University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Faculty Publications in Food Science and Technology

Food Science and Technology Department

12-4-2014

The unseen world: environmental microbial sequencing and identification methods for ecologists

Naupaka Zimmerman

Jacques Izard

Christian Klatt

Jizhong Zhou

Emma Aronson

The unseen world: environmental microbial sequencing and identification methods for ecologists

Naupaka Zimmerman[1,2], Jacques Izard[3,4], Christian Klatt[5], Jizhong Zhou[6,7,8], and Emma Aronson[9*]

Microorganisms inhabit almost every environment, comprise the majority of diversity on Earth, are important in biogeochemical cycling, and may be vital to ecosystem responses to large-scale climatic change. In recent years, ecologists have begun to use rapidly advancing molecular techniques to address questions about microbial diversity, biogeography, and responses to environmental change. Studies of microbes in the environment generally focus on three broad objectives: determining which organisms are present, what their functional capabilities are, and which are active at any given time. However, comprehending the range of methodologies currently in use can be daunting. To provide an overview of environmental microbial sequence data collection and analysis approaches, we include case studies of microbiomes ranging from the human mouth to geothermal springs. We also suggest contexts in which each technique can be applied and highlight insights that result from their use.

Front Ecol Environ 2014; 12(4): 224–231, doi:10.1890/130055

Archaea, bacteria, microeukaryotes, and the viruses that infect them (collectively "microorganisms") are foundational components of all ecosystems, inhabiting almost every imaginable environment and comprising the majority of the planet's organismal and evolutionary diversity. Microorganisms play integral roles in ecosystem functioning; are important in the biogeochemical cycling of carbon (C), nitrogen (N), sulfur (S), phosphorus (P), and various metals (eg Barnard *et al.* 2005); and may be vital to ecosystem responses to large-scale climatic change (Mackelprang *et al.* 2011). Rarely found alone, microorganisms often form complex communities that are dynamic in space and time (Martiny *et al.* 2006). For these and other reasons, ecologists and environmental scientists have become increasingly interested in understanding microbial dynamics in ecosystems. Ecological studies of microbes in the environment generally focus on determining which organisms are present and what functional roles they are playing or could play. Rapid advances in molecular and bioinformatic approaches over the past decade have dramatically reduced the difficulty and cost of addressing such questions (Figure 1; WebTable 1). Yet the range of methodologies currently in use and the rapid pace of their ongoing development can be daunting for researchers unaccustomed to these technologies.

The goals of this article, which originated in an organized session at an Ecological Society of America annual meeting, are (1) to introduce non-specialists to a selection of the approaches currently used to study microbial communities in the environment and (2) to provide examples of their application through a series of case studies. We include examples from diverse microbial habitats – from the human mouth to geothermal springs to soil. The scale of observation in these systems ranges from millimeters to thousands of kilometers. We also suggest possible contexts in which each technique can be used and highlight a number of insights and potential applications.

In a nutshell:

- Recent progress in sequencing technologies coupled with increasing affordability allow for rapid characterization of microbial communities in environmental samples
- We describe several applications of genetic sequencing techniques that can be used to explore ecological questions
- Shotgun sequencing has the potential to reveal novel biodiversity, whereas amplicon sequencing and DNA microarrays of functional and taxonomic genes provide a more targeted approach for understanding community structure and/or function
- Methods quantifying the presence and expression of functional genes facilitate the study of microbial communities' functional capacities

[1]*Department of Biology, Stanford University, Stanford, CA;* [2]*School of Plant Sciences, University of Arizona, Tucson, AZ;* [3]*Department of Microbiology, The Forsyth Institute, Cambridge, MA;* [4]*Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston, MA* *(continued on p231)*

Culture-independent assessment of microbial diversity

One of the first steps in many ecological research projects is to assess which organisms are present in a given environment and to determine their diversity and distribution patterns. Biogeography – the study of the distribution of organisms over space and time – is often a prerequisite to documenting life histories, investigating the impact of imposed treatments, or seeking a mechanistic understanding of microbial roles in the environment (Martiny *et al.* 2006). Traditional methods of studying microorganisms generally relied upon removing a sample from the environment and then cultivating individual strains in the laboratory. This dependence on cultivation made it hard to assess highly diverse microbial assemblages in situ. Furthermore, many microorganisms are resistant to lab-culturing techniques (see Epstein [2013] for a recent review). Despite this obstacle, modern molecular techniques have advanced the understanding of microbial dynamics in natural habitats by culture-independent characterizations of nucleic acid biomarkers (DNA and RNA) instead of direct assays of living organisms.

In some cases, particularly in community diversity studies, it is necessary to assay only particular genomic regions (loci) of interest instead of all the genetic material in a given sample. All such studies require the use of a genetic locus with a nucleotide sequence that does not vary too much (ie is conserved) across diverse taxa but varies enough to allow differentiation among taxa. Conservation is important because it enables the use of a single set of primers – short DNA sequences used to target specific genetic regions – for in vitro replication (amplification) of this locus across broad taxonomic groups during polymerase chain reaction (PCR). This approach is known as "amplicon sequencing" because the target is an amplified fragment. The locus commonly used for studies of bacteria and archaea (among others) is the gene encoding a subunit of ribosomal RNA (16S rRNA; Tringe and Hugenholtz 2008). While studies of eukaryotes typically also rely on sequencing rRNA genes, they often use slightly different loci. Many studies will use either the rRNA gene encoding for the small subunit, sometimes labeled 18S or SSU, or the gene encoding the large subunit (LSU). Still other studies, and particularly those targeting fungi, will use the non-coding Internal Transcribed Spacer (ITS) regions between rRNA genes (Nilsson *et al.* 2009). The amplicon sequencing approach can also be used to target specific functional genes (eg *amoA*, a locus encoding ammonia monooxygenase, an enzyme that catalyzes a necessary step in the process of nitrification) to determine the diversity of organisms capable of performing that function (Smith *et al.* 2009). As finer-level taxonomic boundaries between microorganisms are often unclear, studies of this type commonly use the term "Operational Taxonomic Units" (OTUs) as the unit of interest instead of "species" derived from known cultivable organisms. OTUs are groups of sequences that are classified together based on their sequence similarity and can be attributed

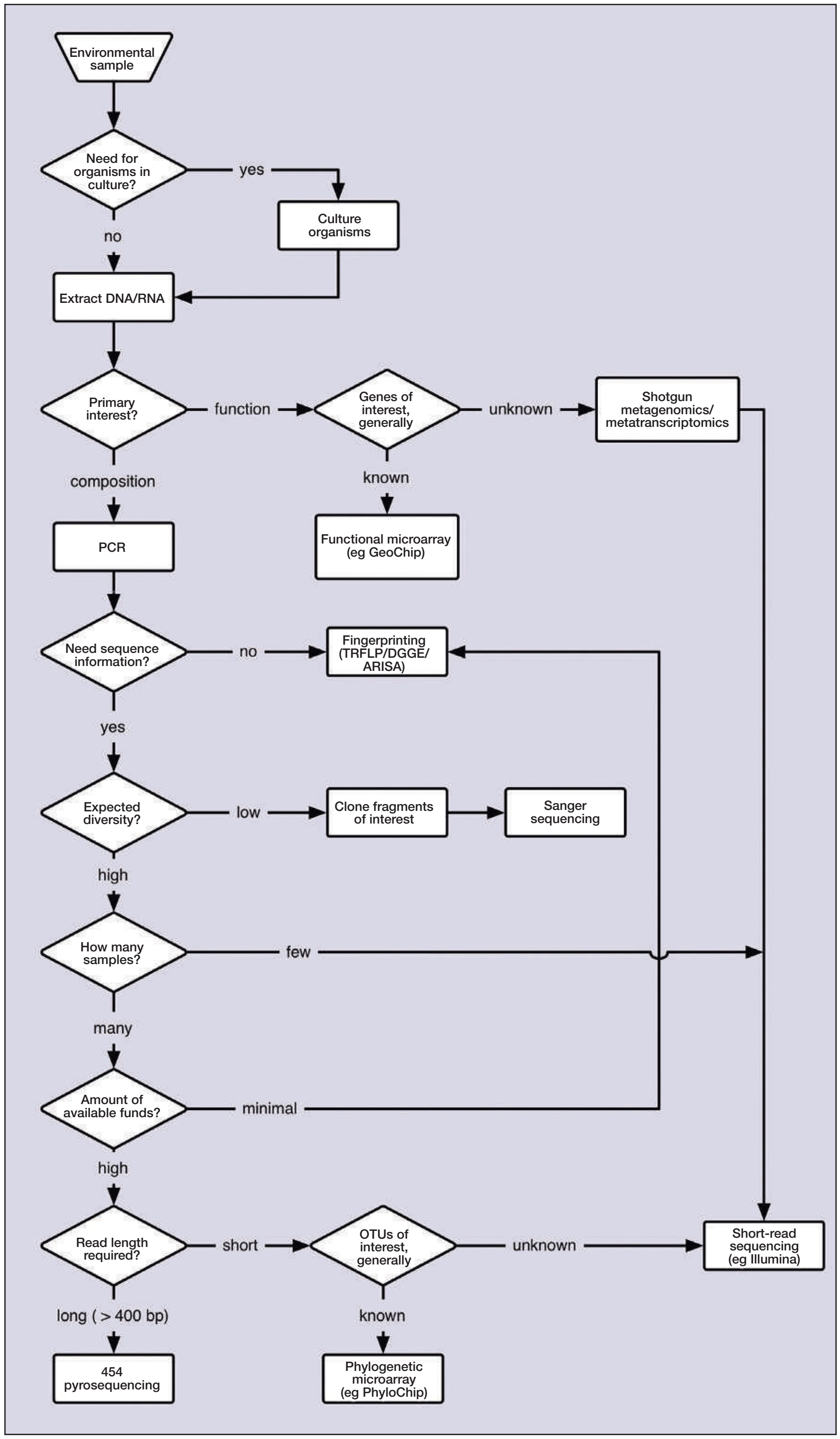

Figure 1. *Decision diagram for choosing a molecular approach for use in microbial ecology studies.*

to species when reference strains are available. Many studies of bacteria, archaea, and microeukaryotes group sequences together as an OTU when they are 97% similar to each other at the 16S rRNA locus; however, it is known that OTUs grouped together based on this level of sequence similarity can contain multiple, ecologically differentiable microorganisms (Koeppel *et al.* 2008).

Sequencing through the Sanger method has been used to identify nucleotide sequences for several decades (Sanger *et al.* 1977) and can generate high-quality, long (750–1000 base pair [bp]) sequences. However, Sanger sequencing cannot be performed on mixed assemblages of DNA without first separating each unique DNA fragment. Therefore, to use this method with amplicons from an environmental sample, researchers must create a clone library. This is accomplished by transforming individual host cells (generally *Escherichia coli*) with a single sequence variant and then growing these cells in separate colonies, thereby isolating and replicating each fragment of interest many times. Although some of the steps in this procedure can be automated, creating and sequencing clone libraries is time-consuming and costly, particularly for diverse microbial communities that may harbor millions of microorganisms.

Advances in sequencing technology over the past decade have largely supplanted the methods described above for many applications. While there are an increasing number of different technologies used for next-generation sequencing (NGS), the two most common approaches in contemporary environmental microbiology studies are 454 pyrosequencing by Roche Inc (Taberlet *et al.* 2012) and the sequencing by synthesis approach by Illumina Inc (Caporaso *et al.* 2012). These methods allow researchers to perform extremely high-throughput amplicon sequencing, resulting in large numbers of sequencing reads for many samples simultaneously. It should be noted, however, that usage of particular sequencing technologies may shift rapidly. As of 2014, researchers still regularly use 454 sequencers. However, Roche announced in October 2013 that it would be terminating its 454 sequencing operations by the end of 2015. Existing 454 sequencers will continue to be supported until then, but approaches such as sequencing by synthesis, used by the various Illumina platforms, and single-molecule sequencing, as used in the Pacific Biosciences or Oxford Nanopore systems, will likely be increasingly relied upon by microbial ecologists in the near future.

In the context of NGS, amplicon sequencing is also sometimes referred to as "barcode sequencing", "barcoding", or "meta-barcoding". The term "barcode" has been used to describe two separate features of the amplicon sequencing process: (1) the sample DNA locus targeted for sequencing, due to its use in taxonomic classification (eg Taberlet *et al.* 2012); and (2) the unique sequence of nucleotides (also called a "tag") incorporated into each amplified DNA fragment in a sample via PCR, and used to identify individual samples within a single "multiplexed" (multiple sample) run (eg Mackelprang *et al.* 2011). In this review, the terms "amplicon" and "tag" are used for the former and latter examples, respectively. The latter tags allow multiple samples, each containing many

Panel 1. Landscape distribution of fungal endophytes revealed through pyrosequencing

Foliar fungal endophytes – microfungi that inhabit the asymptomatic leaf tissue of plants – are ubiquitous in all plant species yet surveyed and have been shown to exhibit many diverse ecological roles, including latent saprotrophy (Voriskova and Baldrian 2013), latent pathogenicity (Rodriguez *et al.* 2009), and mutualism (Arnold *et al.* 2003). However, relatively little is known about the way that abiotic factors influence the makeup of these communities across a landscape. One reason for this is that these communities can be hyperdiverse, particularly in the tropics (Arnold *et al.* 2000), and fully sampling them is expensive and time-consuming. Tagged amplicon pyrosequencing alleviates both of these difficulties. Zimmerman and Vitousek (2012) used this approach to sequence the endophyte communities of 130 trees across an environmental matrix in Hawaii (Figure 2); this study design allowed them to decouple the effects of elevation, rainfall, and substrate age on these communities. They found high diversity at the landscape scale (> 4000 fungal OTUs), primarily driven by between-site differences, which were strongly correlated to both elevation and rainfall.

Figure 2. *'Ōhi'a lehua (*Metrosideros polymorpha*) on Mauna Loa volcano, Hawaii.*

different DNA fragments, to be sequenced together and later separated via computer programs (ie through bioinformatics).

Relied upon for nearly two decades, genetic microarrays are another technique used to assess the diversity of genetic material in a given sample. Microarrays (often given proprietary names such as GeoChips or PhyloChips) are prefabricated microchips lined with short, known sequences ("probes") that provide information on the phylogenetic relationships between organisms, the functional capacity of those organisms (in the case of sequences encoding for particular genes of interest), or both. When a sample is applied, sequences in the sample that match the probes hybridize with them (based on sequence complementarity) and fluoresce in proportion to the number of matching sequence fragments present. These different levels of fluorescence can then be quantified, providing information on gene diversity and abundance (Zhou *et al.* 2011b). One of the more commonly used functional microarrays is the GeoChip, which is so named because it primarily targets genes with biogeochemical functions.

Examples of NGS used to characterize the makeup of microbial communities in the environment include studies describing the landscape distribution of fungal endophytes in Hawaii (Panel 1; Zimmerman and Vitousek 2012) and the microbiota residing on the human tooth (Panel 2; Segata *et al.* 2012). An example of using microarrays for assaying community composition comes from research focusing on microbes involved in methane cycling in a pine forest soil (Panel 3; Aronson *et al.* 2013).

Identifying functional traits and assaying gene expression

Upon gaining some understanding of the composition of a microbial community, researchers often seek to assay the possible ecological functions of its members. For example, what is the connection between the diversity observed in a microbial community and its functional capability, and how do these microorganisms interact with the flux of energy and nutrients in their ecosystems? The techniques commonly used in the investigation of microbial community functions include functional gene microarrays and what has become known as "shotgun sequencing", so-called because the extracted nucleic acids are broken up randomly into small fragments of up to several hundred base pairs and subsequently sequenced. Here, as in a shotgun blast, targeting and precision are traded for breadth. In comparison to the more targeted amplicon sequencing approach, shotgun sequencing is used when researchers are interested in a comprehensive, non-targeted sampling of the DNA or RNA (genomics or transcriptomics, respectively) from a given organism. This type of approach can reveal the functional potential of microorganisms and, since it is non-targeted, can be used to identify novel genes or pathways that encode these functions.

Panel 2. Moving toward a greater scrutiny of the human microbiota through pyrosequencing

With an estimated 300 million bacterial cells on a single tooth surface (Haffajee 2009), the greater number of sequences provided by tagged pyrosequencing (relative to clone libraries and the Sanger approach) enhances our ability to investigate population structure, shifts over time, and multiple habitats per subject. Observing diversity through ecological indices is the first step, but the challenge lies in explaining observed shifts in both abundant and rare members of the microbiota. In other words, what is the importance of specific OTUs with high variability per body habitat among a population? Are any of them useful biomarkers? Many approaches are influenced by the abundance of highly prevalent organisms, while a combination of statistical tests allow for the detection of both low and high abundance markers (Segata *et al.* 2011). The latest approach, using linear discriminant analysis effect size (LEfSe) software, enables quantification of the impact of different microbial markers along the human tooth in areas affected by differing levels of oxygen exposure, resulting in proportional differences of aerotolerants and anaerobes (Segata *et al.* 2012). The circular cladogram (Figure 3, reprinted from Segata *et al.* 2012) is based on the Ribosomal Database Project (RDP) Taxonomy (Cole *et al.* 2009) and shows taxa differentially represented between the supragingival (above the gum line; in red [SupP]) and the subgingival (below the gum line; in green [SubP]) plaque. This demonstrates the extensive preferential organization, even at these highly related sites. At the class level, Actinobacteria, Bacilli, Gamma-proteobacteria, Beta-proteobacteria, and Flavobacteria are characteristic of the supragingival plaque, whereas Fusobacteria, Clostridia, Epsilon-proteobacteria, Spirochaetes, Bacteroidia, and unclassified Bacteroidetes are biomarkers for the subgingival plaque.

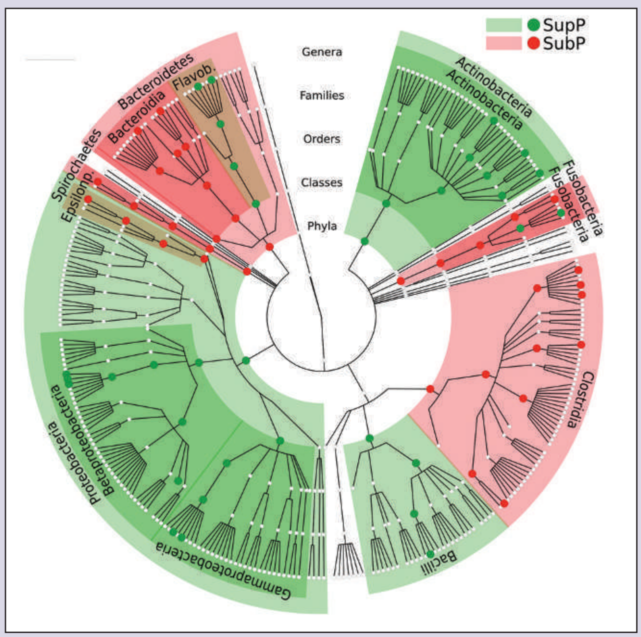

Figure 3. Cladogram showing the different but overlapping microbial communities on the exposed portion of the tooth (SupP) and below the gum line (SubP).

Panel 3. Identifying the primary drivers of greenhouse-gas cycles with PhyloChips

An important trace gas responsible for at least 20% of the current greenhouse effect, methane (CH_4) is produced and consumed by soil microorganisms across the globe, with rates of flux driven by environmental conditions in the soil, measured by means of soil chambers (Figure 4) and gas chromatography. Aronson *et al.* (2013) used a genetic microarray to investigate the target microbes involved in the CH_4 cycle in a pine forest soil under different experimental conditions. The third generation PhyloChip, a 16S rRNA gene microarray designed to provide information on almost 60 000 different OTUs (Hazen *et al.* 2010), was used to quantify the community. Gene arrays are uniquely suited to identify the relative representation of particular microorganisms between locations, time-points, and treatments. This study focused on the diversity of methanotrophic (CH_4 consuming) and methanogenic (CH_4 producing) microorganisms in the soil and their association with variations in CH_4 flux into and out of the soil (Aronson *et al.* 2013). Differences were detected in target communities across sites, time-points, and N treatments, particularly among the less common members of the community. These differences in the least common OTUs (the so-called rare biosphere) might have been obscured by the more common soil microbial phyla in a clone-based screening method and possibly even using high-throughput sequencing without sufficient depth or replication (Zhou *et al.* 2011a).

Figure 4. *Soil collar for in situ methane quantification.*

Meta-omic (eg metagenomic, metatranscriptomic) approaches involve characterization of complex samples representing multiple organisms or entire communities (versus genomics or transcriptomics, which focus on sequences from a single organism). The challenges in metagenomics are twofold: to identify and quantify gene function from short random sequences, and then to link those functional gene fragments to other genes that allow for taxonomic inference. Determination of gene function is often accomplished by matching short sequence fragments to already-published and annotated reference genomes. These annotations, in turn, are generally based on the results of manipulative genetics experiments that have characterized the function of individual genes. The difference between metagenomics (DNA) and metatranscriptomics (RNA) is that the latter includes information about which organisms in the sample are currently active and what they are doing. This inference is based on the observation that the residence time of RNA is much shorter than DNA in both microbial cells and in, for example, the soil matrix; therefore RNA extracted from an environmental sample theoretically represents a snapshot of recent levels of gene expression in the organisms from that sample.

Extremely deep meta-omics studies are being used for understanding dynamics in complex communities, but the computational resources required for such projects (particularly in soil) remain prohibitive (however see Fierer *et al.* 2012). Despite the ongoing limitations imposed by computational hardware, development of new pipelines (sets of software programs or scripts, each performing one step in a multi-step process) and algorithms for data processing has progressed more rapidly. While new pipelines are developed seemingly every month, several – for example, WebCarma (Gerlach *et al.* 2009), QIIME (Caporaso *et al.* 2010), Galaxy (Goecks *et al.* 2010), and MG-RAST (Meyer *et al.* 2008) – have reached wider use.

Here, we highlight two studies that assay the functional potential of microbes: one that uses a GeoChip microarray to examine the effect of soil warming on microbial communities in the US Great Plains (Zhou *et al.* 2011b; Panel 4) and another that uses shotgun metagenomic sequencing to investigate the composition and functional

Panel 4. Investigating ecosystem function with microarrays

Understanding the mechanisms of biospheric feedbacks to climate change is critical to project future climate warming. Although microorganisms catalyze most biosphere processes related to fluxes of greenhouse gases, little is known about the microbial role in regulating future climate change. The GeoChip functional gene microarray can quantify the presence of genes involved in C, N, S, and P cycling, organic contaminant degradation, metal resistance, antibiotic resistance, stress responses, virulence, and bacterial phage-mediated lysis, among others. In Zhou *et al.* (2011b), the GeoChip was used in conjunction with high-throughput sequencing in an analysis of soil from a long-term experimental warming site in the US Great Plains. The goal was to explore the role of microbial mediation in C-cycle feedbacks to climate warming. First, long-term experimental warming induced a decline in temperature sensitivity of heterotrophic soil respiration by 14.5% in comparison to the control, largely attributable to functional adjustments in soil microbial communities. Second, warming significantly stimulated functional genes for labile C decomposition but did not affect genes for recalcitrant C decomposition, although both labile and recalcitrant C input to soil increased under warming. Such differential impacts on microbial functional groups may promote long-term stability of ecosystem C. Third, warming stimulated functional genes for nutrient cycling, possibly favoring plant growth and vegetation C uptake. These results indicate that microorganisms critically regulated the ecosystem C-cycle feedback to climate warming, with important implications for C–climate modeling.

potential of microbial communities in geothermal springs (Klatt *et al*. 2011, 2013; Panel 5).

■ Caveats and technical biases

New molecular techniques allow for comprehensive assessment of the structure and function of microbial communities; however, there are several caveats associated with their use. Many of these, including DNA/RNA extraction biases and PCR biases, are not specific to NGS techniques. Nucleic acid extraction includes a step intended to disrupt the cell membranes and release nucleic acids, but this disruption can be incomplete, causing researchers to inadvertently exclude cells that are more resistant to being lysed (Kim and Bae 2011). PCR bias can occur whenever a particular amplified locus (such as the gene encoding the 16S rRNA subunit) is used as the basis of comparisons; the specific choice of primers can intentionally or unintentionally limit the gene variants amplified. Furthermore, sequence artifacts known as chimeras can be introduced during PCR. Chimeras represent fusions of multiple disparate templates from the original DNA pool and can sometimes account for as much as 45% of the dataset (Ashelford *et al*. 2006). Many, but not all, of these chimeric sequences can be removed bioinformatically (ie with programs such as ChimeraSlayer; Haas *et al*. 2011).

Beyond the vagaries of extraction and PCR, each of the different microarray or sequencing technologies has its own particular biases. While microarrays make it possible to assay thousands or millions of fragments simultaneously, inaccuracies can be introduced as a result of biases in how fragments anneal to the array (Gentry *et al*. 2006). Particular sequencing technologies have characteristic biases as well: 454 sequencers (and those based on certain other technologies, such as the Ion Torrent sequencers produced by Life Technologies) can have difficulty quantifying long runs of homopolymers (repeats of the same base in a given sequence; Zhou *et al*. 2011a; Loman *et al*. 2012), whereas Illumina sequencers can exhibit a bias when sequencing guanine–cytosine (GC) rich regions (Minoche *et al*. 2011).

Even after the sequence data are retrieved from the sequencer, bioinformatic challenges remain. In shotgun sequencing, because the sequence fragments are often several hundred base pairs or less, they can be difficult to assemble into complete genes. The function of many genes is still unknown, despite having their full sequence; comparing sequence similarity with that of known genes or motifs (sequences that encode a putatively functional domain) is often the only approach available. Unfortunately, this approach does not always provide definitive mapping from sequence to function. Additional studies of individual pathways will be required, to fill in the gaps in our understanding of microbial biochemistry and to better understand dynamic microbial communities.

■ Rapid technological change and data management

Most research published thus far on microbial community ecology using NGS has been based on sequences from Roche's 454 or one of the several Illumina platforms (eg GAIIx, HiSeq, MiSeq). However, less commonly used

Panel 5. Identifying functional traits of microbial assemblages with shotgun sequencing

Microbial communities from extreme environments (such as those inhabiting hypersaline environments, acid mine drainage, or geothermal springs) may exhibit lower species diversity and complexity as compared with communities inhabiting mesophilic habitats, including many soils or marine environments. With less diverse communities, relatively modest levels of shotgun metagenomic sequencing have successfully revealed functional genes that can readily be assigned to dominant members of these communities (Kunin *et al*. 2008; Inskeep *et al*. 2010). The research described by Klatt *et al*. (2011, 2013) utilized metagenomics to discern the functional attributes of dominant community members in phototrophic microbial mat communities inhabiting geothermal springs in Yellowstone National Park (Figure 5) and partitioned the members of these communities into interacting functional guilds. Two of the dominant members in these communities were only distantly related to any cultured organisms, so that it was not possible to discern which taxonomic group they belonged to. Instead, their genomic signatures, characterized by oligonucleotide frequency patterns, were used to cluster sequence data together with others that originated from the same taxonomic group. This technique provided the means to link 16S rRNA genes with functional genes involved in phototrophy, which established the presence of two previously uncharacterized phototrophic bacteria.

D Ward

Figure 5. *A phototrophic microbial mat community – inhabiting a geothermal spring located in the Mammoth Hot Springs area of Wyoming's Yellowstone National Park – exhibits a conspicuous and colorful representation of the microbial diversity observed across geochemical gradients at millimeter scales.*

techniques may soon become more widespread. Emerging technologies include the longer reads of the bench-top Illumina MiSeq (up to 300 bp per read); single molecule sequencers made by PacBio, which can produce reads of up to 10 kilobase pairs (kbp) but have higher per-base error rates (Mosher *et al.* 2013); and sequencers that rely on non-optical sensing, such as those manufactured by Ion Torrent (Rothberg *et al.* 2011) and Oxford Nanopore (Clarke *et al.* 2009).

To illustrate how rapidly methods are evolving, consider the changes that have occurred in one of the global centers of genome-level sequencing, the US Department of Energy's (DOE's) Joint Genome Institute (JGI) in Walnut Creek, California. Within just the past decade, JGI has transitioned from using only Sanger sequencers, to Sanger in conjunction with 454, to Illumina in conjunction with 454, to an Illumina/PacBio pipeline for their genome sequencing projects. More detailed discussions of these technologies and their applications are available (Glenn 2011; Loman *et al.* 2012; Shendure and Aiden 2012; Thomas *et al.* 2012; Segata *et al.* 2013).

The methods described here produce datasets that are often gigabytes to terabytes in size; thus, archival storage has become increasingly resource-intensive. The canonical repository for genetic sequences in the US is the National Center for Biotechnology Information (NCBI). NCBI now hosts several databases that serve as repositories for sequence data: two examples are the Sequence Read Archive (SRA; Leinonen *et al.* 2011) for data from high-throughput sequencing projects and the Gene Expression Omnibus (GEO; Barrett *et al.* 2013) for gene expression studies. However, these databases are not curated, and the quality of the metadata can vary markedly between submitted projects. Other locations where NGS data can be deposited and made publicly available include MG-RAST (Meyer *et al.* 2008), a server-side metagenomics processing resource run by the US DOE's Argonne National Laboratory, and a repository associated with the Quantitative Insights Into Microbial Ecology (QIIME) pipeline (Caporaso *et al.* 2010). There are also efforts underway to streamline the management of, and access to, publicly funded data. Ongoing initiatives include JGI's Integrated Microbial Genomes and Metagenomes (IMG/M; Markowitz *et al.* 2011), the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA; Sun *et al.* 2010), and the US DOE's KBase platform (http://kbase.us).

Conclusions

Current methods for elucidating microbial community composition and function are based on reference datasets, which are predominantly created from genes identified through cloning, targeted Sanger sequencing, and/or genome assemblies. Thus, despite the huge potential of new sequencing technologies, culture-based research and archiving of type specimens must continue; without these foundational data, results with the potential to yield important ecological insights will remain undecipherable. The challenges are substantial, yet the field is also seeing rapid improvement in sequencing technologies and the development of more efficient and rigorous data analysis tools, coupled with a growing number of ecologists being trained in microbial methods. Just as a better understanding of the human microbiome has enabled a new set of medical treatments and more accurate diagnoses, our growing understanding of microbial communities in the environment will lead to a better understanding of how ecosystems function and will provide new opportunities to test and formulate ecological theory. The ongoing application of this growing field will provide new approaches for environmental remediation and for the sustainable management of natural and agricultural ecosystems worldwide.

Acknowledgements

This work was funded in part by an NSF Graduate Research Fellowship and a Gordon and Betty Moore Postdoctoral Fellowship from the Life Sciences Research Foundation to NZ; a grant from the National Cancer Institute (CA166150) to JI; an NSF IGERT in Geobiological Systems (DGE 0654336) to CK; a grant from the DOE Office of Biological and Environmental Research (DE-AC02-05CH11231, as part of ENIGMA, a Scientific Focus Area) to JZ; and a NASA Graduate Student Researchers Program Fellowship and a NOAA Climate and Global Change Fellowship to EA. CK would also like to thank C Hendrix and S Guenther (Center for Resources at Yellowstone National Park) for permitting assistance in Yellowstone National Park (Permit No YELL-0129).

References

Arnold AE, Maynard Z, Gilbert GS, *et al.* 2000. Are tropical fungal endophytes hyperdiverse? *Ecol Lett* **3**: 267–74.

Arnold AE, Mejia LC, Kyllo DA, *et al.* 2003. Fungal endophytes limit pathogen damage in a tropical tree. *P Natl Acad Sci USA* **100**: 15649–54.

Aronson EL, Dubinsky EA, and Helliker BR. 2013. Effects of nitrogen addition on soil microbial diversity and methane cycling capacity depend on drainage conditions in a pine forest soil. *Soil Biol Biochem* **62**: 119–28.

Ashelford KE, Chuzhanova NA, Fry JC, *et al.* 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl Environ Microb* **72**: 5734–41.

Barnard R, Leadley PW, and Hungate BA. 2005. Global change, nitrification, and denitrification: a review. *Global Biogeochem Cy* **19**: GB2007.

Barrett T, Wilhite SE, Ledoux P, *et al.* 2013. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res* **41**: D991–95.

Caporaso JG, Kuczynski J, Stombaugh J, *et al.* 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–36.

Caporaso JG, Lauber CL, Walters WA, *et al.* 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**: 1621–24.

Clarke J, Wu HC, Jayasinghe L, *et al.* 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**: 265–70.

Cole JR, Wang Q, Cardenas E, *et al.* 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**: D141–45.

Epstein SS. 2013. The phenomenon of microbial uncultivability. *Curr Opin Microbiol* **16**: 636–42.

Fierer N, Leff JW, Adams BJ, *et al.* 2012. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *P Natl Acad Sci USA* **109**: 21390–95.

Gentry TJ, Wickham GS, Schadt CW, *et al.* 2006. Microarray applications in microbial ecology research. *Microb Ecol* **52**: 159–75.

Gerlach W, Junemann S, Tille F, *et al.* 2009. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* **10**: 430.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**: 759–69.

Goecks J, Nekrutenko A, Taylor J, and Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**: R86.

Haas BJ, Gevers D, Earl AM, *et al.* 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**: 494–504.

Haffajee AD. 2009. Plaque microbiology in (periodontal) health and disease. In: Henderson B, Curtis MA, Seymor RM, and Donos N (Eds). Periodontal medicine and system biology. Chichester, UK: Wiley-Blackwell.

Hazen TC, Dubinsky EA, DeSantis TZ, *et al.* 2010. Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science* **330**: 204–08.

Inskeep WP, Rusch DB, Jay ZJ, *et al.* 2010. Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS ONE* **5**: e9773.

Kim K-H and Bae J-W. 2011. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microb* **77**: 7663–68.

Klatt CG, Inskeep WP, Herrgard MJ, *et al.* 2013. Community structure and function of high-temperature chlorophototrophic microbial mats inhabiting diverse geothermal environments. *Front Microbiol* **4**: 106.

Klatt CG, Wood JM, Rusch DB, *et al.* 2011. Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *ISME J* **5**: 1262–78.

Koeppel A, Perry EB, Sikorski J, *et al.* 2008. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. *P Natl Acad Sci USA* **105**: 2504–09.

Kunin V, Raes J, Harris JK, *et al.* 2008. Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol Syst Biol* **4**: 198.

Leinonen R, Sugawara H, Shumway M, *et al.* 2011. The sequence read archive. *Nucleic Acids Res* **39**: 19–21.

Loman NJ, Misra RV, Dallman TJ, *et al.* 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* **30**: 434–39.

Mackelprang R, Waldrop MP, Deangelis KM, *et al.* 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480**: 368–71.

Markowitz VM, Chen IM, Chu K, *et al.* 2011. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* **40**: D123–29.

Martiny JBH, Bohannan BJM, Brown JH, *et al.* 2006. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102–12.

Meyer F, Paarmann D, D'Souza M, *et al.* 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Minoche A, Dohm J, and Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* **12**: R112.

Mosher JJ, Bernberg EL, Shevchenko O, *et al.* 2013. Efficacy of a 3rd generation high-throughput sequencing platform for analyses of 16S rRNA genes from environmental samples. *J Microbiol Meth* **95**: 175–81.

Nilsson RH, Ryberg M, Abarenkov K, *et al.* 2009. The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiol Lett* **296**: 97–101.

Rodriguez RJ, White Jr JF, Arnold AE, *et al.* 2009. Fungal endophytes: diversity and functional roles. *New Phytol* **182**: 314–30.

Rothberg JM, Hinz W, Rearick TM, *et al.* 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348–52.

Sanger F, Nicklen S, and Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *P Natl Acad Sci USA* **74**: 5463–67.

Segata N, Boernigen D, Tickle TL, *et al.* 2013. Computational meta'omics for microbial community studies. *Mol Syst Biol* **9**: 1–15.

Segata N, Haake SK, Mannon P, *et al.* 2012. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol* **13**: R42.

Segata N, Izard J, Waldron L, *et al.* 2011. Metagenomic biomarker discovery and explanation. *Genome Biol* **12**: R60.

Shendure J and Aiden EL. 2012. The expanding scope of DNA sequencing. *Nat Biotechnol* **30**: 1084–94.

Smith AM, Heisler LE, Mellor J, *et al.* 2009. Quantitative phenotyping via deep barcode sequencing. *Genome Res* **19**: 1836–42.

Sun S, Chen J, Li W, *et al.* 2010. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res* **39**: D546–51.

Taberlet P, Prud'homme SM, Campione E, *et al.* 2012. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Mol Ecol* **21**: 1816–20.

Thomas T, Gilbert JA, and Meyer F. 2012. Metagenomics – a guide from sampling to data analysis. *Microb Inform Exp* **2**: 3.

Tringe SG and Hugenholtz P. 2008. A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol* **11**: 442–46.

Voriskova J and Baldrian P. 2013. Fungal community on decomposing leaf litter undergoes rapid successional changes. *ISME J* **7**: 477–86.

Zhou J-Z, Wu L-Y, Deng Y, *et al.* 2011a. Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* **5**: 1303–13.

Zhou J-Z, Xue K, Xie JP, *et al.* 2011b. Microbial mediation of carbon-cycle feedbacks to climate warming. *Nat Clim Change* **2**: 106–10.

Zimmerman NB and Vitousek PM. 2012. Fungal endophyte communities reflect environmental structuring across a Hawaiian landscape. *P Natl Acad Sci USA* **109**: 13022–27.

[5]*Department of Forest Ecology and Management, Swedish University of Agricultural Sciences, Umeå, Sweden;* [6]*Institute for Environmental Genomics, and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK;* [7]*Earth Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA;* [8]*State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China;* [9]*Department of Plant Pathology and Microbiology, University of California, Riverside, Riverside, CA* [*]*(emma.aronson@ucr.edu)*