2014) The Unculturables: targeted isolation of bacterial species associated with canine periodontal health or disease from dental plaque. BMC Microbiol 14:196. doi:10.1186/1471-2180-14-196
28. Bochner BR (2009) Global phenotypic characterization of bacteria. FEMS Microbiol Rev 33:191–205. doi:10.1111/j.1574-6976.2008.00149.x
29. Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, Kanigan T, Lewis K, Epstein SS (2010) Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. Appl Environ Microbiol 76:2445–2450. doi:10.1128/AEM.01754-09
30. Omsland A, Cockrell DC, Howe D, Fischer ER, Virtaneva K, Sturdevant DE, Porcella SF, Heinzen RA (2009) Host cell-free growth of the Q fever bacterium *Coxiella burnetii*. Proc Natl Acad Sci USA 106:4430–4434. doi:10.1073/pnas.0812074106
31. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420:806–810. doi:10.1038/nature01240
32. Rappé MS, Connon SA, Vergin KL, Giovannoni SJ (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418:630–633. doi:10.1038/nature00917
33. Harris JK, Kelley ST, Pace NR (2004) New perspective on uncultured bacterial phylogenetic division OP11. Appl Environ Microbiol 70:845–849. doi:10.1128/AEM.70.2.845-849.2004
34. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208–211. doi:10.1038/nature14486
35. Luef B, Frischkorn KR, Wrighton KC, Holman H-YN, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG, Downing KH, Comolli LR, Banfield JF (2015) Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun. doi:10.1038/ncomms7372
36. Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, Thomas BC, Banfield JF (2013) Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. mBio 4:00708-13. doi:10.1128/mBio.00708-13
37. Hanke A, Hamann E, Sharma R, Geelhoed JS, Hargesheimer T, Kraft B, Meyer V, Lenk S, Osmers H, Wu R, Makinwa K, Hettich RL, Banfield JF, Tegetmeyer HE, Strous M (2014) Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between alpha- and gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat. Front Microbiol 5:231. doi:10.3389/fmicb.2014.00231
38. Rohwer F (2003) Global phage diversity. Cell 113:141
39. DeLong EF, Wickham GS, Pace NR (1989) Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. Science 243:1360–1363
40. Pace NR, Stahl DA, Lane DJ, Olsen GJ (1986) The analysis of natural microbial populations by ribosomal RNA sequences. In: Marshall KC (ed) Adv Microb Ecol. Springer, Berlin, pp 1–55
41. Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271
42. Schloss PD, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68:686–691. doi:10.1128/MMBR.68.4.686-691.2004
43. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369–394. doi:10.1146/annurev.micro.57.030502.090759
44. Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, Swings J (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60:407–438
45. Thompson CC, Amaral GR, Campeão M, Edwards RA, Polz MF, Dutilh BE, Ussery DW, Sawabe T, Swings J, Thompson FL (2015) Microbial taxonomy in the post-genomic era: rebuilding from scratch? Arch Microbiol 197:359–370. doi:10.1007/s00203-014-1071-2
46. National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications (2007) The new science of metagenomics: revealing the secrets of our microbial planet. National Academies Press (US), Washington (DC)
47. Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G (2013) High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods 95:401–414. doi:10.1016/j.mimet.2013.08.011
48. Bragg L, Tyson GW (2014) Metagenomics using next-generation sequencing. Methods Mol Biol Clifton NJ 1096:183–201. doi:10.1007/978-1-62703-712-9_15
49. Zhou J, He Z, Yang Y, Deng Y, Tringe SG, Alvarez-Cohen L (2015) High-Throughput metagenomic technologies for complex microbial community analysis: open and closed formats. mBio 6:e02288-14. doi:10.1128/mBio.02288-14
50. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5:R245–R249. doi:10.1016/S1074-5521(98)90108-9
51. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66:2541–2547
52. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA 99:14250–14255. doi:10.1073/pnas.202488399
53. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43. doi:10.1038/nature02340
54. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso sea. Science 304:66–74. doi:10.1126/science.1093857

55. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. Trends Genet 30:418–426. doi:10.1016/j.tig.2014.07.001
56. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632. doi:10.1038/nature06810
57. Trindade-Silva AE, Rua C, Silva GGZ, Dutilh BE, Moreira APB, Edwards RA, Hajdu E, Lobo-Hajdu G, Vasconcelos AT, Berlinck RGS, Thompson FL (2012) Taxonomic and functional microbial signatures of the endemic marine sponge *Arenosclera brasiliensis*. PLoS ONE 7:e39905. doi:10.1371/journal.pone.0039905
58. Sharon I, Banfield JF (2013) Genomes from metagenomics. Science 342:1057–1058. doi:10.1126/science.1247023
59. Karlin S, Mrázek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:3899–3913
60. Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. Annu Rev Genet 38:771–792. doi:10.1146/annurev.genet.38.072902.094318
61. Bailly-Bechet M, Danchin A, Iqbal M, Marsili M, Vergassola M (2006) Codon usage domains over bacterial chromosomes. PLoS Comput Biol 2:e37. doi:10.1371/journal.pcbi.0020037
62. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, Banfield JF (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science 337:1661–1665. doi:10.1126/science.1224041
63. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun. doi:10.1038/ncomms5498
64. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31:533–538. doi:10.1038/nbt.2579
65. Dutilh BE, Huynen MA, Strous M (2009) Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. Bioinforma Oxf Engl 25:2878–2881. doi:10.1093/bioinformatics/btp377
66. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C (2014) Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146. doi:10.1038/nmeth.3103
67. Dutilh BE, Schmieder R, Nulton J, Felts B, Salamon P, Edwards RA, Mokili JL (2012) Reference-independent comparative metagenomics using cross-assembly: crAss. Bioinforma Oxf Engl 28:3225–3231. doi:10.1093/bioinformatics/bts613
68. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. doi:10.7717/peerj.603
69. Kang DD, Froula J, Egan R, Wang Z (2014) A robust statistical framework for reconstructing genomes from metagenomic data. BioRxiv 011460. doi: 10.1101/011460
70. Beitel CW, Froenicke L, Lang JM, Korf IF, Michelmore RW, Eisen JA, Darling AE (2014) Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ 2:e415. doi:10.7717/peerj.415
71. Burton JN, Liachko I, Dunham MJ, Shendure J (2014) Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3 Bethesda Md 4:1339–1346
72. Marbouty M, Cournac A, Flot J-F, Marie-Nelly H, Mozziconacci J, Koszul R (2014) Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. eLife 3:e03318. doi:10.7554/eLife.03318
73. Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods 6:673–676. doi:10.1038/nmeth.1358
74. Horan K, Shelton CR, Girke T (2010) Predicting conserved protein motifs with Sub-HMMs. BMC Bioinform 11:205. doi:10.1186/1471-2105-11-205
75. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinform Oxf Engl 26:2460–2461. doi:10.1093/bioinformatics/btq461
76. Berendzen J, Bruno WJ, Cohn JD, Hengartner NW, Kuske CR, McMahon BH, Wolinsky MA, Xie G (2012) Rapid phylogenetic and functional classification of short genomic fragments with signature peptides. BMC Res Notes 5:460. doi:10.1186/1756-0500-5-460
77. Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. doi:10.1186/gb-2014-15-3-r46
78. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y-H, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5:e77. doi:10.1371/journal.pbio.0050077
79. Ghai R, Hernandez CM, Picazo A, Mizuno CM, Ininbergs K, Díez B, Valas R, DuPont CL, McMahon KD, Camacho A, Rodriguez-Valera F (2012) Metagenomes of Mediterranean Coastal Lagoons. Sci Rep. doi:10.1038/srep00490
80. Mizuno CM, Rodriguez-Valera F, Ghai R (2015) Genomes of planktonic acidimicrobiales: widening horizons for marine actinobacteria by metagenomics. mBio 6:e02083-14. doi:10.1128/mBio.02083-14
81. Mande SS, Mohammed MH, Ghosh TS (2012) Classification of metagenomic sequences: methods and challenges. Brief Bioinform 13:669–681. doi:10.1093/bib/bbs054
82. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. Science 335:587–590. doi:10.1126/science.1212665
83. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923
84. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. doi:10.1038/nmeth.3176
85. Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. Genome Res 13:693–702. doi:10.1101/gr.634603
86. Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. DNA Res Int J Rapid Publ Rep Genes Genomes 12:281–290. doi:10.1093/dnares/dsi015
87. Chan C-KK, Hsu AL, Halgamuge SK, Tang S-L (2008) Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9:215. doi:10.1186/1471-2105-9-215

88. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF (2009) Community-wide analysis of microbial genome sequence signatures. Genome Biol 10:R85. doi:10.1186/gb-2009-10-8-r85
89. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179. doi:10.1038/nature14447
90. Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW, Horn M, Daims H, Bartol-Mavel D, Wincker P, Barbe V, Fonknechten N, Vallenet D, Segurens B, Schenowitz-Truong C, Médigue C, Collingro A, Snel B, Dutilh BE, Op den Camp HJM, van der Drift C, Cirpus I, van de Pas-Schoonen KT, Harhangi HR, van Niftrik L, Schmid M, Keltjens J, van de Vossenberg J, Kartal B, Meier H, Frishman D, Huynen MA, Mewes H-W, Weissenbach J, Jetten MSM, Wagner M, Le Paslier D (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. Nature 440:790–794. doi:10.1038/nature04647
91. Mokili JL, Dutilh BE, Lim YW, Schneider BS, Taylor T, Haynes MR, Metzgar D, Myers CA, Blair PJ, Nosrat B, Wolfe ND, Rohwer F (2013) Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. PLoS ONE 8:e58404. doi:10.1371/journal.pone.0058404
92. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21:1616–1625. doi:10.1101/gr.122705.111
93. Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J 6:81–93. doi:10.1038/ismej.2011.78
94. Verhees CH, Kengen SWM, Tuininga JE, Schut GJ, Adams MWW, de VOS WM, van der OOST J (2003) The unique features of glycolytic pathways in Archaea. Biochem J 375:231. doi:10.1042/BJ20021472
95. Yutin N, Puigbò P, Koonin EV, Wolf YI (2012) Phylogenomics of prokaryotic ribosomal proteins. PLoS ONE 7:e36972. doi:10.1371/journal.pone.0036972
96. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM (2006) Sequencing genomes from single cells by polymerase cloning. Nat Biotechnol 24:680–686. doi:10.1038/nbt1214
97. Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, Saw JH, Senin P, Yang C, Chatterji S, Cheng J-F, Eisen JA, Sieracki ME, Stepanauskas R (2009) Assembling the marine metagenome, one cell at a time. PLoS ONE 4:e5299. doi:10.1371/journal.pone.0005299
98. Dupont CL, Rusch DB, Yooseph S, Lombardo M-J, Alexander Richter R, Valas R, Novotny M, Yee-Greenbaum J, Selengut JD, Haft DH, Halpern AL, Lasken RS, Nealson K, Friedman R, Craig Venter J (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J 6:1186–1199. doi:10.1038/ismej.2011.189
99. Hess M, Sczyrba A, Egan R, Kim T-W, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, Mackie RI, Pennacchio LA, Tringe SG, Visel A, Woyke T, Wang Z, Rubin EM (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331:463–467. doi:10.1126/science.1200387
100. McLean JS, Lombardo M-J, Ziegler MG, Novotny M, Yee-Greenbaum J, Badger JH, Tesler G, Nurk S, Lesin V, Brami D, Hall AP, Edlund A, Allen LZ, Durkin S, Reed S, Torriani F, Nealson KH, Pevzner PA, Friedman R, Venter JC, Lasken RS (2013) Genome of the pathogen Porphyromonas gingivalis recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform. Genome Res 23:867–877. doi:10.1101/gr.150433.112
101. Kalyuzhnaya MG, Zabinsky R, Bowerman S, Baker DR, Lidstrom ME, Chistoserdova L (2006) Fluorescence in situ hybridization-flow cytometry-cell sorting-based method for separation and enrichment of type I and type II methanotroph populations. Appl Environ Microbiol 72:4293–4301. doi:10.1128/AEM.00161-06
102. Podar M, Abulencia CB, Walcher M, Hutchison D, Zengler K, Garcia JA, Holland T, Cotton D, Hauser L, Keller M (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. Appl Environ Microbiol 73:3205–3214. doi:10.1128/AEM.02985-06
103. Lasken RS (2012) Genomic sequencing of uncultured microorganisms from single cells. Nat Rev Microbiol 10:631–640. doi:10.1038/nrmicro2857
104. Martin-Cuadrado A-B, Garcia-Heredia I, Moltó AG, López-Úbeda R, Kimes N, López-García P, Moreira D, Rodriguez-Valera F (2014) A new class of marine Euryarchaeota group II from the mediterranean deep chlorophyll maximum. ISME J. doi:10.1038/ismej.2014.249
105. Meyer F, Paarmann D, Souza MD', Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server: a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinform 9:386. doi:10.1186/1471-2105-9-386
106. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res 42:D560–D567. doi:10.1093/nar/gkt963
107. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto J-M, dos Santos Quintanilha MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Léonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sørensen S, Tap J, Tims S, Ussery DW, Yamada T, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, Ehrlich SD, MetaHIT Consortium (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32:822–828. doi:10.1038/nbt.2939
108. Aziz RK, Dwivedi B, Akhter S, Breitbart M, Edwards RA (2015) Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. Virology 6:381. doi:10.3389/fmicb.2015.00381
109. Silva GGZ, Cuevas DA, Dutilh BE, Edwards RA (2014) FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ 2:e425. doi:10.7717/peerj.425
110. Rolfsson Ó, Palsson BO (2015) Decoding the jargon of bottom-up metabolic systems biology. BioEssays 37:588–591. doi:10.1002/bies.201400187
111. Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M, Johnson J, Montgomery R, Ferriera S, Beeson K, Williamson SJ, Tovchigrechko A, Allen AE, Zeigler LA, Sutton G, Eisenstadt E, Rogers Y-H, Friedman R, Frazier M, Venter JC (2010) Genomic and functional adaptation in surface ocean

planktonic prokaryotes. Nature 468:60–66. doi:10.1038/nature09530
112. Scholz MB, Lo C-C, Chain PS (2012) Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Curr Opin Biotechnol 23:9–15. doi:10.1016/j.copbio.2011.11.013
113. Unterseher M, Jumpponen A, Öpik M, Tedersoo L, Moora M, Dormann CF, Schnittler M (2011) Species abundance distributions and richness estimations in fungal metagenomics: lessons learned from community ecology. Mol Ecol 20:275–285. doi:10.1111/j.1365-294X.2010.04948.x
114. Allen HK, Bunge J, Foster JA, Bayles DO, Stanton TB (2013) Estimation of viral richness from shotgun metagenomes using a frequency count approach. Microbiome 1:5. doi:10.1186/2049-2618-1-5
115. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, Owens S, Gilbert JA, Wall DH, Caporaso JG (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci 109:21390–21395. doi:10.1073/pnas.1215210110
116. Hettich RL, Sharma R, Chourey K, Giannone RJ (2012) Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. Curr Opin Microbiol 15:373–380. doi:10.1016/j.mib.2012.04.008
117. Hoffmann KH, Rodriguez-Brito B, Breitbart M, Bangor D, Angly F, Felts B, Nulton J, Rohwer F, Salamon P (2007) Power law rank–abundance models for marine phage communities. FEMS Microbiol Lett 273:224–228. doi:10.1111/j.1574-6968.2007.00790.x
118. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R, Felts B, Haynes M, Liu H, Lipson D, Mahaffy J, Martin-Cuadrado AB, Mira A, Nulton J, Pašić L, Rayhawk S, Rodriguez-Mueller J, Rodriguez-Valera F, Salamon P, Srinagesh S, Thingstad TF, Tran T, Thurber RV, Willner D, Youle M, Rohwer F (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4:739–751. doi:10.1038/ismej.2010.1
119. Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM (2004) A taxa–area relationship for bacteria. Nature 432:750–753. doi:10.1038/nature03073
120. Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, Luo H, Wright JJ, Landry ZC, Hanson NW, Thompson BP, Poulton NJ, Schwientek P, Acinas SG, Giovannoni SJ, Moran MA, Hallam SJ, Cavicchioli R, Woyke T, Stepanauskas R (2013) Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc Natl Acad Sci 110:11463–11468. doi:10.1073/pnas.1304246110
121. Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, Maritz JM, Reeves D, Gandara J, Chhangawala S, Ahsanuddin S, Simmons A, Nessel T, Sundaresh B, Pereira E, Jorgensen E, Kolokotronis S-O, Kirchberger N, Garcia I, Gandara D, Dhanraj S, Nawrin T, Saletore Y, Alexander N, Vijay P, Hénaff EM, Zumbo P, Walsh M, O'Mullan GD, Tighe S, Dudley JT, Dunaif A, Ennis S, O'Halloran E, Magalhaes TR, Boone B, Jones AL, Muth TR, Paolantonio KS, Alter E, Schadt EE, Garbarino J, Prill RJ, Carlton JM, Levy S, Mason CE (2015) Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Syst. doi:10.1016/j.cels.2015.01.001
122. Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289:1902–1906. doi:10.1126/science.289.5486.1902
123. Fuhrman JA (1992) Novel major archaebacterial group from marine plankton. Nature 356:148–149. doi:10.1038/356148a0
124. DeLong EF (1992) Archaea in coastal marine environments. Proc Natl Acad Sci USA 89:5685–5689
125. Jurgens G, Glöckner F-O, Amann R, Saano A, Montonen L, Likolammi M, Münster U (2000) Identification of novel Archaea in bacterioplankton of a boreal forest lake by phylogenetic analysis and fluorescent in situ hybridization. FEMS Microbiol Ecol 34:45–56. doi:10.1111/j.1574-6941.2000.tb00753.x
126. Schleper C, Holben W, Klenk HP (1997) Recovery of crenarchaeotal ribosomal DNA sequences from freshwater-lake sediments. Appl Environ Microbiol 63:321–323
127. Preston CM, Wu KY, Molinski TF, DeLong EF (1996) A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. Proc Natl Acad Sci USA 93:6241–6246
128. Friedrich MW, Schmitt-Wagner D, Lueders T, Brune A (2001) Axial differences in community structure of Crenarchaeota and Euryarchaeota in the highly compartmentalized gut of the soil-feeding termite *Cubitermes orthognathus*. Appl Environ Microbiol 67:4880–4890
129. Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437:543–546. doi:10.1038/nature03911
130. Nicol GW, Schleper C (2006) Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle? Trends Microbiol 14:207–212. doi:10.1016/j.tim.2006.03.004
131. Wu Y, Conrad R (2014) Ammonia oxidation-dependent growth of group I.1b Thaumarchaeota in acidic red soil microcosms. FEMS Microbiol Ecol 89:127–134. doi:10.1111/1574-6941.12340
132. Kowalchuk GA, Stephen JR (2001) Ammonia-oxidizing bacteria: a model for molecular microbial ecology. Annu Rev Microbiol 55:485–529. doi:10.1146/annurev.micro.55.1.485
133. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol 6:245–252. doi:10.1038/nrmicro1852
134. Cavender-Bares J, Kozak KH, Fine PVA, Kembel SW (2009) The merging of community ecology and phylogenetic biology. Ecol Lett 12:693–715. doi:10.1111/j.1461-0248.2009.01314.x
135. Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2009) Relative roles of niche and neutral processes in structuring a soil microbial community. ISME J 4:337–345. doi:10.1038/ismej.2009.122
136. McGill BJ, Enquist BJ, Weiher E, Westoby M (2006) Rebuilding community ecology from functional traits. Trends Ecol Evol 21:178–185. doi:10.1016/j.tree.2006.02.002
137. Rosindell J, Hubbell SP, Etienne RS (2011) The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol Evol 26:340–348. doi:10.1016/j.tree.2011.03.024
138. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, Dornelas M, Enquist BJ, Green JL, He F, Hurlbert AH, Magurran AE, Marquet PA, Maurer BA, Ostling A, Soykan CU, Ugland KI, White EP (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 10:995–1015. doi:10.1111/j.1461-0248.2007.01094.x
139. Venkataraman A, Bassis CM, Beck JM, Young VB, Curtis JL, Huffnagle GB, Schmidt TM (2015) Application of a neutral community model to assess structuring of the human lung microbiome. mBio 6:e02284-14. doi:10.1128/mBio.02284-14
140. Mendes LW, Kuramae EE, Navarrete AA, van Veen JA, Tsai SM (2014) Taxonomical and functional microbial community

selection in soybean rhizosphere. ISME J 8:1577–1587. doi:10.1038/ismej.2014.17
141. Caruso T, Chan Y, Lacap DC, Lau MCY, McKay CP, Pointing SB (2011) Stochastic and deterministic processes interact in the assembly of desert microbial communities on a global scale. ISME J 5:1406–1413. doi:10.1038/ismej.2011.21
142. Ferrenberg S, O'Neill SP, Knelman JE, Todd B, Duggan S, Bradley D, Robinson T, Schmidt SK, Townsend AR, Williams MW, Cleveland CC, Melbourne BA, Jiang L, Nemergut DR (2013) Changes in assembly processes in soil bacterial communities following a wildfire disturbance. ISME J 7:1102–1111. doi:10.1038/ismej.2013.11
143. Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T (2011) Bacterial community assembly based on functional genes rather than species. Proc Natl Acad Sci 108:14288–14293. doi:10.1073/pnas.1101591108
144. Levin SA, Carpenter SR, Godfray HCJ, Kinzig AP, Loreau M, Losos JB, Walker B, Wilcove DS (2009) The Princeton guide to ecology. Princeton University Press, Princeton
145. McCann KS (2000) The diversity–stability debate. Nature 405:228–233. doi:10.1038/35012234
146. Naeem S, Li S (1997) Biodiversity enhances ecosystem reliability. Nature 390:507–509. doi:10.1038/37348
147. Yachi S, Loreau M (1999) Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. Proc Natl Acad Sci U S A 96:1463–1468
148. Yasuda K, Oh K, Ren B, Tickle TL, Franzosa EA, Wachtman LM, Miller AD, Westmoreland SV, Mansfield KG, Vallender EJ, Miller GM, Rowlett JK, Gevers D, Huttenhower C, Morgan XC (2015) Biogeography of the intestinal mucosal and lumenal microbiome in the rhesus macaque. Cell Host Microbe 17:385–391. doi:10.1016/j.chom.2015.01.015
149. Cadotte M, Albert CH, Walker SC (2013) The ecology of differences: assessing community assembly with trait and evolutionary distances. Ecol Lett 16:1234–1244. doi:10.1111/ele.12161
150. Lynch RC, Darcy JL, Kane NC, Nemergut DR, Schmidt SK (2014) Metagenomic evidence for metabolism of trace atmospheric gases by high-elevation desert Actinobacteria. Front Microbiol. doi:10.3389/fmicb.2014.00698
151. Grimm V, Wissel C (1997) Babel, or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion. Oecologia 109:323–334. doi:10.1007/s004420050090
152. Wittebolle L, Marzorati M, Clement L, Balloi A, Daffonchio D, Heylen K, De Vos P, Verstraete W, Boon N (2009) Initial community evenness favours functionality under selective stress. Nature 458:623–626. doi:10.1038/nature07840
153. Consortium THMP (2012) Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. doi:10.1038/nature11234
154. Walter J, Ley R (2011) The human gut microbiome: ecology and recent evolutionary changes. Annu Rev Microbiol 65:411–429. doi:10.1146/annurev-micro-090110-102830
155. Greenblum S, Carr R, Borenstein E (2015) Extensive strain-level copy-number variation across human gut microbiome species. Cell 160:583–594. doi:10.1016/j.cell.2014.12.038
156. Faust K, Lahti L, Gonze D, de Vos WM, Raes J (2015) Metagenomics meets time series analysis: unraveling microbial community dynamics. Curr Opin Microbiol 25:56–66. doi:10.1016/j.mib.2015.04.004
157. Erkus O, de Jager VC, Spus M, van Alen-Boerrigter IJ, van Rijswijck IM, Hazelwood L, Janssen PW, van Hijum SA, Kleerebezem M, Smid EJ (2013) Multifactorial diversity sustains microbial community stability. ISME J 7:2126–2136. doi:10.1038/ismej.2013.108
158. Hurwitz BL, Westveld AH, Brum JR, Sullivan MB (2014) Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. Proc Natl Acad Sci USA 111:10714–10719. doi:10.1073/pnas.1319778111
159. Yang Y, Gao Y, Wang S, Xu D, Yu H, Wu L, Lin Q, Hu Y, Li X, He Z, Deng Y, Zhou J (2014) The microbial gene diversity along an elevation gradient of the Tibetan grassland. ISME J 8:430–440. doi:10.1038/ismej.2013.146
160. Xia LC, Steele JA, Cram JA, Cardon ZG, Simmons SL, Vallino JJ, Fuhrman JA, Sun F (2011) Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. BMC Syst Biol 5(Suppl 2):S15. doi:10.1186/1752-0509-5-S2-S15
161. Castro A, Silva M, Quirino B, Kruger R (2013) Combining "Omics" strategies to analyze the biotechnological potential of complex microbial environments. Curr Protein Pept Sci 14:447–458. doi:10.2174/13892037113149990062
162. Larsen PE, Field D, Gilbert JA (2012) Predicting bacterial community assemblages using an artificial neural network approach. Nat Methods 9:621–625. doi:10.1038/nmeth.1975
163. Bucci V, Xavier JB (2014) Towards predictive models of the human gut microbiome. J Mol Biol. doi:10.1016/j.jmb.2014.03.017
164. Manor O, Levy R (2014) Borenstein E mapping the inner workings of the microbiome: genomic- and metagenomic-based study of metabolism and metabolic interactions in the human microbiome. Cell Metab. doi:10.1016/j.cmet.2014.07.021
165. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW (2012) A whole-cell computational model predicts phenotype from genotype. Cell 150:389–401. doi:10.1016/j.cell.2012.05.044
166. Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR (2015) Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc Natl Acad Sci U S A 112:6449–6454. doi:10.1073/pnas.1421834112
167. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28:977–982. doi:10.1038/nbt.1672
168. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10:291–305. doi:10.1038/nrmicro2737
169. Dutilh BE, Backus L, Edwards RA, Wels M, Bayjanov JR, van Hijum SAFT (2013) Explaining microbial phenotypes on a genomic scale: GWAS for microbes. Brief Funct Genomics 12:366–380. doi:10.1093/bfgp/elt008
170. Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BØ (2013) Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. Proc Natl Acad Sci 110:20338–20343. doi:10.1073/pnas.1307797110
171. Bizukojc M, Dietz D, Sun J, Zeng A-P (2010) Metabolic modelling of syntrophic-like growth of a 1,3-propanediol producer, *Clostridium butyricum*, and a methanogenic archeon, *Methanosarcina mazei*, under anaerobic conditions. Bioprocess Biosyst Eng 33:507–523. doi:10.1007/s00449-009-0359-0
172. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA (2007) Metabolic modeling of a mutualistic microbial community. Mol Syst Biol. doi:10.1038/msb4100131
173. Klitgord N, Segrè D (2010) Environments that induce synthetic microbial ecosystems. PLoS Comput Biol 6:e1001002. doi:10.1371/journal.pcbi.1001002
174. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121. doi:10.1038/nprot.2009.203

175. Fleming RMT, Thiele I, Provan G, Nasheuer HP (2010) Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. J Theor Biol 264:683–692. doi:10.1016/j.jtbi.2010.02.044
176. Zomorrodi AR, Maranas CD (2012) OptCom: a multi-level OPTIMIZATION framework for the metabolic modeling and analysis of microbial communities. PLoS Comput Biol 8:e1002363. doi:10.1371/journal.pcbi.1002363
177. Khandelwal RA, Olivier BG, Röling WFM, Teusink B, Bruggeman FJ (2013) Community flux balance analysis for microbial consortia at balanced growth. PLoS ONE. doi:10.1371/journal.pone.0064567
178. Chiu H-C, Levy R, Borenstein E (2014) Emergent biosynthetic capacity in simple microbial communities. PLoS Comput Biol 10:e1003695. doi:10.1371/journal.pcbi.1003695
179. Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T (2010) Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics 26:i255–i260. doi:10.1093/bioinformatics/btq183
180. Kim MK, Lun DS (2014) Methods for integration of transcriptomic data in genome-scale metabolic models. Comput Struct Biotechnol J 11:59–65. doi:10.1016/j.csbj.2014.08.009
181. Kronfeld M, Dräger A, Aschoff M, Zell A (2009) On the benefits of multimodal optimization for metabolic network modeling. CiteSeerX. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.149.5657. Accessed 10 June 2015
182. Chen J, Zheng H, Liu H, Niu J, Liu J, Shen T, Rui B, Shi Y (2007) Improving metabolic flux estimation via evolutionary optimization for convex solution space. Bioinforma Oxf Engl 23:1115–1123. doi:10.1093/bioinformatics/btm050
183. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD (2012) Mathematical optimization applications in metabolic networks. Metab Eng 14:672–686. doi:10.1016/j.ymben.2012.09.005
184. Zakrzewski P, Medema MH, Gevorgyan A, Kierzek AM, Breitling R, Takano E (2012) MultiMetEval: comparative and multi-objective analysis of genome-scale metabolic models. PLoS ONE 7:e51511. doi:10.1371/journal.pone.0051511
185. Shoaie S, Karlsson F, Mardinoglu A, Nookaew I, Bordel S, Nielsen J (2013) Understanding the interactions between bacteria in the human gut through metabolic modeling. Sci Rep. doi:10.1038/srep02532
186. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, Lang AH, Bonilla G, Kar A, Leiby N, Mehta P, Marx CJ, Segrè D (2014) Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. Cell Rep 7:1104–1115. doi:10.1016/j.celrep.2014.03.070
187. Stern A, Mick E, Tirosh I, Sagy O, Sorek R (2012) CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. Genome Res 22:1985–1994. doi:10.1101/gr.138297.112
188. Modi SR, Lee HH, Spina CS, Collins JJ (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature 499:219–222. doi:10.1038/nature12212
189. Sanguino L, Franqueville L, Vogel TM, Larose C (2015) Linking environmental prokaryotic viruses and their host through CRISPRs. FEMS Microbiol Ecol. doi:10.1093/femsec/fiv046
190. Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074
191. Pride DT, Sun CL, Salzman J, Rao N, Loomer P, Armitage GC, Banfield JF, Relman DA (2011) Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. Genome Res 21:126–136. doi:10.1101/gr.111732.110
192. Weitz JS, Poisot T, Meyer JR, Flores CO, Valverde S, Sullivan MB, Hochberg ME (2013) Phage-bacteria infection networks. Trends Microbiol 21:82–91. doi:10.1016/j.tim.2012.11.003
193. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS (2011) Statistical structure of host-phage interactions. Proc Natl Acad Sci U S A 108:E288–E297. doi:10.1073/pnas.1101595108
194. Flores CO, Valverde S, Weitz JS (2013) Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. ISME J 7:520–532. doi:10.1038/ismej.2012.135
195. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, Mira A (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7:828–836. doi:10.1038/nrmicro2235
196. Suttle CA (2007) Marine viruses–major players in the global ecosystem. Nat Rev Microbiol 5:801–812. doi:10.1038/nrmicro1750
197. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB, Tara Oceans Coordinators (2015) Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science 348:126. doi:10.1126/science.1261498