Oomic Approaches in Microbial Ecology: Charting the Unknown

Analysis of whole-community sequence data is unveiling the diversity and function of specific microbial groups within uncultured phyla and across entire microbial ecosystems

Brett J. Baker and Gregory J. Dick

“Somewhere, something incredible is waiting to be known.”

—Carl Sagan

Much like astronomy, the microbial sciences are confronted with staggering unknowns and are reliant on tools and technology to probe the frontiers of knowledge. The first realizations of just how extraordinarily diverse and unexplored microorganisms are came after analyzing microbial ribosomal RNA (rRNA) gene sequences directly from environmental samples. Norm Pace of the University of Colorado, Boulder, and his collaborators 20 years ago helped to uncover the array of microorganisms from hot springs and comparable environments. However, the novel microorganisms that they revealed are not restricted to extreme environments. In fact, we need to look no farther than the dirt outside our homes and workplaces, the lakes and streams that provide us with water to drink, and the surfaces on and within our bodies. Pace’s approach, applied to many environments, led many other microbiologists to realize that many phyla in the natural world are waiting to be discovered. Indeed, these ongoing efforts to learn more about microorganisms from so many different environments continue to alter our views of the tree of life (Fig. 1 and 2).

Advances in DNA sequencing techniques made it possible to move from analyzing rRNA genes to entire genomes. Although this approach, called metagenomics, is based on random shotgun DNA sequencing of microbial communities and has proved powerful for interrogating microbial communities, it does not provide information about which genes are being expressed at any specific moment. However, whole-community transcriptomics and proteomics are helping to address questions about gene expression. Here we highlight how such whole-community omic analyses provide unique perspectives of microbial diversity and discuss the importance of using de novo assembly to analyze this type of data.

Understanding Microbial Communities in the Wild

Nearly a decade ago, Jill Banfield of the University of California, Berkeley, and her collaborators applied random shotgun DNA sequencing to microbial biofilms from extremely acidic waters in the Iron Mountain mine in California, assembling genomes for the most abundant of those microbes. Although they relied on Sanger sequencing, whose yields are modest compared to other DNA-sequencing techniques that are now available, their efforts showed that genome sequences could be reconstructed directly from environmental samples, bypassing cultivation while

SUMMARY

➤ Investigators are deducing genomes, transcriptomes, and proteomes of novel microbes, based on analysis of DNA sequence data from complex environmental samples.

➤ Because genome sequence data fall short of covering the microbial diversity in nature, database-dependent approaches to analyzing that diversity have important limitations and biases.

➤ Recent glimpses, using omics approaches, into the uncultured microbial biosphere have started to reveal fascinating organisms that expand our spectrum of knowledge of biology.

➤ Using whole-community omics to track microbial communities in nature can resolve the roles of novel uncultured groups and shed light on fundamental links between ecological and evolutionary processes.
providing insights into the uncultured members of that microbial community.

Subsequent studies illustrate that metagenomics contributes on at least three distinct levels to our understanding of microbial diversity. Taken together, these three levels provide insights into diversity that are inherent to microbial communities in the wild and that are readily tracked by whole-community genomic approaches.

First, metagenomic analysis reveals entirely new microbial groups that might otherwise be overlooked if one were relying on rRNA or other traditional analytic techniques. For example, Banfield and her collaborators in 2006 detected groups of archaea called ARMAN (archaeal Richmond Mine acidophilic nanoorganisms), that were not previously recognized because their 16S rRNA genes are mismatched with standard archaeal PCR primers.

Second, metagenomics highlights how even well-known microbial taxa harbor novel genes, which sometimes are located within hypervariable regions called genomic islands. Such novel genes may be expressed at high levels in particular environments, suggesting that they are important even if their functions are not known, according to Edward DeLong of the Massachusetts Institute of Technology in Cambridge and his collaborators.

Third, metagenomics can reveal genomic variability at the strain level within natural populations of bacteria, archaea, and viruses.

**FIGURE 1**

Phylogenetic tree reflecting our knowledge of the diversity of Bacteria. Sequences are clustered into bacterial phyla based on the Silva classification with manual curations and additions.
Such variability between slightly different strains can contribute in important ways to community functions.

**Data Assembly Is Critical when Analyzing Microbial Communities**

Investigators now have two main options for analyzing whole-community genomic and transcriptomic datasets. One approach involves comparing individual DNA sequences, or reads, to those in available databases. This approach helps them to determine the function of specific genes and the organisms in which those genes are functioning (grey arrows, Fig. 3).

One ongoing problem with taking this “read mapping” or “recruitment” approach is that many or, in some cases, most of the reads may fail to match sequences in public databases such as GenBank and MG-RAST. Although these and other genomic databases are growing rapidly, they vastly underrepresent microbial diversity in nature. Genomes from cells grown in culture or other sources that populate such databases are mere snapshots of the microbial world, and fall short of the full genomic repertoire. Put another way, much of the DNA and RNA sequence data from environmental samples is novel and thus unlike what is available in major databases.

There are at least two critical challenges facing anyone interpreting reads that do have matches in databases. First, comparisons of DNA sequences from the environment to those in databases typically use low thresholds of sequence similarity, or BLAST scores, to define positive matches. This approach can result in finding considerable diversity within reads mapping to single database sequences (Fig. 3). For example, we found that many bacterial reads readily map to an archaeal genome with commonly used pa-
FIGURE 3

Schematic of approaches for analyzing whole-community genomic and transcriptomic sequence data, database-dependent (on top, grey arrows) and assembly-based (on the bottom). Multi-colored lines represent DNA sequences (“reads”) from different microbial genomes. The thickness of arrows roughly represents number of reads. Database mapping and searching of both DNA (genomic) and RNA (transcriptomic) reads is biased by the completeness of the database being used to map and assign function to the reads. Both transcriptomic and genomic assemblies enable the reconstruction of natural community members and genes that are commonly overlooked by database-dependent approaches.
rameters. Second, inferring functions of genes based on environmental reads can be difficult because read lengths are considerably shorter than full-length genes.

The other major approach to analyzing whole microbial communities involves assembling individual reads into ever-larger fragments on the basis of sequence overlaps (Fig. 3). While this approach may seem daunting, especially when working with diverse microbial communities in which sequence coverage of individual taxa may be limited, it can prove fruitful, and it is now commonly used to reconstruct near-complete genomes for microorganisms within samples from environments such as seawater and groundwater.

This analytic approach can lead to useful insights. For instance, genes that are linked via the genome can also reveal how genes associate on finer scales, such as operons, providing potential insights into how those genes function. Moreover, this approach enables investigators to predict what metabolic pathways may be functioning within uncultured organisms. Perhaps most importantly, metagenomic assembly can help investigators to piece together novel genomic regions as well as full genomes for microorganisms that are absent from databases. Even sets of fragments that do not belong to other near-complete genomes can be linked, or “binned,” with one another on the basis of signatures such as tetranucleotide frequencies. However, binning is not currently possible when relying on individual reads.

Metagenomic assemblies can also enhance analyses of parallel metatranscriptomic data. We took this approach to look at ammonia-oxidizing archaeal populations within deep-sea hydrothermal plume communities from the Gulf of California. From these samples, we assembled several similar genotypes of archaea and could differentiate the transcriptional activity of each of these types under different regimes of ammonium concentration. Thus, sequences from different organisms can be assembled and used to estimate their abundance and transcriptional activity in the community (Fig. 4). Without assembly to resolve these different types, all of those reads might be assigned mistakenly to only one or a few entries in public databases, thus collapsing queries into database-dependent compartments that do not fully reflect the diversity within the natural environment.

Complementary metagenomic and metatranscriptomic datasets can also be used for assessing...
the relative abundance and activity of different microbial populations. We find extreme differences in DNA and RNA compositions from the same community. Some of the most abundant mRNA sequences belong to rare members of the community.

Relying on metagenomic data when analyzing metatranscriptomics has pitfalls, however. For example, novel, highly expressed genes from low-abundance organisms may not be captured in metagenomic datasets or public databases. To better characterize these novel mRNA sequences, we assembled transcript libraries from the deep sea to reconstruct whole operons from uncultured microbial populations. Some of the most abundant transcripts are those from rare—that is, accounting for less than 1% of the community—bacteria that are involved in oxidizing nitrite. There is value in doing transcriptomic and genomic assemblies in parallel.

**Filling in the Tree of Life**

During the past several decades, rRNA gene surveys have uncovered a substantial number of new microbial phyla or divisions. Few members of these phyla have been cultured; thus, our knowledge of them is limited to abundances of single marker genes in the environment and, in some instances, visualization of cells with rRNA-targeted fluorescent probes.

How do we begin to understand the physiology and ecology of these mysterious lineages, which some experts call biological dark matter?
In lieu of traditional culture-based analyses, genome sequence analysis of environmental samples is providing insights into their lifestyles and evolutionary history.

A few years ago, for example, we determined the genomic sequences of the members of two uncultured archaeal phyla, now referred to as Parvarchaeota (two genomes, ARMAN-4 and 5) and Micrarchaeota (ARMAN-2). The latter sequence was obtained directly from whole-community assemblies, while the two Parvarchaeota sequences depended in part on enriching our samples for ultrasmall (less than 500 nm in diameter) cells. Recognizing that their genomes indicated that these cells interact with others, we determined that other nearby archaeal species mysteriously penetrate their cell walls.

This degree of genomic information remains a rarity for Archaea, most of whose phyla are poorly characterized (Fig. 2). However, massive sequencing of natural microbial communities is bearing fruit more quickly in terms of furnishing genomes for candidate Bacteria phyla. For example, Banfield and her collaborators recently assembled genomes for members of several such phyla, including OD1, BD1-5, OP11, and a new group PER (Fig. 1).

Single-cell genomics (SCG) is another approach to filling gaps in the tree of life. SCG depends on isolating individual cells from microbial communities, and then amplifying individual whole genomes before determining their DNA sequences. Of the known Bacteria phyla, many uncultured phyla are published and others are in the works. Some of the novel bacterial genomes, including OP1, OP9, and OP3, as well as several archaea such as Misc. Crenarchaeota Group and Marine Benthic group D, were determined by SCG.

Although powerful, SCG provides mere snapshots of the genomic content of single cells. High-throughput methods make it possible to take many snapshots, but only community-wide approaches provide a direct measure of their full diversity. This diversity can be tracked further using community transcriptomic and proteomic approaches. Ultimately, community genomics and SCG are complementary, with the latter providing references for fragment recruitment and nucleotide compositional binning, while metagenomic data can provide quantitative information and a wider lens for addressing questions concerning whole communities and their dynamics.

Outlook

Microbiologists now have the analytic tools to reconstruct genomes, transcriptomes, and proteomes of entire microbial communities. Because so much of what is out there is uncharacterized, we are facing many challenges before we more fully understand how novel genes and proteins from members within such communities function and how they determine the ecological roles of novel phyla.

Now that we can also quantify gene expression in nature and how it varies in different niches, we can begin to assess how these novel microorganisms within communities behave and interact at the system level. Such data may yield clues about how to culture cells within novel microbial groups, leading to experiments that will link gene sequences to their functions. However, even if we could culture all members from a particular microbial community, removing them from that environment will change how each of them behaves. Moreover, because ecology and evolution are linked, our efforts to understand the latter will very much depend on our ability to track microbial genomes and their expression patterns in complex assemblages in nature.

Brett Baker is a Ph.D. student and Gregory Dick is an assistant professor at the University of Michigan in the Department of Earth and Environmental Sciences.

Acknowledgments

We thank Jillian Banfield for insights and sequences for the phylogenetic trees. This article is funded by grants from the Alfred P. Sloan Foundation (BR2013-027), the Gordon and Betty Moore Foundation (GBMF2609) to Dr. Gregory Dick, and by a University of Michigan Rackham Pre-doctoral fellowship to Brett Baker.

Suggested Reading


