**MINIREVIEW**

**Linking bacterial identities and ecosystem processes: can ‘omic’ analyses be more than the sum of their parts?**

Sergio E. Morales1,2 & William E. Holben1,3

1Microbial Ecology Program, Division of Biological Sciences, The University of Montana, Missoula, MT, USA; 2Department of Crops and Agroenvironmental Sciences, College of Agricultural Sciences, University of Puerto Rico, Mayagüez, Puerto Rico; and 3Montana – Ecology of Infectious Diseases Program, The University of Montana, Missoula, MT, USA

**Correspondence:** Sergio E. Morales, Microbial Ecology Program, Division of Biological Sciences, The University of Montana, Missoula, MT 59812-1006, USA. Tel.: +1 406 243 6365; fax: +1 406 243 4184; e-mail: sergio.morales@mso.umt.edu

Received 23 November 2009; revised 24 May 2010; accepted 14 June 2010. Final version published online 19 July 2010.

DOI: 10.1111/j.1574-6941.2010.00938.x

**Editor:** Ian Head

**Keywords**
metagenomics; metatranscriptomics; quantitative PCR; ecosystem processes; genomes; stable isotope probing.

**Abstract**

A major goal in microbial ecology is to link specific microbial populations to environmental processes (e.g. biogeochemical transformations). The cultivation and characterization of isolates using genetic, biochemical and physiological tests provided direct links between organisms and their activities, but did not provide an understanding of the process networks in situ. Cultivation-independent molecular techniques have extended capabilities in this regard, and yet, for two decades, the focus has been on monitoring microbial community diversity and population dynamics by means of rRNA gene abundances or rRNA molecules. However, these approaches are not always well suited for establishing metabolic activity or microbial roles in ecosystem function. The current approaches, microbial community metagenomic and metatranscriptomic techniques, have been developed as other ways to study microbial assemblages, giving rise to exponentially increasing collections of information from numerous environments. This review considers some advantages and limitations of nucleic acid-based ‘omic’ approaches and discusses the potential for the integration of multiple molecular or computational techniques for a more effective assessment of links between specific microbial populations and ecosystem processes in situ. Establishing such connections will enhance the predictive power regarding ecosystem response to parameters or perturbations, and will bring us closer to integrating microbial data into ecosystem- and global-scale process measurements and models.

**Introduction and overview**

The term ‘metagenomics’ was first used over a decade ago (Handelsman et al., 1998) and, indeed, was first proposed (Pace et al., 1985) and applied even earlier (Schmidt et al., 1991). Metagenomic approaches have enabled us to understand the genomic potential of the entire microbial community in an ecosystem by cloning and analyzing microbial community DNA directly extracted from environmental samples. This approach bypassed PCR, thus reducing the potential bias introduced by amplification, allowing for the direct sampling of all DNA within an environment. Subsequent sequencing of cloned inserts enabled large genome fragments to be analyzed, providing insights into the genomic organization, regulation and potential function of unknown organisms (Stein et al., 1996). Yet, due to the vast diversity of microorganisms encountered in many environments, this approach was limited due to the sequencing technologies available at the time. For example, a conservative estimate indicated that for a typical surface soil community, 10.6 Mbp (million base pairs) of sequence (~106 BAC clones with 100 kbp inserts) would be needed to capture its metagenome effectively (Handelsman et al., 1998). The large amount of cloning and sequencing required for such a task was prohibitive during the early development of metagenomics. More recently, researchers have transitioned to new direct and high-throughput sequencing technologies (e.g. 454 pyrosequencing), often bypassing cloning steps that were essential previously. Direct sequencing increases the depth of microbial community DNA sequences analyzed, probing deeper into the metagenome of an ecosystem, and to date, has been applied to a variety of
environments ranging from the termite gut (Warnecke et al., 2007), to the human intestinal tract (Gill et al., 2006), to marine environments (Nealon & Venter, 2007). This direct, deep-sequencing, metagenomic approach has been accepted widely and has even been termed ‘omic’ in acknowledgement of the amount of effort required to achieve a comprehensive coverage (Handelsman, 2005). Despite vast sequencing efforts, the complete coverage of a metagenome based on multiple-fold redundancy or the complete assembly of individual genomes within a community remains largely unattained. The exceptions have been extremely simple communities such as those existing under acid mine drainage conditions (Tyson et al., 2004). Such difficulties are expected, given the typical microbial community complexity and unevenness (i.e. a few numerically predominant populations and vast numbers of low-abundance ones). To achieve complete sequencing coverage, the sequence library analyzed must surpass the size of the metagenome of the community by 100–1000-fold, especially if information about minority populations within the community is desired (Riesenfeld et al., 2004). If we apply this principle to a bacterial community with just 200 unique species (total metagenome of 1 Gbp based on an average 5 Mbp per genome), a library of 100 Gbp would be required for the comprehensive coverage of the metagenome, and this under the assumption of an equal representation of all 200 taxa, which is unrealistic for most natural microbial communities. Given that (1) most microbial communities studied to date exhibit variable evenness (Sogin et al., 2006), (2) the species-level diversity of microorganisms in most environments is estimated to be in the range of hundreds to thousands (Eckburg et al., 2005; Hong et al., 2006; Morales et al., 2009), and (3) there are errors and biases introduced by DNA extraction methods (Frostegard et al., 1999; Martin-Laurent et al., 2001; LaMontagne & Michel, 2002; Yang et al., 2007; Feinstein et al., 2009; Inceoglu et al., 2010), even 100 Gbp is a gross underestimate of the sequencing depth needed to meet the challenge of obtaining full coverage in environmental microbial metagenomics. Despite these challenges, environmental metagenomics has advanced and is now moving into its next logical progression, spawning the subfield of ‘metatranscriptomics’. In metatranscriptomics, a direct cDNA cloning approach is applied to delve into the composition of RNA transcripts in environmental samples, thereby detecting active genes undergoing transcription. This approach provides certain advantages over DNA-based metagenomics, the most important being the reduction of community complexity by focusing on the active populations in a sample. Initial studies explored the active functional diversity of marine and freshwater bacterioplankton communities (Poretsky et al., 2005). In doing so, they paved the way for subsequent larger studies, one focusing on eukaryotes in forest soils by cloning double-stranded cDNA (Bailly et al., 2007), and later by direct sequencing of reverse-transcribed total community RNA from soil (Urish et al., 2008). As with the original work using metagenomics, metatranscriptomics has led to novel findings including unique small RNAs (Shi et al., 2009). The ability of nucleic acid-based ‘omic’ approaches (hereafter referred simply as ‘omic’) to delve into the unknown genetic potential of microorganisms has spurred an increase in the use of these techniques. Coupling of direct ‘omic’ approaches to direct high-throughout sequencing is producing massive amounts of data, replacing the previous limitation of sequencing depth with the challenge of analyzing terabytes (trillions of bytes) of sequence information (Snyder et al., 2009). A comprehensive discussion of the potential and the limitations of the nascent area of microbial community analysis through metagenomics and metatranscriptomics would be an excellent focus, and indeed multiple articles are available (Riesenfeld et al., 2004; Tringe & Rubin, 2005; Warnecke & Hugenholtz, 2007; Cardenas & Tiedje, 2008; DeLong, 2009; Warnecke & Hess, 2009). In this review, we focus on the shared benefits and shortcomings of sequence-based ‘omic’ approaches in general. We also discuss the power and potential of integration of multiple techniques aimed at linking specific bacterial populations (and ultimately whole communities) and the ecosystem processes/services they perform. Lastly, we attempt to envision potential future directions in molecular microbial ecology.

**Advantages and successful applications of nucleic acid-based ‘omic’ approaches in microbial ecology**

Previous reviews have discussed the advantages of ‘omic’ approaches (Riesenfeld et al., 2004), and the potential insights gained by their application (Tringe & Rubin, 2005). Both metagenomics and metatranscriptomics have been heralded as technologies that can facilitate the exploration of the unculturable majority of microorganisms, their metabolic capabilities and functional roles. In doing so, we believed the information would guide cultivation and enrichment techniques. In practice, ‘omic’ approaches, in particular metagenomics, have been most successful when applied to simple communities such as those present in acid mine drainage environments. Cloning of DNA from acid mine drainage led to the reconstruction of the near-complete genomes of Leptospirillum group II and Ferroplasma type II (Tyson et al., 2004). By creating these whole-genome reconstructions, the evolution of metabolic and regulatory networks within the context of a known ecosystem can be better understood. Direct analysis of genomic or transcriptomic material also eliminates the potential for the intense
selective pressure imposed by laboratory cultivation, which can produce phenotypes alien to the original ecosystem (Elena & Lenski, 2003). Yet, despite increases in the depth of sequencing, the application of ‘omic’ techniques to highly complex ecosystems using randomly based ‘shotgun’ approaches produces sequence fragments, which, although sufficiently redundant to enable identification of genes or membership in gene families, typically lack sufficient overlap for genome assembly efforts (especially in the absence of a framework of reference genomes on which to align sequences from environmental samples).

A more focused gene- or process-centered approach could still be quite powerful. One of the most undervalued uses of ‘omic’ approaches is that of being a ‘discovery tool’ when applied to simple and targeted systems. Warnecke et al. (2007) used fosmid libraries generated with prokaryotic DNA extracted from the hindgut of termites and focused on cellulose hydrolysis-associated genes. Using this approach, they identified putative functional genes and expressed them in vitro, allowing for the confirmation of their functional role. They further linked their metagenomic data to metaproteomic data (proteomic analysis with clarification of gut fluid from the same sample) demonstrating in vitro expression of several cellulosases and xylanases, in addition to putative sugar-metabolizing proteins. Using a conservative estimate based on the detection of complete protein-coding domains, over 100 gene modules related to cellulose hydrolysis were identified. These modules contained over 700 glycoside hydrolase catalytic domains belonging to 45 different families including putative cellulosases and hemicellulases. Not only did they link a gene to a specific process, they went on to use the assembled metagenomic data to link specific functions (i.e. ecosystem processes; in this case, lignocellulose degradation in the termite hindgut) to specific bacterial populations. This termite hindgut system was dominated by diverse species of Treponema (Spirochaetes), with species of the phylum Fibrobacteres being linked to putative genes encoding endoglucanases and nitrogenases. Although a relatively simple system when compared with soils or oceans, it highlights the power of combining approaches toward a specific target.

The pursuit of specific targets (e.g. families of genes) is particularly effective and has been applied in the search for novel natural products with biotechnological potential (e.g. enzymes or antimicrobials). Although novel natural product discovery may seem tangential to microbial ecology in its purest sense, such discoveries prompt subsequent analyses of their ecological importance and effects. They also serve as a good example of how targeted approaches can yield discoveries within well-defined contexts. Metagenomic analyses have revealed novel gene products, including biocatalysts (enzymes), small sensory molecules and other bioactive molecules (Rondon et al., 2000; Daniel, 2004; Steele et al., 2009). As illustrated above, the ability of ‘omic’ approaches to probe into the unknown functional capabilities of microorganisms has garnered interest beyond microbial ecology, with particular interest from scientists looking for industrial applications to microbial products (Lorenz & Eck, 2005).

Although very well suited for identifying products with industrial or commercial potential, screening of activities encoded within metagenomic libraries highlights a very efficient method of linking a particular process to a given gene. Wang et al. (2000) isolated four novel natural products by expressing large fragments of metagenomic DNA in a Streptomyces surrogate host by combining biological and chemical screenings. Metagenomic studies in other systems have led to the discovery of antitumor molecules. The antitumor molecule pederin was first identified in metagenomic analyses of microorganisms associated with the Paederus beetle, and was determined to be a polyketide produced by an uncultured Pseudomonas symbiont (Piel, 2002; Piel et al., 2004a, b). Similar work with the marine sponge Theonella swinhoei identified polyketide biosynthesis genes closely related to those needed for pederin biosynthesis (Piel et al., 2004a, b). As was the case for pederin, these genes were encoded by an uncultured bacterial symbiont. Although the discoveries of both symbionts and antitumor molecules are noteworthy, the fact that these discoveries were linked to symbionts belonging to the genus Pseudomonas and most closely related to Pseudomonas aeruginosa also highlights the power of this approach. Despite pseudomonads being one of the best-represented organisms in culture collections and the apparent ubiquity of culturable P. aeruginosa strains, pure culture isolation is not always tractable for all strains, thus limiting the rate of novel discoveries using traditional culture-based approaches. This can be overcome using ‘omic’ approaches, but without targeted screening, such discoveries can be lost within the giga- to terabytes of data produced today. This targeted approach also allows for direct links between processes and populations, a topic to be discussed in the following sections.

Despite the power of these screening methods, they are still in their infancy, and as with most techniques, there are challenges. Screening methods are primarily affected by the choice of the host, which has been discussed elsewhere (Uchiyama & Miyazaki, 2009). The lack of suitable hosts for expression limits the potential for expressing every gene harbored within a collection. The use of alternate hosts will likely yield novel products and new avenues of research, bringing us back to our reliance on culturing approaches to provide new organisms that can serve as hosts.

Current state of technology

The increase in the use of nucleic acid-based ‘omic’ approaches is intrinsically linked to recent developments in high-throughput, highly parallel sequencing platforms.
These technologies have surpassed early metagenomic coverage predictions producing sequence libraries numbering in the hundreds of thousands to over a billion reads (Mardis, 2008). Three main sequencing systems have been in the forefront of recent sequencing innovations: the Roche/454 FLX System (also referred to as pyrosequencing) (Margulies et al., 2005), the Illumina/Solexa Genome Analyzer System (Bentley, 2006) and the Applied Biosystems SOLiD™ System (Ondov et al., 2008). Their outstanding success results from their ability to produce massive amounts of data at a fraction of the cost of traditional cloning and Sanger dideoxy sequencing approaches. However, these novel sequencing platforms come with other costs relative to the Sanger approach: namely shorter reads and higher error rates. Although highly attractive for their economic and time efficiencies, some of the limitations of these new sequencing platforms thus give rise to new problems (discussed in subsequent sections) to ‘omic’ approaches. These limitations have not slowed the application of these techniques, with many research groups acquiring ‘omic’ data from a variety of both simple (e.g. hot springs) and complex (e.g. soil, human microbiome) systems. As of September 2009, there were 79 completed metagenomic studies and 129 in process (refer to the Genomes Online Database at http://genomesonline.org/index.htm). While some of the successes with assembling metagenomic data from simple systems have already been noted, meaningful assembly of genomes or even lengthy contigs from complex microbial communities will require much more in-depth sequencing efforts than have been performed to date. Larger sequencing projects will also require the development of novel computational approaches and enhanced capabilities over what are currently available to the scientific community [e.g. cloud computing (Bateman & Wood, 2009; Field et al., 2009; Fricke et al., 2009; McPherson, 2009)].

**Limitations of the current ‘omic’ approaches in microbial ecology**

**Size matters**

A primary limitation of all three high-throughput sequencing systems is the relatively small read length generated at sequencing. Currently, 454 pyrosequencing reliably offers ~425 bp reads and an approximately 1% error rate; Illumina generates reads in the 125 bp range with a similar error rate, while the Applied Biosystems SOLiD™ System offers 25–35 bp reads and a substantially higher error rate (Bentley, 2006; Huse et al., 2007; Dohm et al., 2008; Shendure & Ji, 2008), although new chemistries and platforms are already increasing these limits. Higher error rates are a relatively minor problem when sequencing single genomes where multifold coverage enables error reduction through the building of a consensus sequence. The same error rate becomes a significant obstacle with environmental ‘omic’ approaches in which achieving even onefold coverage of a mixture of genomes is unlikely. The second, and in our opinion most important limitation, is sequence length. Wommack et al. (2008) compared phylogenetic and functional placements of metagenomic data based on size for 100–750 base lengths of DNA fragments. In this analysis, reads ≥ 600 bp were processed *in silico* to simulate the read lengths and depth of coverage typically associated with analyzing a microbial or viral metagenome using high-throughput, short-read sequencing approaches. They demonstrated reduced ability to detect homologous sequences in GenBank when query sequences were short (ranging down to 100 bases). The same study also suggested that increasing sampling depth (which can be achieved using high-throughput sequencing) or increasing read length up to 400 bases, the longest reported read length for any technique at that time (Schuster, 2008), still did not allow for reliable placement of homologous sequences compared with the longer reads. Only in cases where strong homology to previously described genes was found did short reads reliably identify sequence homologs. This is especially of concern, given that the study was based on a 100 base-read minimum, while the SOLID approach typically produces read lengths of ~33 bases. Many studies have been published that tally small reads into bins based on putative functions. However, recent trends have moved toward assembled reads in an attempt to increase the overall sequence length, a strategy that would help address the problem of short reads and increase the information linked to each gene, but it has not been simple. Short read lengths have other limitations. In order to cover the length of a whole genome, even when using a reference genome for scaffolding, many overlapping sequences have to be present when using short reads. The computational capabilities required to process and assemble genomes generated from these enormous numbers of short fragments has become the limiting factor, although progress is being made [see Nature Methods supplement “Focus on next-generation sequencing data analysis” (http://www.nature.com/nmeth/journal/v6/n11s/index.html)]. Before even attempting to assemble genomes, sequence processing must undergo image analysis, signal processing, background subtraction, base calling and quality assessment (Mardis, 2008), each with a bias associated with fragment size. Established protocols have been based on existing approaches and capabilities designed on premises more relevant to previous longer read technologies. The need for the corresponding enhancement of computational approaches to support the new sequencing technologies and the short, but extremely numerous sequence reads they produce, has led to the creation of specific programs capable of dealing with the
limitations of high-throughput sequencing technologies. Until advances in these technologies increase the length or decrease the error rates in sequence reads, users might take advantage of the many genome assembler programs that have been developed specifically to take advantage of the depth of coverage produced by the new high-throughput sequencing platforms to compensate for the higher error rates (a selection of recent programs can be found in Dohm et al., 2007; Jeck et al., 2007; Sundquist et al., 2007; Zerbino & Birney, 2008).

**Unknown functions**

The massive amounts of data generated by large-scale sequencing efforts and made widely available through the Internet and dedicated servers seem to promise a new era of metagenomics and metatranscriptomics. It is reasonable to anticipate that these data will provide insights into environmental processes and allow the detection of active microbial populations (and their actively transcribed genes) selected for by the environment, and this at a total community level rather than based on individual populations. To date, though, what has been accomplished can be generally characterized by very large collections of sequences, mostly unassembled, that are most often used to assign putative functions to genes via *in silico* comparisons, a task for which short reads have been shown to be poorly suited (Wommack et al., 2008). A common outcome from these collections is the cataloguing of numerous uncharacterized genes, a large portion of them identified as being among the most predominant in that environment (Tringe et al., 2005; Gill et al., 2006; Gilbert et al., 2008). This outcome is a product of the low genomic diversity currently represented in sequence databases that have historically relied on cultured organisms. As is well documented (Rappé & Giovannoni, 2003; Floyd et al., 2005; Janssen, 2006; Jones et al., 2009; Spain et al., 2009), uncultured microorganisms represent the vast majority of microbial diversity, and so these databases grossly underrepresent the functional potential of microorganisms. This is changing rapidly as a result of the numerous current microbial genome-sequencing efforts supported by such programs as the DOE-supported Genomic Science Program (http://genomicscience.energy.gov/) and the Genomic Encyclopedia of Bacteria and Archaea Project (http://www.jgi.doe.gov/programs/GEBA/), as well as the NIH-sponsored Human Microbiome Project (http://nihroadmap.nih.gov/hmp/).

Unclassified genetic diversity may also arise from 'genomic erosion'. In this phenomenon, genomic 'debris' (e.g. pseudogenes and otherwise inactive genes or gene segments) can be detected and erroneously represented as operational genes. Despite, or most likely due to, evolutionary pressures that control genomic composition, bacterial genomes have been shown to encode fewer operational genes than originally predicted (Ochman & Davalos, 2006). That is, predicted ORFs putatively coding for a protein have a fairly high likelihood of being pseudogenes (nonfunctioning versions of protein-coding genes). This occurs due to a mutational bias that allows nonfunctional regions to be maintained in the genome for some time until genetic drift and other evolutionary pressures eventually eliminate the entire gene. Although it is highly unlikely that all novel genes discovered via 'omic' efforts represent genomic debris, assignment of gene function to recovered sequences based solely on homology to known functional genes should be performed and interpreted with care. As with many of the challenges already discussed, the distinction between gene erosion products vs. true metabolic potential and suggested expression of mRNA sequences can be made with high confidence only if larger sequences can be assembled from the relatively short sequence reads currently being obtained. Linking organisms to environmental processes and ecosystem services is a primary goal of the microbial ecology community (Buckley & Roberts, 2007). However, it seems that due to the complexity of most microbial consortia, ‘omic’ approaches without further advancements in cultivation techniques or other integrated methodologies are at this time insufficient for unraveling their myriad physiologies (Janssen, 2006).

**Jumbled genomes**

As noted above, nucleic acid-based ‘omic’ approaches are most informative when the source of DNA or RNA is well known (e.g. from an environmental isolate or enrichment) or community genomic complexity is low (e.g. low-diversity environments), both exceptions rather than the rule when studying communities of microorganisms. Processing community nucleic acids for shotgun-based sequencing produces a difficult challenge; comprehensive extraction of nucleic acids from a sample, although typically the desired outcome, leads to a highly complex mixture of nucleic acid fragments that are further fragmented through the generation of short sequencing reads or cloning and sequencing of small fragments. This high degree of fragmentation of previously contiguous sequence removes much, if not all, of the valuable genetic context of the surrounding sequence that must be restored by assembly. However, effective assembly requires sufficient fragment overlap to assemble contiguous sequences that are both correct and informative with regard to genome structure. It is at this level of assembling large sequences from small fragments derived from complex DNA or RNA pools that appreciation and interpretation of the true genetic variability inherent in the system is undermined (Raes et al., 2007). In the case of highly divergent bacterial lineages, identification of fragments and their classification into groups are relatively
simple. However, intraspecies and even intragenus variability is likely overlooked in the assembly process due to the complexity of the sample template and the lack of sufficient sequence coverage to allow reliable contig assembly and disentanglement of mixtures of highly or even poorly related genomes (Vietes et al., 2009).

Although organisms of the same species share many genetic and phenotypic characteristics, they are not necessarily identical. Studies have explicitly demonstrated substantial genomic variability within a single species, not only at the single gene level, but over the entire genome (Welch et al., 2002). Related species or subspecies share what is termed a pan-genome, named to represent a global gene repertoire of a bacterial species: core genome + dispensable genome (Medini et al., 2005). The core genome represents the pool of genes shared by all strains of the same bacterial species, while the dispensable or the accessory genome represents the pool of genes present in some, but not all, strains of the same bacterial species. Studies in two separate genera have shown a conserved portion of genome within a single genus ranging from 42% to 80% of the genome, with the remainder found in only some of the component species of that genus (Welch et al., 2002; Tettelin et al., 2005).

For environmental ‘omic’ approaches, this genomic phenomenon raises the question of how to assign identity to randomly mixed small sequence fragments in the presence of high genomic and transcriptomic diversity, especially under high sequencing error rates. At least two scenarios arise in current research that can lead to erroneous genome or transcriptome reconstructions: (1) multiple related strains with subtle genomic/transcriptomic variability are present, but are not recognized as individual strains or transcripts in the assembly process, thus abolishing the detection of species-level genetic richness, (2) multiple different genome arrangements are present and real, but one is selected over the others based on the order of discovery. The latter phenomenon is similar to the errors encountered when comparing divergent sequences using sequential pairwise alignments where the first analyzed sequence serves as a reference. These potential pitfalls can be avoided where relevant reference strains are available. Marine bacteria, like their counterparts in most ecosystems, are genetically diverse (Rocap et al., 2003), and thus susceptible to the problems discussed above when analyzed using ‘omic’ approaches. Zehr et al. (2007) overcame those problems, and assembled genome segments from metagenomic data. They compared the metagenome fragments to a reference strain and determined low genomic diversity ( < 1% nucleotide sequence divergence) among Crocosphaera watsonii strains. This approach depended on the availability of a sequenced genome from a cultured isolate and, as mentioned earlier, underscores our continued reliance on culture-based approaches. However, for this type of study, a reference genome could have been established in the absence of cultivation if the genome was correctly assembled by means of more directed approaches, which will be discussed later in this review.

Disconnected pathways

The noncontiguous sequence fragments resulting from small sequence fragments and the low coverage of a sample often lead to disconnected genetic elements (e.g. operons). As parts of an organized system, gene sequences alone provide little information in the physiological or the ecological sense without understanding the regulatory framework that controls their expression (Sorek & Cossart, 2009). Also, the surrounding sequences often provide a context to discern whether putative gene sequences are capable of producing a viable protein. Unless successfully assembled into sequences representing operons or other genetic organizational units, short sequence reads merely produce a collection of genetic fragments that may have high identity with parts of genes or gene families, but do not provide sufficient information to ascertain an ecological function or guarantee gene product functionality.

As in whole-genome assembly from sequence reads, considerable information can be derived from the chromosomal ‘neighborhood’ of a gene. Operons tend to harbor genes that are involved in a shared function (Lawrence & Roth, 1996). Assigning function based on a small fragment of a single gene is at best a guess, but when an organized collection of genes points to a shared function, its putative placement and role is bolstered. Although the physical proximity of genes is indicative of operons, the recognition of transcriptional regulation elements is key to defining true operons. Short sequence reads must be assembled into larger (multiple-kbp) contiguous sequences if one is to identify the regulatory elements that provide additional evidence for putative functions. While it is well established, for example, that the primary sigma (σ) factor in bacteria (belonging to the σ70 family) is essential for the expression of housekeeping genes under normal growth conditions (Paget & Helmann, 2003), accessory proteins and alternate sigma factors can specifically control transcriptional regulation, providing clues into roles and regulatory networks under defined conditions (Browning & Busby, 2004).

Providing the genetic context in environmental sequence analyses is key to developing better hypotheses linking specific genes to given functions, and for the direct interpretation and validation of hypothesized metabolic networks. By understanding regulatory processes and pathway organization, we are provided with additional information to judge important aspects of the ecology of microbial systems and their genetic repertoire such as: (1) How essential are certain genes for survival? (2) What conditions...
control their expression? (3) How are multiple pathways, potentially in different populations, interconnected and to what degree do they affect each other’s gene expression patterns? As it stands now, given the high microbial and genetic diversity in almost every environment, current capabilities in ‘omic’ approaches are seemingly insufficient to provide enough assembled contiguous sequences, much less entire genomes, to resolve such pathways, their regulation and their interactions in the greater context of total community function and ecology.

**New directions**

Linking ecosystem processes to individual bacterial species or guilds is one of the central goals in microbial ecology (Buckley & Roberts, 2007). Advancements in microbial community ‘omic’ approaches have focused primarily on two fronts: (1) development of advanced software capable of effectively processing the increasing sequence loads and shorter reads (Table 1) and (2) complementation or integration of ‘omic’ approaches with other methods that aim to either test hypotheses developed from initial studies or to reduce the complexity of total microbial community samples, thus allowing for the focused coverage of portions of a community (Fig. 1).

**Software improvements**

Within the last decade, large-scale ‘omic’ approaches have evolved to the point where the original primary limitation, namely depth of sequencing, has been replaced with the challenge of analyzing ever-increasing sequence datasets (both the size of individual datasets and the total collection of datasets being amassed through time). These computational challenges include the correct assembly of genomes, functional characterization of small genetic fragments and phylogenetic placement in the absence of traditional marker sequences (e.g. the 16S rRNA gene). Although the list and capabilities of available programs have grown prodigiously (Table 1), we will limit our discussion to new tools developed to aid in sequence data analyses aiming to overcome the above-noted challenges.

A huge challenge in linking functions to specific populations is the assignment of phylogenies to specific genetic fragments in the absence of markers such as the 16S rRNA gene. The use of phylogenetic anchors (Riesenfeld et al., 2004) had limited this approach to sequence fragments physically linked to the genes used for phylogenetic reconstructions. New programs have since focused on genetic patterns within sequences (e.g. GC composition) that point toward a common ancestor. McHardy et al. (2007) developed a program, **PHYLOPYTHIA**, that allows users to assign genomic fragments into phylogenetic clades with high accuracy. The program assigns sequences to classes according to the nucleotide composition and based on annotations of 340 completed genomes. **PHYLOPYTHIA** is designed to accurately classify fragments > 3 kb, but was also shown to be effective at assigning a higher-level (e.g. phylum level) classification to smaller fragments. The program’s designers note that complications arise from horizontal gene transfer (HGT). The program can correctly identify fragments of DNA acquired by HGT, having for example correctly assigned HGT-acquired portions of bacterial genomes as archaeal. This was possible due to the availability of large assembled genomic fragments, but this ability is reduced significantly when working with smaller fragments. Although the concept of a chimeric genome sequence due to assembly error was not studied, the authors posit that because the program relies on the depth of the training dataset (the number of available sequenced and annotated genomes), increased availability of sequenced genomes is required to increase accuracy. Fortunately, current efforts like that in the Human Microbiome Project to fully sequence 1200 microbial genomes from the human gastrointestinal tract will produce enriched datasets for at least certain environments in a relatively short order. Increased datasets from known source organisms will also benefit other software programs such as the **SEQUENCE ORTHOLOGY-BASED ANNOTATION USING SUBSYSTEMS TECHNOLOGY (MG-RAST)** (Meyer et al., 2009). This phylogenetic assignment tool improves similarity-based binning approaches by using more parameters than just the usual bit scores. By first identifying an appropriate taxonomic level of assignment, followed by an orthologous target search, it increases assignment accuracy, especially for sequences originating from new species and genera that would otherwise go unclassified. This increase in placement ability is welcome, but so little is known about the level of HGT in most ecosystems that the most desirable approach for the classification of sequences remains the assembly of large genomic fragments.

Processing of large amounts of data has also spurred the development of ‘pipelines’ that provide users with high-throughput analytical tools for ‘omic’ data. The Ribosomal Database Project (RDP) has provided tools for the analysis of small subunit rRNA, and has recently developed a pipeline dedicated to the analysis of ultra high-throughput rRNA sequencing data (Cole et al., 2009). **METAGENOME RAPID ANNOTATION USING SUBSYSTEMS TECHNOLOGY (MG-RAST)** (Meyer et al., 2008) is another pipeline providing for the general analysis of genomic data (i.e. including, but not limited to rRNA sequences). **MG-RAST** is an open-source server created from a modification of the original **RAST** server, itself designed to produce high-quality annotations of both complete and draft microbial genomes. **MG-RAST** is a fully automated service and provides users with annotated sequence fragments, their phylogenetic classification,
metabolic reconstructions and access to comparison tools for multiple samples. Functional assignments of ORFs are performed by comparison with both protein and nucleotide databases [including the SEED framework (Overbeek et al., 2005), NCBI BLAST (Altschul et al., 1997), SQLite (http://www.sqlite.org/) and SUN GRID ENGINE (http://gridengine.sunsource.net/)], as well as the aforementioned RDP database and Greengenes (DeSantis et al., 2006) for rRNA sequences.

### Table 1. Software available for processing, analyzing and managing ‘omic’ datasets

<table>
<thead>
<tr>
<th>Program name</th>
<th>Access</th>
<th>Input</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAMERA</td>
<td>Website</td>
<td>NA</td>
<td>The CAMERA website is a central hub that serves as a repository of metagenomic datasets and bioinformatics tools. CAMERA is making accessible raw environmental sequence data, associated metadata, precomputed search results and high-performance computational resources.</td>
<td>Seshadri et al. (2007)</td>
</tr>
<tr>
<td>IMG/M</td>
<td>Server</td>
<td>NA</td>
<td>Data management and analysis system for completed (meta)genomes within the server. The system provides for data exploration and visualization tools, as well as comparative genome analysis.</td>
<td>Markowitz et al. (2008)</td>
</tr>
<tr>
<td>JCOAST</td>
<td>Freeware</td>
<td>NA</td>
<td>JCOAST allows users to analyze and compare (meta)genome sequences from default databases. Individual, cross genome-, and metagenome analyses are facilitated allowing for mining, annotation and interpretation of (meta)genomic data.</td>
<td>Richter et al. (2008b)</td>
</tr>
<tr>
<td>MEGAN</td>
<td>Freeware</td>
<td>Concatenated file of BLAST results for all targets.</td>
<td>Laptop-based program designed to compute and explore the taxonomical content of the metagenomic dataset based on the NCBI taxonomy. It allows for the general organization and visualization of data in the absence of contig assembly.</td>
<td>Huson et al. (2007)</td>
</tr>
<tr>
<td>METASIM</td>
<td>Freeware</td>
<td>NA</td>
<td>METASIM is a sequencing simulator that allows users to create a mock dataset based on reference genomes and adjustable software settings. These ‘synthetic’ reads can be used as proxies for the diverse taxonomical composition of typical metagenome data sets, and allow for standardized test scenarios for planning sequencing projects or for testing of metagenomic software.</td>
<td>Richter et al. (2008a)</td>
</tr>
<tr>
<td>MG-RAST</td>
<td>Server</td>
<td>Sequence data (raw unassembled or assembled contigs) in: FASTA format or 454 reads.</td>
<td>Online pipeline for analysis of metagenomic data capable of producing automated functional assignments of sequences by comparing both protein and nucleotide databases. It also allows for phylogenetic and metabolic reconstructions, and the ability to compare the metabolism and annotations of one or more metagenomes.</td>
<td>Meyer et al. (2008)</td>
</tr>
<tr>
<td>PHYLOPYTHIA</td>
<td>Online Server</td>
<td>Genomic fragments in FASTA format.</td>
<td>Composition-based classifier capable of accurately classifying genomic fragments into taxonomic ranks. Classification is based on 340 completed genomes in addition to up to 100 kb of training sequence from a sample-specific population. Fragments should be ( \geq 1 \text{ kb} ) for high-resolution classification, although smaller fragments can be classified at higher taxonomic ranks.</td>
<td>McHardy et al. (2007)</td>
</tr>
<tr>
<td>RDP's PYROSEQUENCING PIPELINE</td>
<td>Server</td>
<td>Sequence and Quality files in FASTA format.</td>
<td>The RDP website and the pyrosequencing pipeline provide a suit of tools for analyzing, processing and submitting 16S rRNA gene pyrosequencing data. Tools include an aligner, sequence classifier, rarefaction and diversity calculator.</td>
<td>Cole et al. (2009)</td>
</tr>
<tr>
<td>SORTITEMS</td>
<td>Freeware</td>
<td>BLASTX output from metagenomic reads.</td>
<td>Similarity-based binning software for classification of (meta)genomic data. It allows users to assign genes into bins based on similarity for cataloguing of large sequence datasets.</td>
<td>Monzoorul Haque et al. (2009)</td>
</tr>
<tr>
<td>STRAINER</td>
<td>Freeware</td>
<td>Fully assembled chromosome or genome, a contig from an assembler or the genome of a related organism. Reference sequences can be input as either FASTA or GenBank formatted files.</td>
<td>Software package for the analysis and visualization of genetic variation in populations and reconstruction of strain variants from otherwise coassembled sequences from metagenomes. The program reveals the degree of clustering among closely related sequence variants and provides a rapid means to generate gene and protein sequences for functional, ecological and evolutionary analyses. Source code available and allows for the incorporation of new algorithms.</td>
<td>Eppley et al. (2007)</td>
</tr>
</tbody>
</table>

NA, not applicable.
The MG-RAST server processes both assembled and unassembled sequences, and provides guidelines based on the goal of the study. Because its design is based on the SEED subsystem, MG-RAST’s placement accuracy is higher for core metabolism and pathogenesis (the foci of its training data).

**New strategies and approaches**

At this time, most recent ‘omic’ studies can be thought of as preliminary experiments during the development and maturation of new approaches and technologies. Further, much of the data that have been amassed can be thought of as a large set of observations in need of corroboration through the application of mechanistically different or synergistically complementary approaches. Studies that have succeeded in this have often been rewarded with novel discoveries supported by multiple approaches. Leininger et al. (2006) combined quantitative PCR (qPCR), reverse transcription qPCR (RT-qPCR) and pyrosequencing to demonstrate the activity of specific archaeal groups *in situ*. This study effectively linked and demonstrated the dominance of crenarchaeota over bacterial ammonia oxidizers in the soils studied. By focusing on the strengths of each technique and integrating approaches towards a focused target (i.e. ammonia oxidation), they managed to establish clear connections between populations and ecosystem-level functions. Other researchers, as described below, have applied an integrated approach combining multiple methods including stable isotope labeling, metagenomics, metatranscriptomics, qPCR and RT-qPCR, among others, to distill community information and to target subgroups.

A direct way to select for active populations consuming a given substrate has been the application of stable isotope labeling (hereafter referred to as stable isotope probing or SIP). Originally used for targeting sulfate-reducing bacteria (Sørensen et al., 1981; Parkes et al., 1989), SIP uses $^{13}$C-labeled growth substrates to link microbial processes to specific populations by the incorporation and the subsequent recovery of ‘heavy’ $^{13}$C-labeled cellular components (Boschker et al., 1998; Radajewski et al., 2003; Dumont & Murrell, 2005). Once incorporated, the ‘labeled’ components can be separated and used to identify populations utilizing the given substrate and incorporating the isotopically labeled carbon into cellular biomass (Radajewski et al., 2000; Manefield et al., 2002; Lueders et al., 2003a,b; Lu & Conrad, 2005; Vandenkoornhuyse et al., 2007; Chen et al., 2008). The ability to target active groups carrying out a particular task is strengthened when combined with metagenomics and high-throughput sequencing, facilitating the detection of even the low-abundance species that represent the vast majority of microbial biodiversity in most ecosystems.
SIP-based metagenomic analyses are also benefitted by the ability to pool sequencing templates from several samples using bar-coding (Hamady et al., 2008), which can allow high-throughput processing of samples taken using multiple substrates. SIP has also been used in combination with magnetic-bead capture hybridization to target rRNA in order to link substrate utilization to specific phylogenetic groups (Miyatake et al., 2009). Future uses will undoubtedly take advantage of metatranscriptomic approaches and will likely expand on substrates used to include, among others, nitrogen (Radajewski et al., 2003). The major limitations to SIP are technical, including unintended isolation of high G+C content $^{12}$C-DNA in the $^{13}$C-DNA fraction and specific experimental system limitations (Radajewski et al., 2003), but information is available to guide the new user (Gray & Head, 2001; Radajewski et al., 2003; Dumont & Murrell, 2005; Neufeld et al., 2007; Chen & Murrell, 2010).

Specific groups with functions not directly recognized through substrate utilization can still be targeted through molecular sieving. This refers to the retrieval of a subpopulation from the total community, in this case by means of selective or targeted meta-omic approaches. The studies discussed hereafter have used various methods of enriching specific targets within complex mixtures of total community nucleic acids. While direct cloning and screening by hybridization can be used in support of targeted gene sequencing, it also involves the generation of large clone libraries capable of providing the depth of coverage to ensure adequate representation of populations, especially those present in low abundance (Demaneche et al., 2009). Subtractive hybridization approaches present an alternative to reduce the time, resource requirements and expenses posed by direct cloning and screening by hybridization. Subtractive hybridization has been applied successfully to total environmental RNA to minimize the recovery of highly abundant rRNA sequences, effectively enriching for mRNA transcripts in samples before RT-PCR and cDNA cloning in metatranscriptomics (Shrestha et al., 2009).

Subtractive hybridization has been further evolved by a competition assay called genome fragment enrichment (GFE), in which DNA fragments are isolated from one metagenomic DNA sample (target) by hybridizing it to another reference sample to which it is being compared (Shanks et al., 2006, 2007). This approach reduces the complexity of the sample and enriches for sequences that are specific to one sample or the other, thus focusing analyses on the differences in the genetic complement of the two environments. Although effective, the GFE method only recovers relatively small fragments. Meyer et al. (2007) implemented a different approach that allows the recovery of full-length ORFs. This is accomplished using driver DNA (partial target gene fragments amplified from community DNA), which is used as a hybridization probe for the recovery of DNA fragments covering the full-length ORF. Such technologies, if applied in combination with large-scale sequence analysis, can target specific functions (e.g. denitrification, sulfidogenesis, methanogenesis, etc.) that can more directly link microbial guilds and their component populations to ecosystem processes.

Possible alternatives to these approaches include the exploitation of DNA- or RNA-binding proteins for selecting pools of genes sharing transcriptional control mechanisms. Using nucleic acid-binding proteins, molecular biologists and geneticists have, with good success, identified and characterized DNA–protein or RNA–protein interactions. Microbial ecologists might benefit from modifying these protocols to enrich for certain targets from total community nucleic acid pools. Chromatin immunoprecipitation (ChIP) is traditionally used to determine the location of DNA-binding sites on a genome by probing it with a single protein (Nelson et al., 2006). Typically, DNA–protein complexes are precipitated and amplified by PCR, but given the new sequencing capabilities, researchers can directly sequence the DNA recovered. In this way, transcription factors (e.g. $\sigma$ factors, or positive and negative regulators) could be used to enrich for the corresponding regulated DNA. Primary $\sigma$ factors can be used to enrich for DNA associated with core functions, while alternative $\sigma$ factors could be used to enrich for at least some DNA associated with accessory functions. Because response to environmental cues can involve positive or negative regulation via specific alterations in the composition of the cellular pool of RNA polymerases, it should be possible to utilize the cell’s own regulatory mechanisms to select for specific genomic fragments.

RNA transcripts can also be targeted through RNA immunoprecipitation (Gilbert & Svejstrup, 2006), a modified version of the ChIP technique. RNA–protein interactions can be used to select for subsets of the RNA transcript pool before reverse transcription and sequencing. The Rho factor is the most common RNA-binding protein identified in bacteriophages, Bacillus, Klebsiella, Escherichia, Salmonella and other enteric bacteria (Ciampi, 2006), as well as in more recently uncovered lineages such as the Planctomyces (Glöckner et al., 2003). This protein is required for Rho-dependent transcription termination in both bacteria and phages. It is a homohexameric protein that binds to mRNA at the Rho-binding site, known as the Rho utilization site (rut). Rho sites have been targeted previously using ChIP and microarrays to study the role of Rho termination in controlling the production of small RNAs and tRNAs, as well as novel antisense transcripts in Escherichia coli (Peters et al., 2009). As with other transcription factors, universally conserved binding sites for Rho are not likely, but the conservation of, for example, a GC-rich rut region might be suitable for the enrichment of transcripts, reducing the complexity of the mRNA pool before sequencing.
A more direct approach to detect single organisms and determine their metabolic capabilities is the isolation and sequencing of single cells (Marcy et al., 2007). A microfluidic device can be used to isolate individual microbial cells and subsequently amplify their genome in the absence of cultivating. Individual TM7 cells derived from the human mouth have been thus isolated and > 1000 genes have been amplified and sequenced from single cell templates. One surprising finding from the above-mentioned analysis of this poorly understood phylum was the abundance of novel gene sequences with no close affiliation to sequenced organisms, highlighting the need for deeper databases with adequate representation of organisms from outside the known dominant or well-characterized groups.

Two common features can be identified in all of the approaches and studies described in the preceding paragraphs: (1) they aim to corroborate specific hypotheses through integrated approaches and (2) they aim to reduced complexity before sequencing in order to maximize the possibility of assembling overlapping fragments and even genome reconstruction from mixed community DNA.

Concluding remarks

Soledly on the large body of data they have provided, the nascent ‘omic’ approaches have already made unquestionable contributions to microbial ecology. It is, however, important to develop an understanding of the advantages and limitations of each method and the utility of the data generated for addressing certain questions. Additionally, in order to avoid searching through billions of nucleic acid sequence fragments looking for the proverbial needle in the haystack, it is important to narrow the searches in our studies in order to increase the efficiency of sequence assembly, thereby facilitating the linkage of ecosystem processes to individual microbial populations.

In our opinion, microbial ecologists should focus on developing clear and concise experimental hypotheses and goals that go beyond simple exploratory research to focus on experimental design over technological prowess. Special focus should be given to temporal and manipulative studies that track changes, at the genome and transcriptome level, linked to given contextual environmental variables in an attempt to understand the response of the microbial community to major and minor changes in its environment in order to link genetic information to ecological function in situ. In a word, it is context that is lacking in much of the data already available. Future work should ensure that accompanying environmental metadata are collected and provided so that meaningful connections can be made between the environment and the populations controlling and controlled by it. Our hope is that ‘omic’-enabled research in microbial ecology will be driven by questions and hypotheses that will allow us to be more effective at linking ecosystem processes to bacterial populations and communities, properly casting new technological advances as crucial, powerful means to achieve research goals rather than as ends in themselves.

References


Toward linking bacterial identities and ecosystem processes


Toward linking bacterial identities and ecosystem processes


