Chapter 21

Bioinformation and 'Omic Approaches for Characterization of Environmental Microorganisms

Emily B. Hollister, John P. Brooks and Terry J. Gentry

21.1 INTRODUCTION

In biology, the term 'ome generally refers to the entirety or totality of a collection of specific things. For example, a biome is a collection of living organisms, and a genome refers to the collection of genes within a single organism. 'Omics, then, are fields of study that deal with these collections and involve the characterization and consideration of multiple molecules simultaneously. When botanist Hans Winkler proposed the term "genome" to describe a collection of chromosomes in the 1920s, he probably had no idea how widely the 'ome suffix would come to be used. We commonly study genomes of individual organisms or the metagenomes of communities in order to: (1) understand functional potential; (2) discern phylogenetic relationships; and (3) evaluate heredity (e.g., horizontal gene transfer) at the DNA level. The 'omics concept extends well beyond DNA, however, and can include RNA transcripts, proteins and metabolites, and these are often referred to as the 'omics cascade (Figure 21.1). In this cascade:

- The genome (or metagenome) contains information about *what can happen* (i.e., functional potential);
- The transcriptome (or metatranscriptome) contains information about *what appears to be happening* (i.e., which genes are being expressed);
- The proteome (or metaproteome) contains information about the molecules that *make things happen*; and
- The metabolome contains information about *what has happened recently or is currently happening*.

Although the 'omics cascade captures many of the major 'omics disciplines under study today, a variety of other 'omics have emerged in recent years. Some are subdisciplines of the major 'omics fields mentioned above (e.g., glycomics, lipidomics, interactomics), while others remain emerging concepts, and have yet to be embraced as standalone disciplines in mainstream science.

In addition, the field of bioinformatics has developed to provide the statistical and computational approaches necessary for evaluating the increasingly large and complex datasets that 'omics technologies are producing. In fact, with the rapid expansion in technologies such as DNA sequencing, the analysis and interpretation of 'omics datasets are often the most challenging parts of 'omics -based experiments. In this chapter, we will discuss the primary

I.L. Pepper, C.P. Gerba, T.J. Gentry: Environmental Microbiology, Third edition. DOI: http://dx.doi.org/10.1016/B978-0-12-394626-3.00021-1

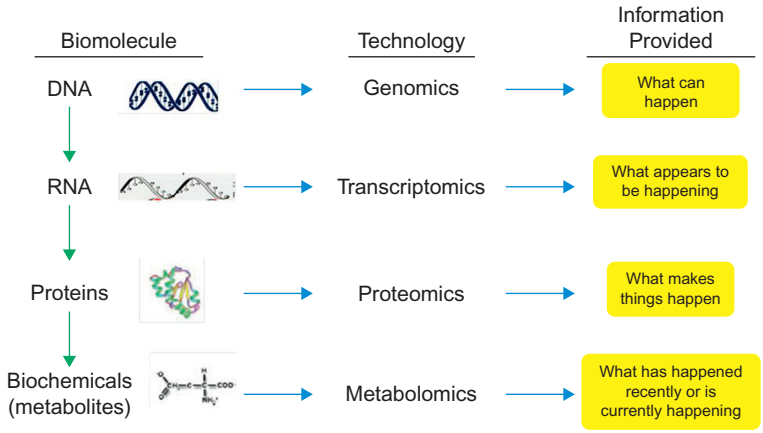


FIGURE 21.1 Overview of 'omics-based approaches for characterizing environmental microorganisms. Adapted from Zhang *et al.* (2010).

'omics-based methods currently being used to characterize environmental microorganisms, and also approaches for analyzing and interpreting the "bioinformation" that these studies generate.

21.2 GENOMICS AND COMPARATIVE GENOMICS

The term genome describes the total collection of an organism's hereditary information. Genomes are often encoded as DNA and stored in chromosomes, mitochondria, plasmids and/or chloroplasts. However, for many viruses, the genome is composed of RNA only. Advances in DNA sequencing technologies have resulted in the ability to produce vast amounts of sequence information. Where sequencing was once limited to specific gene targets or relatively short DNA fragments, it is now routinely applied to whole genomes. The first whole genome sequence of a free-living organism, *Haemophilus influenzae*, was completed in 1995 (Fleischmann *et al.*, 1995). According to the Genomes Online Database (GOLD, see Table 13.2) as of July 2013, nearly 7000 genomes had been sequenced (in complete or draft stage), and thousands more were listed as ongoing projects. The availability of such large quantities of genome sequence information has spawned a field of study known as comparative genomics. Comparative genomics studies seek to identify similarities and differences in the genes and gene content of various organisms, and a variety of data management systems and analysis platforms have evolved aid in these efforts. The Joint Genome Institute (JGI) provides such a platform in their Integrated Microbial Genomes (IMG) system (Markowitz *et al.*, 2010).

By examining the similarities and differences among genomes, comparative genomics attempts to draw inferences with respect to the function of particular genes, identify regulatory regions and find evidence of evolution and/or genetic exchange, by providing insights into the mobility of chromosomal sections and lateral gene transfer. For example, bacterial and archaeal thermophiles often share the same habitats, and there is abundant evidence from genomic analysis that lateral gene transfer is common in the group. Specifically, the *Thermotoga maritima* genome has been estimated to have approximately 20% of genes that have primary homology to hyperthermophilic Archaea, principally *Pyrococcus* spp. (Nelson *et al.*, 1999). When comparative genomic approaches were used to study the thermophilic carboxydotroph, *Carboxydothermus hydrogenoformans*, a variety of interesting features, including conserved genes involved in sporulation and a *Rhodosporillum rubrum*-like carbon monoxide dehydrogenase operon, were discovered (Wu *et al.*, 2005). In addition, it was revealed that approximately 30% of the open reading frames in the genome have high similarity to genes in methanogenic Archaea. This observed sequence similarity has led researchers to hypothesize that extensive lateral genetic exchange has occurred between *C. hydrogenoformans* and methanogens (González and Robb, 2000). The close association of methanogens and carboxydotrophic bacteria in the environment suggests that at the very least there is a high potential for exchange of metabolites between the two groups. These examples illustrate the power of comparative genomics in taking nucleic acid sequences and inferring functionality of individual genes as well as potential interactions and genetic exchanges between members of a particular microbial community.

An emerging area of comparative genomics is single-cell genomics (Laskin, 2012). One of the major benefits of nucleic acid-based methods is the ability to circumvent the need to culture microorganisms before they can be characterized, thus enabling the characterization of difficult-(or impossible)-to-culture microorganisms. However, when applied to environmental samples containing diverse communities of microorganisms, these approaches can usually only provide information for a handful of genes (e.g., 16S rRNA), or at best partially assembled genomes for the most dominant organisms in the samples. However, new techniques such as microfluidics and

microencapsulation are allowing researchers to isolate and grow individual microorganisms (Zengler *et al.*, 2005; Wessel *et al.*, 2013). When combined with whole-genome amplification methods, these approaches are now enabling researchers to obtain sufficient DNA from one initial microbial cell to determine its entire genome, and thus get a better understanding of its potential environmental function—without ever isolating it on traditional laboratory media (Figure 21.2)! This is particularly powerful when used in combination with other methods such as FISH (Section 13.3.5) to target and select for specific groups of microorganisms that may be less abundant and thus would largely be missed with shotgun sequencing-based metagenomics approaches (Podar *et al.*, 2007).

21.3 METAGENOMICS

As discussed in Section 13.6.2, the term metagenomics was first coined by Handelsman *et al.* (1998) in reference to the collective gene content of a community of microorganisms (e.g., those in a soil sample). The definition of metagenomics has since been expanded by the scientific community to generally include any technique that is based upon analysis of DNA extracted from environmental samples. This broader definition of metagenomics would include 16S rRNA sequencing and related phylogenetic fingerprinting techniques; however, it should be noted that some researchers do not consider these methods (e.g., 16S rRNA sequencing) to be true "metagenomic" techniques.

Over the past two decades, metagenomics-based assays have become the standard for characterizing microbial communities, and have been used in countless studies to determine the structure, function and metabolic potential of microbial communities in a wide variety of environments (Table 19.1). The largest application has been 16S rRNA gene sequencing for determining bacterial diversity and community composition, although a variety of other marker genes have been used, and an increasing number of studies are randomly sequencing environmental DNA.

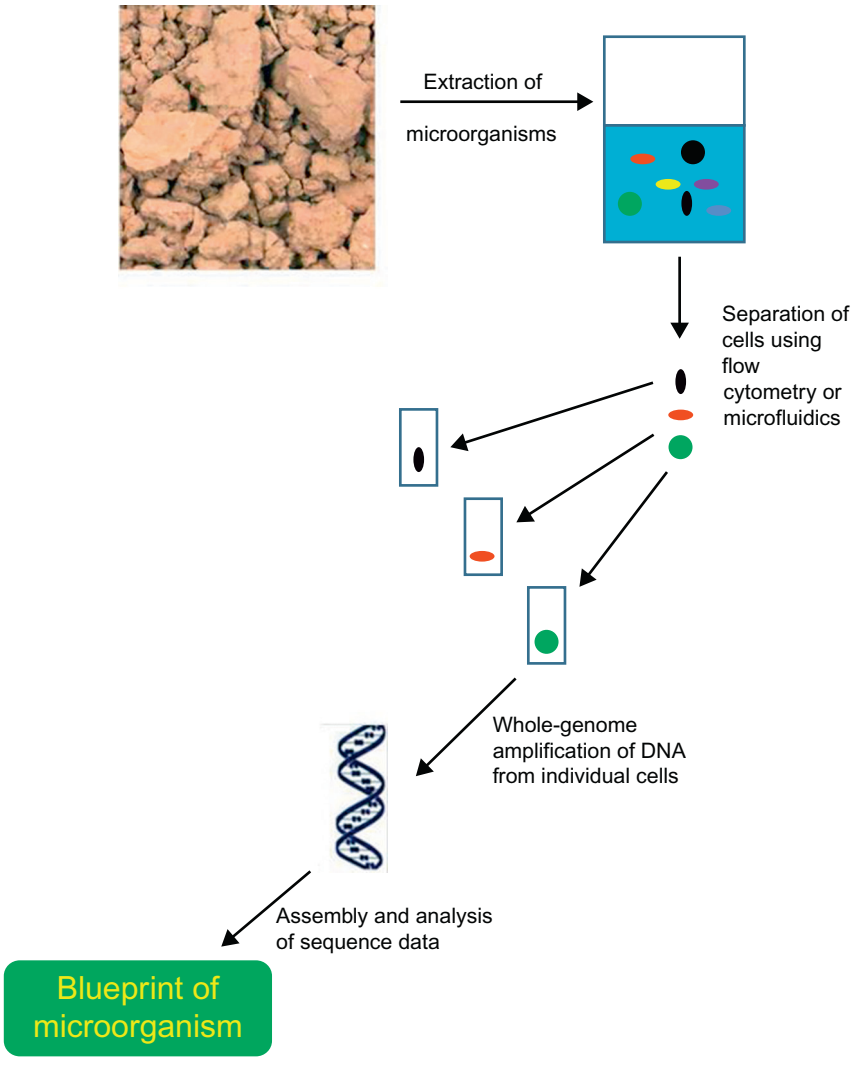


FIGURE 21.2 Overview of a single-cell genomics-based approach for characterizing the genomes of environmental microorganisms.

The ability of metagenomics-based methods to characterize environmental microorganisms without having to first isolate and culture them has allowed the discovery of many previously unknown microorganisms and elucidation of their environmental functions, such as the major contributions of Archaea to ammonia oxidation in a variety of ecosystems (Section 4.4.3).

Although earlier metagenomics studies began with cloning environmental DNA into vectors prior to functional analysis or DNA sequencing, most metagenomics studies today go directly from DNA extraction to sequencing (Figure 13.13). If specific genes are targeted (e.g., 16S rRNA), they can be amplified prior to sequencing (see Section 13.4). Alternatively, the extracted DNA can be sequenced without amplification of any specific genes. This approach is often described as shotgun sequencing. In this process, community DNA is extracted and fractionated into small pieces (if necessary) and sequenced directly via high-throughput sequencing (e.g., 454, Illumina and similar platforms). Following sequencing and processing for quality control (see Section 21.7.1), the reads are either: (1) directly compared to databases for taxonomic and/or functional annotation; or (2) assembled together into longer stretches of DNA which can provide better information since they then represent larger portions of the genome(s) (see Section 21.7.2). Commonly used databases include those available from the National Center for Biotechnological Information (NCBI) and the Metagenomics Analysis Server (MG-RAST) (Section 21.7.2.3; Table 21.1). If higher-order functional identification is required, genes can be categorized using a database such as the Kyoto Encyclopedia of Genes and Genomes (Kanehisa *et al*., 2004); such databases facilitate identification of specific functional and enzymatic pathways. At the moment, the assembly of metagenomics data from environmental samples is extremely challenging due to the complexity of microbial communities in these environments, and the lack of a good set of reference sequences from a diverse microbial community to serve as a scaffold for assembling the sequences (Thomas *et al*., 2012). In general, assembly of metagenomics data is limited to only extremely dominant members of simple communities such as those in acid mine drainage (Case Study 21.1 and Figure 21.3; Tyson *et al*., 2004) or contaminated groundwater (Hemme *et al*., 2010). Another challenge for assembly is the relatively short read-lengths ($<$500 bp) of many currently used sequencing methods. This not only makes assembly more difficult, but it also makes direct annotation of the reads more difficult since they often contain only partial gene sequences. However, the development of newer sequencing technologies, such as that of Pacific Biosciences, promise the ability to provide longer reads ($>$3000 bp) that will encompass entire genes, and possibly even operons, and will thus allow for better taxonomic classification and/or functional prediction. Additionally, the large sequence datasets produced can be computationally challenging to analyze. However, a variety of analysis pipelines and software programs have been developed, and are continually being updated, that facilitate and are standardizing the processing and analysis of these types of datasets (see Section 21.7).

Case Study 21.1 Metagenomics-based Characterization of Dominant Microorganisms in an Acid Mine Drainage Biofilm

One of the first studies to reconstruct putative genomes of environmental microorganisms solely from metagenomic sequence data was the work by Tyson *et al*. (2004) on an acid mine drainage community in California, U.S.A. Although the site was extremely acidic (pH 0.83), an extensive biofilm existed on the surface of water from the mine. Using fluorescence *in situ* hybridization (FISH) and 16S rRNA sequencing, the scientists determined that the biofilm community was relatively simple, and was dominated ($\approx$75% of community) by a single group of related bacteria, *Leptospirillum* group II. The scientists then cloned and sequenced the extracted DNA followed by assembly of the reads. Due to the simplicity of the biofilm community, the sequences were successfully assembled into near-complete genomes for two groups of Bacteria and Archaea: *Leptospirillum* group II and *Ferroplasma* type II (Figure 21.3), as well as partial assembly of three other genomes. Both of the near-complete genomes contained putative genes commonly found in microorganisms living in similar, extreme sites including genes for efflux of heavy metals and various other detoxification mechanisms. A number of novel cytochrome genes, which were potentially involved in iron oxidation, were also detected. Since the site was in the deep subsurface, it received little-to-no inputs of carbon and nitrogen from the surface, and therefore would require at least some of the members of the microbial community to fix both carbon and nitrogen. Genes for carbon fixation were found in the *Leptospirillum* group II genome, but *Ferroplasma* type II appeared to require external sources of carbon. Interestingly, neither *Leptospirillum* group II nor *Ferroplasma* type II contained genes for nitrogen fixation, suggesting that other members of the community most likely fulfilled this vital role for the community. The metagenomic sequencing data from this study provided some initial insights into the metabolism of dominant microorganisms in the biofilm community. In addition, the biofilm is an ideal model community since it is: (1) a relatively simple community dominated by a few microbial populations; and (2) contains a large amount of biomass per unit volume. This enabled a variety of other ′omics methods including transcriptomics and proteomics to be used to validate and expand insights into the ecology of the acid mine drainage biofilm community.

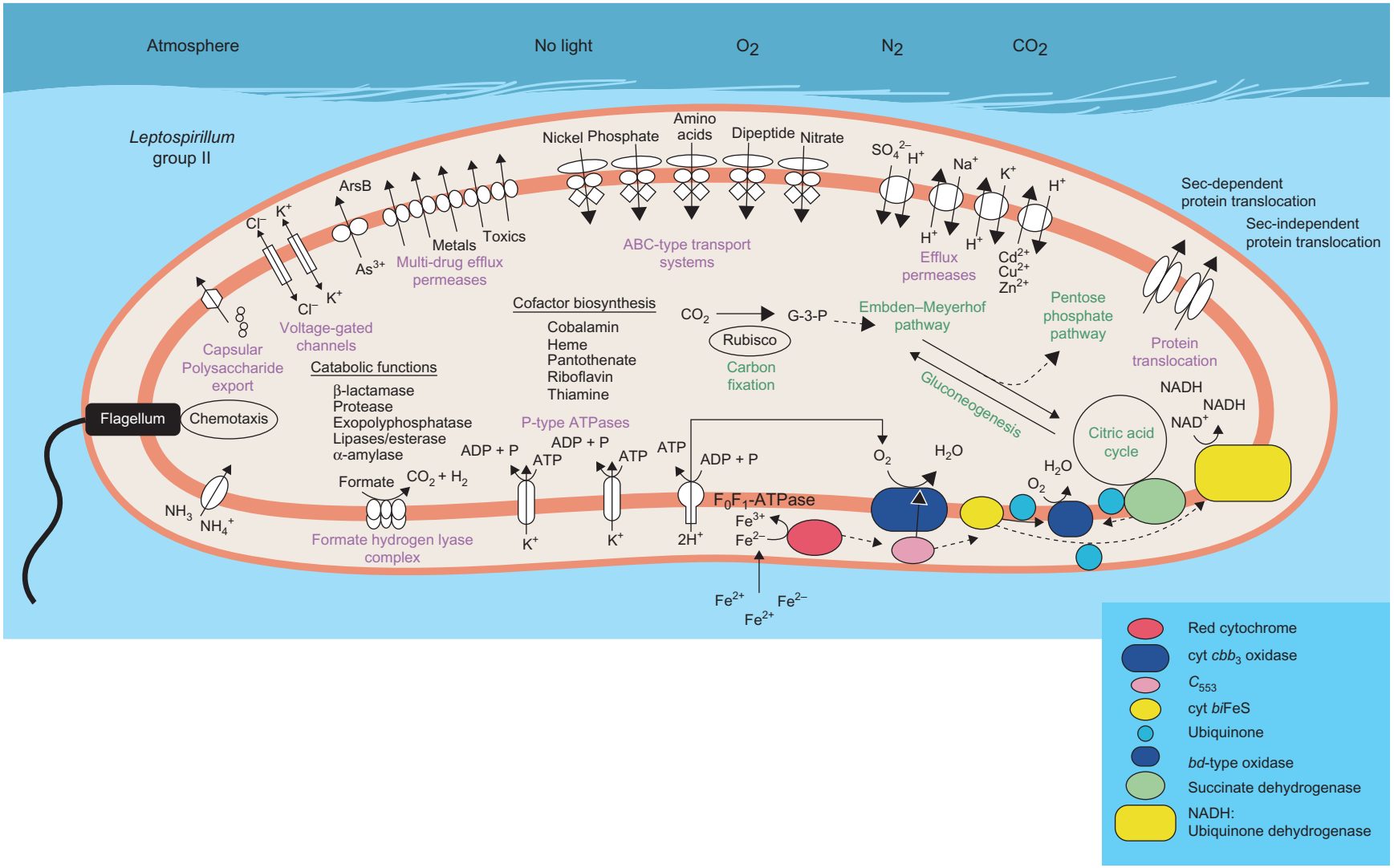


FIGURE 21.3 Metagenomic reconstruction of microbial metabolism in an acid mine drainage community. Constructed from the annotation of 2180 ORFs identified in the assembled *Leptospirillum* group II genome (63% with putative assigned function). The cell diagram is shown within a biofilm that is attached to the surface of an acid mine drainage stream (viewed in cross-section). From Tyson *et al*. (2004).

Despite the unprecedented insight that metagenomics is allowing into the diversity, structure and genetic potential of microbial communities, it should be recognized that the function of genes, from metagenomics data, is inferred bioinformatically rather than tested empirically. However, this initial characterization and prediction of a microbial community's genomic capabilities can serve as the platform for further characterization using other additional 'omics-based assays such as transcriptomics and proteomics, which can verify whether these putative genes are expressed and produce the predicted proteins (Case Study 21.1).

21.4 TRANSCRIPTOMICS

Modern genomic techniques such as metagenomics yield vast amounts of data; however, this data represents the DNA potential of a biological system, not necessarily the expressed phenotype. To unlock the expressed fraction of genomics, one must turn to RNA or protein expression, transcriptomics (a.k.a. metatranscriptomics) and proteomics, respectively. Since RNA, specifically mRNA, represents the product of DNA transcription, it is a logical target for transcriptomics-based analyses. Many metatranscriptomics analyses are less hypothesis driven and may be considered more exploratory in nature. Conversely, some transcriptomics studies focus investigation on expression of targeted genes, and additionally rely on other 'omics to complete the picture (see Case Study 21.2). A number of studies applying transcriptomics to various environmental matrices are available for more in-depth discussion beyond the scope of this section: Carvalhais *et al.* (2012) (review of transcriptomics and soil); de Menezes *et al.* (2012) (transcriptomics and organic contaminant degradation); and Kyle *et al.* (2010) (transcriptomics applied to *E. coli* survival on food).

Overall, transcriptomics analyses have been conducted on a number of sample matrices. Much of the original transcriptomics work was conducted with clinical fecal samples (Gosalbes *et al.*, 2011), which given similar caveats as environmental samples, provided for an applicable template for the analysis of soil, water and plant rhizosphere matrices. Much like sample collection for DNA, care must be taken when collecting mRNA; however, mRNA is notoriously labile. mRNA will typically persist in an environmental sample for no more than a few minutes following collection. Additionally, the mRNA half-life may vary for different environments and microorganisms, and by gene function, with housekeeping genes yielding more stable mRNA products (Selinger *et al.*, 2003). For this reason, samples must be preserved within minutes, if not seconds, of collection. There are a number of collection protocols, including commercial kits (easily standardized) and "homemade" traditional approaches, which often yield larger quantities and higher quality RNA, though standardization may be more difficult if conducting latitudinal studies.

Often, sample collection involves immediate freezing in liquid nitrogen in order to prevent enzymatic RNA degradation. While this may be possible when working in a laboratory or greenhouse environment, it may not be

Case Study 21.2 Combining 'Omics: Metatranscriptomics and Metabolomics

Combining 'omic analyses yields more useful data than a single analysis in many cases. For example, the application of transcriptomics- and metabolomics-based analyses can reveal the relationships between genes and their final functional activity. At the most basic level, one analysis may provide useful insight while the other may not; a more complex analysis may reveal intricate relationships between transcriptional control and metabolic function. A study by Ishii *et al.* (2007) aimed to marry the two analyses in the study of common environmental (substrate abundance and reduction) and genetic (missing enzymatic pathways) pressures imposed on *Escherichia coli* K-12. Global responses were measured using a combination of qRT-PCR (quantitative real-time PCR) to measure targeted mRNA transcripts, and liquid chromatography and time-of-flight mass spectrometry to measure metabolome response. Additionally, DNA microarrays and 2D-differential gel electrophoresis were used to measure relative gene and protein expression, respectively. From these data, the scientists generated an expression index, which took data, separately, from each analysis type and scaled the responses to permit comparisons across all analyses. The analyses revealed gradual increases in mRNA and protein levels using both targeted and global analyses when placing *E. coli* under high growth rate conditions. Interestingly, metabolites did not significantly increase. Reducing substrate availability additionally demonstrated few changes in metabolites compared to the control. Finally, the authors disrupted the enzymatic network by disrupting individual genes; but only subtle changes were noted in mRNA and protein expression of central carbon enzymatic pathways. The study demonstrated two approaches which allow *E. coli* to quickly react to genetic and environmental changes. The results of the study suggest that *E. coli* has built-in structural redundancy (in enzymatic pathways), which absorbs sudden changes in available substrate as well as loss of single gene function. Results also suggest that *E. coli* maintained the same metabolic rate (as demonstrated by metabolomics) while up-regulating enzyme expression (as demonstrated by targeted and global transcriptomics). This study demonstrates the stability that *E. coli*'s enzymatic pathways provide along with the ability to rapidly respond to environmental pressures. Discovery of this information was only made possible through use of the multiple 'omics approach, in which one assay demonstrated changes in the system, while the other assay was incapable of detecting responses.

feasible for environmental work. These situations may necessitate the use of RNA stabilizing buffers such as the MO BIO LifeGuard™ Soil Preservation Solution (MO BIO Laboratories, Inc., Carlsbad, CA). These buffers facilitate the collection of mRNA from environmental samples with immediate preservation while in the greenhouse or field. While this step preserves the total RNA in a biological system, the extraction of RNA from intact cells is still necessary prior to analysis. As with DNA extraction procedures, most RNA extraction from commercial kits involves bead-beating technology and the capture of RNA in a stable buffer which can be frozen and subsequently analyzed. However, copurifying soil and fecal humic acids and contaminating organic molecules and metals can affect the quality of the final RNA products (see Chapter 8).

Once mRNA is safely collected and preserved, it needs to be converted to cDNA (complementary DNA; Section 13.4.5). However, mRNA is often present as a small fraction of the total RNA (mostly rRNA and tRNA). Therefore, mRNA is often enriched or selectively isolated from total RNA. As with sample collection and RNA extraction, there are a number of commercial approaches available, including the use of exonuclease treatment (targeting rRNA), and subtractive hybridization using magnetic beads coupled with oligos specific for rRNA and tRNA, which are subsequently removed from the solution. However, in environmental and clinical samples, eukaryotic mRNA may be present at high levels; in these cases, eukaryotic mRNA can be removed by targeting mRNA containing 3′ poly-A tails (Bailly *et al*., 2007). Following mRNA enrichment, cDNA is most often the template of choice for most downstream applications. In these cases, reverse transcriptase and either specific primers or random oligos are applied, as in most other methods requiring cDNA synthesis (see Section 13.4.5).

As with DNA metagenomics work, the choice of the sequencing system depends on the length of the intended sequence product and anticipated coverage needed for a specific biological system. Currently, most metatranscriptomics work is conducted using 454 or Illumina systems, the former producing larger sequence products ($\approx$ 500 bp), while the latter provides for smaller sequences ($\approx$ 150 bp), but a larger number of products ($<$1 Gb vs. 600 Gb). Each system satisfies different study objectives as longer reads are used to map repetitive sequence regions, while some studies require deeper coverage depth. As sequencing methods continue to develop, other platforms will likely be adopted for use in metatranscriptomics.

Following sequencing, bioinformatic analysis removes poor quality and short read sequences. Sequence ends are also trimmed and data analyzed for the presence of rRNA sequences (which can still be present, despite mRNA enrichment), which are promptly removed from the library. Typically, sequences are compared to available databases which assign gene function and identification. However, most metatranscriptomic projects include comparisons of gene relative frequency, and whether a gene is up- or down-regulated. In this case, gene frequencies are normalized to gene abundances from a control metagenome, preferably from the same environmental matrix. Similarly, control metatranscriptomes allow for comparison to treated samples or to various time points, depending on the study objectives. As with metagenomic work, assembly may also be necessary, though the complexity of environmental samples may prohibit this. Various assemblers are available and consist of programs commonly employed in metagenomic work such as Genovo (Laserson *et al*., 2011) and Newbler (454 Life Sciences, Branford, CT, U.S.A.). A transcriptomic specific assembler such as Velvet (Zerbino and Birney, 2008) can also be used (see Section 21.7 for additional details on bioinformatics).

21.5 PROTEOMICS

Although DNA- and RNA-based methods can provide tremendous insights into the environmental roles of microorganisms, proteins, not genes, are directly responsible for the majority of microbial processes. Therefore, measurement of these proteins (i.e., enzymes) can provide a more direct measurement of microbial activity. The proteins produced by a given microorganism under a given set of conditions are collectively referred to as the proteome. In contrast to the genome, the proteome is much more variable (like the transcriptome) with different proteins being produced depending upon the stage of cell metabolism and the environmental stimuli present.

Studying the proteome has the potential to provide unique information about cell function, and the mechanisms behind cell responses to different stimuli. Specifically, proteomics-based approaches allow identification of proteins that are differentially expressed and, thus, likely to be important in the microbial response to environmental conditions. Proteomics-based studies of environmental effects on microorganisms typically involve the following:

- Exposure of microorganisms to a condition of interest
- Isolation of proteins from each population
- Separation of proteins
- Protein identification

The first two steps in proteomics-based studies are relatively easy. There are a host of effective methods available to isolate and purify the heterogeneous protein mixtures made by microorganisms. However, separating the proteins contained within these complex mixtures represents one of the most challenging aspects of proteomics. Two strategies to separate proteins are commonly used: two-dimensional polyacrylamide gel

electrophoresis (2D-PAGE) and liquid chromatography-mass spectrometry (LC-MS). In 2D-PAGE, proteins are first separated according to their isoelectric points (pI), the pH at which the protein has no net charge. The second dimension of 2D-PAGE separates proteins based on their masses using a polyacrylamide gel. The resulting gel contains many spots, each ideally containing a single protein that can be identified using mass spectrometry-based methods described below.

Alternatively, LC can be used for protein separation. In this approach, proteins from a given population are pooled and digested enzymatically into their constituent peptides. These peptides are separated by LC (see Section 11.2.1.1) which allows for the separation of molecules based on charge or hydrophobicity. Proteins in the original population of cells are identified on the basis of these peptides as described below. LC-based separation of proteins can be more readily automated, and may be more reproducible than 2D-PAGE.

Once separated, proteins must be identified to gain insight into mechanisms by which microorganisms interact with the environment. Mass spectrometry is currently the tool of choice for this task. Intact proteins are broken down enzymatically (i.e., digested) into smaller peptides and analyzed by mass spectrometry. Once accurate masses of the peptides are obtained, the protein from which the peptides originated can be identified. This approach to protein identification is known as peptide mass fingerprinting (PMF). When PMF fails, other types of mass spectrometry can be used to obtain direct amino acid sequence data that can be useful for protein identification. As differentially expressed proteins are identified, the investigator gains insight into mechanisms by which the microorganism responds to a particular environmental condition (Westermeier and Naven, 2002).

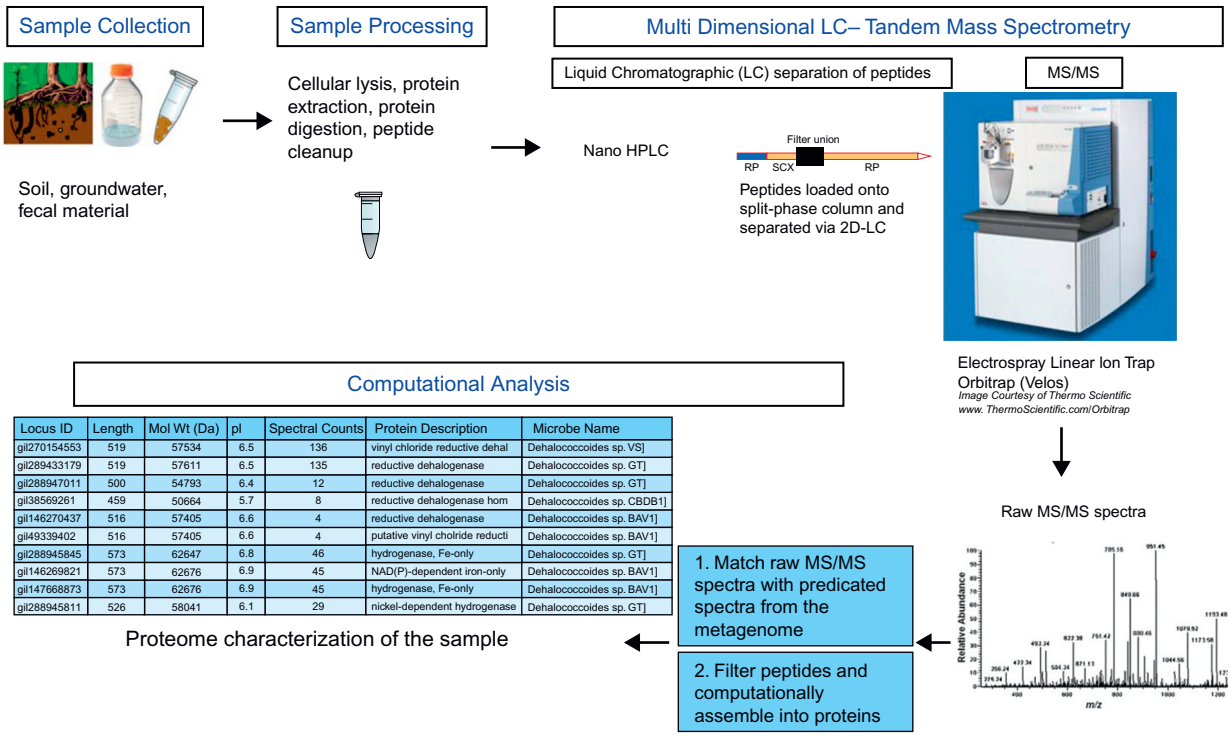
Studies have demonstrated that the comprehensive, high-throughput nature of proteomics-based approaches is also well suited to elucidating biodegradative pathways. For example, Kim *et al.* (2004) examined biodegradation pathways of an aromatic-degrading pseudomonad (*Pseudomonas* sp. K82) using 2D-PAGE followed by mass spectrometric identification of proteins. Using this approach, the investigators discovered three metabolic pathways, each of which was induced to a different degree by three different aromatic compounds.

As with recent research in metagenomics, applications of proteomics to microbial ecosystems are emerging and offer promise to link microbial species within complex communities to function (Hettich *et al.*, 2013). Termed metaproteomics or community proteomics, these approaches are designed to isolate as many proteins as possible from a microbial community to learn more about which microorganisms perform what tasks within a community (Figure 21.4). For example, Ram *et al.* (2005) used metaproteomics to investigate and characterize an acid mine drainage biofilm community similar to the one described in Case Study 21.1 and Figure 21.3. As with most proteomics-based approaches, this approach was facilitated by genomic sequence data (Figure 21.5). Specifically, the authors constructed a database of 12,148 predicted protein sequences from the similar biofilm community previously characterized using metagenomics (Tyson *et al.*, 2004). Using this database and an LC-mass spectrometry approach to protein identification, the authors identified 2033 individual proteins. Most were produced by members of the genus *Leptospirillum* and were involved with adaptation to this extremely acidic ($pH \approx 0.8$), metal-laden environment. Many proteins could not be assigned a function, yet were highly prevalent. One of these, which was previously identified by the metagenomics approach as possibly playing a role in iron oxidation, was confirmed to be a novel cytochrome involved in iron oxidation and acid mine drainage formation. A subsequent study found that the proteome changed during development of the biofilm (Mueller *et al.*, 2011). For example, the dominant organism, *Leptospirillum* group II, produced more enzymes for metabolism of 1- and 2-carbon compounds and protein synthesis during early biofilm development, and more stress-related and iron oxidation proteins, likely related to acid mine drainage formation, as the biofilm developed and resources likely became more limiting (Figure 21.6).

Despite the promise of metaproteomics, many impediments to its broader use exist. The need for a universal method to exhaustively extract proteins from complex communities, particularly those indigenous to soil, is of paramount importance. In addition, the sensitivity of detection of existing methods is limited, and approaches are only capable of identifying proteins from microbial populations that comprise $>1\%$ of a community. Furthermore, additional metagenomics data are needed in order to better predict the suite of proteins produced by environmental microbial communities and accurately interpret metaproteomics data (Figure 21.5). Nevertheless, metaproteomics is a developing and promising area of research, and will likely be increasingly used over the next decade to study the activity and functions of environmental microorganisms.

21.6 METABOLOMICS

Metabolomics consists of the study of low molecular weight metabolites. Environmental metabolomics consists of metabolites produced by interactions between microorganisms, small eukaryotes, plants, animals, predators and the presence of abiotic pressures and stimulants.



Locus ID	Length	Mol Wt (Da)	pI	Spectral Counts	Protein Description	Microbe Name
gil270154553	519	57534	6.5	136	vinyl chloride reductive dehal	Dehalococcoides sp. VS]
gil289433179	519	57611	6.5	135	reductive dehalogenase	Dehalococcoides sp. GT]
gil288947011	500	54793	6.4	12	reductive dehalogenase	Dehalococcoides sp. GT]
gil38569261	459	50664	5.7	8	reductive dehalogenase hom	Dehalococcoides sp. CBDB1]
gil146270437	516	57405	6.6	4	reductive dehalogenase	Dehalococcoides sp. BAV1]
gil49339402	516	57405	6.6	4	putative vinyl cholride reducti	Dehalococcoides sp. BAV1]
gil288945845	573	62647	6.8	46	hydrogenase, Fe-only	Dehalococcoides sp. GT]
gil146269821	573	62676	6.9	45	NAD(P)-dependent iron-only	Dehalococcoides sp. BAV1]
gil147668873	573	62676	6.9	45	hydrogenase, Fe-only	Dehalococcoides sp. BAV1]
gil288945811	526	58041	6.1	29	nickel-dependent hydrogenase	Dehalococcoides sp. GT]

FIGURE 21.4 Experimental flowchart for sample preparation and measurement in a metaproteomics experiment. Sample collection and processing steps must be optimized to match the nature of the material to be analyzed, in terms of biomass amount and complexity, matrix composition, sample heterogeneity, etc. The resulting proteome sample is digested with trypsin and loaded onto a biphasic HPLC column for concomitant 2D-separation and MS analysis via nanoelectrospray-based ionization of eluting peptides. Acquisition of parent peptide ion (MS1) mass and fragmentation (MS/MS or MS2) information provides an experimental dataset containing hundreds of thousands of spectra that can be computationally matched to the predicted proteome obtained from metagenomics information. From Hettich *et al.* (2013).

Common metabolites (≤1500 Da) consist of organic acids (e.g., glycolytic intermediates), amino acids (e.g., protein intermediates) and various saccharides (e.g., monosaccharides and cleaved sugars).

As with genomics, transcriptomics and proteomics studies, the goal of metabolomics is often to elucidate the function of a microorganism or microbial community; however, proteomics and metabolomics reveal information related to the "final" genome product. Similarly, metabolomics characterizes the interactions between microbial constituents and their environment, or between microbial and other higher-order ecological organisms such as plants and animals. Metabolomics has been used as an exploratory tool (Dunn, 2008), to uncover the functional status of microbial populations and single cells in their environment, revealing community and ecological structure. Targeted metabolomics enables the user to focus upon a specific metabolite, for instance when a treatment may dictate the up- or down-regulation of a product, while global metabolomics views the biological system and its metabolites as a whole. A number of studies or reviews describing metabolomics and various environmental matrices are listed for further information beyond the scope of this section: Zhang *et al.* (2010) (review); Ito *et al.* (2013) (contaminated feedstock); Liebeke *et al.* (2009) (benchtop single culture study); and Bundy *et al.* (2009) (review).

Metabolites are broken down into two groups: the endometabolome and exometabolome, which are metabolites contained intracellularly and extracellularly, respectively. Like transcriptomics, the study of intracellular metabolites can be more difficult, as these molecules are more fleeting and in a constant state of flux. Metabolome complexity and study objectives involving intra- or extracellular metabolites determine the type of extraction and processing. Once metabolites are extracted, they are subjected to identification with a number of instruments such as gas and liquid chromatography-mass spectrometry, Raman spectroscopy and nuclear magnetic resonance (NMR). In many instances, depending on the complexity of the biological system, the study will call for a combination of two or more of these instruments (Dunn, 2008; Case Study 21.2).

Regardless of platform, a large amount of metabolic data is typically generated. In many instances, the metabolites under investigation are unknown and global in perspective; therefore, query databases are required to deduce the function and purpose of the metabolite. Metabolites are often identified as products or intermediates of environmental

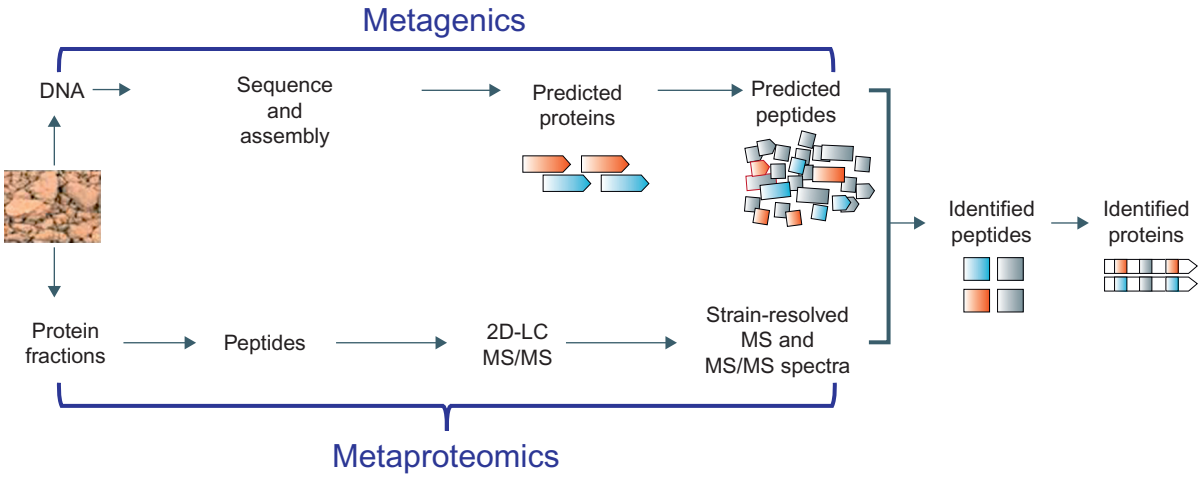


FIGURE 21.5 Integrated use of metagenomics and metaproteomics for characterizing microbial communities. DNA is extracted from biological samples, fragmented and sequenced. The resulting sequence reads are then assembled and/or binned. After gene annotation, the protein-sequence database is constructed and an *in silico* trypsin digest is performed on the predicted proteins, resulting in a peptide database (top). From the same or similar biological samples, total community protein is extracted and then digested using trypsin. Peptide separation by two-dimensional (2D) nano-liquid chromatography (LC) and tandem mass spectrometry (MS/MS) is performed (see Figure 21.4). The spectra are matched to peptides in the database, and after filtering, a list of identified peptides is obtained. Based on their unique occurrence in one protein in the whole database, certain peptides (unique peptides, colored red and blue) can be tracked back to their corresponding proteins and thus permit reliable protein identification. Nonunique peptides (gray) cannot be used to uniquely identify a protein, but these data are used in the calculation of protein coverage and abundance measures. The identified proteins are placed back into the genomic context of the organisms they are derived from to allow for the biological mining of the data. Adapted from VerBerkmoes *et al*. (2009).

populations under stress due to the overall health of a system. Given the relatively novel nature of metabolomics, particularly in environmental sciences, very few databases exist to facilitate identification of environmental metabolites. Common databases consist of the Human Metabolome Database and Kyoto Encyclopedia of Genes and Genomes; commonly used databases can be found at http://www.metabolomicssociety.org/databases.

21.7 BIOINFORMATION

21.7.1 Bioinformatics and Analysis of Marker Gene Data

21.7.1.1 16S rRNA and Other Marker Genes

As discussed in Chapter 13, marker genes, such as ribosomal RNA (rRNA) genes or the internal transcribed spacer (ITS), are frequently used to characterize the composition of bacterial, archaeal and fungal communities. Marker genes are useful because they allow for the relatively rapid characterization of the composition and diversity of microbial communities. The 16S rRNA gene is the most commonly used marker gene for the characterization of Bacteria and Archaea, while the ITS tends to be favored among microbiologists for the characterization of fungi. That notwithstanding, the 18S rRNA and 28S rRNA genes are also commonly used for the characterization of fungal communities, and are frequently employed as an alternative to the ITS region when detailed phylogenetic information is needed.

Recall, good marker genes share the characteristics of:

- Ubiquity—the marker should be present in most, if not all, target species
- Genetic conservation—the sequence of the marker should be conserved sufficiently that it can be targeted with PCR primers
- Variability—in combination with genetic conservation, the marker should also contain regions of sequence that are variable and allow for differentiation between species, among lineages and within populations.

Given these characteristics, marker genes are well suited to serve as targets for sequence-based community surveys. Using high-throughput sequencing platforms, such as 454, Ion Torrent or Illumina, researchers are now able to generate large quantities of sequence information allowing them to describe the structure and diversity of microbial communities of interest.

21.7.1.2 Platforms for Sequence Analysis

Due to the generation of large quantities of marker gene sequences, there is a subsequent need to analyze and

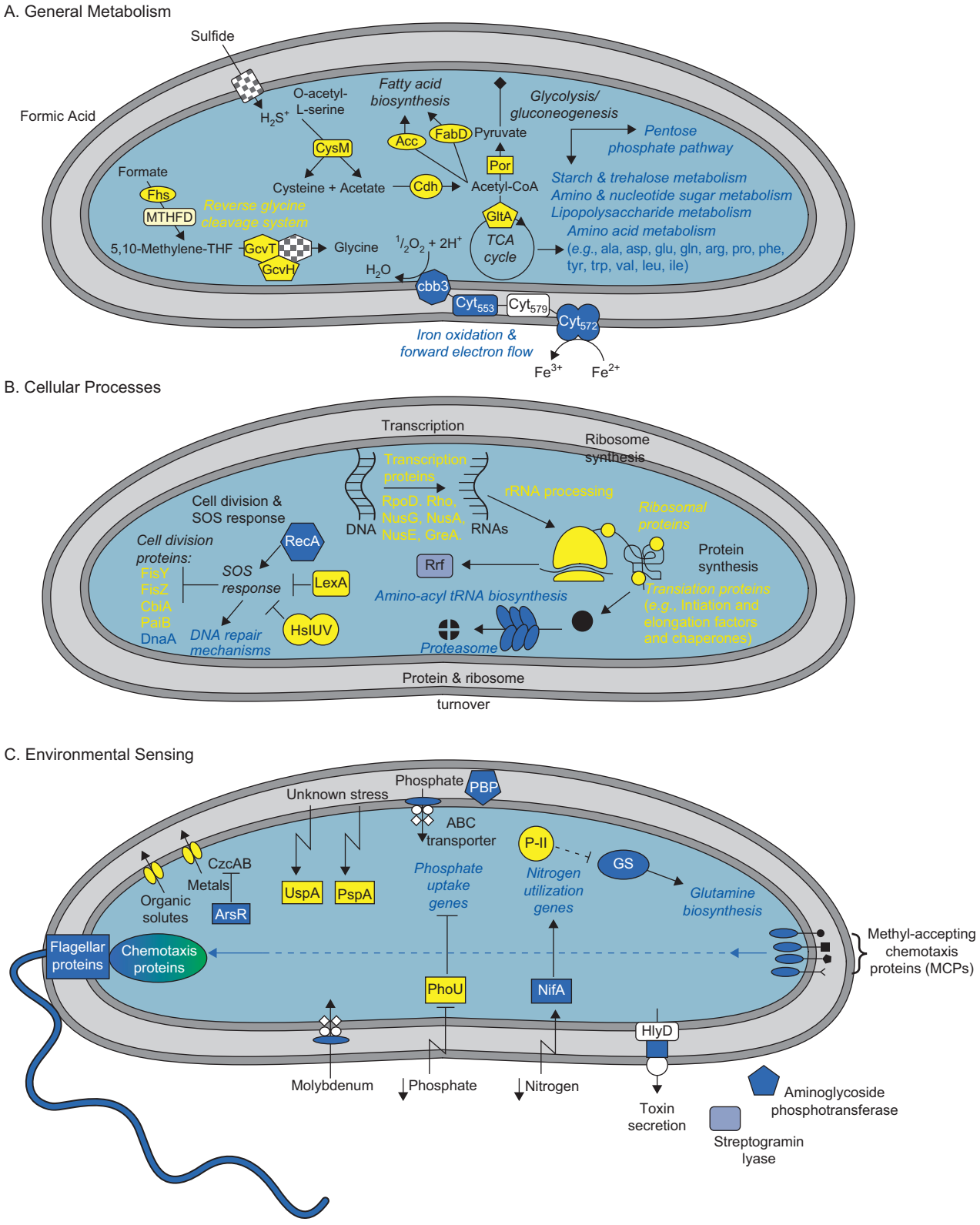


FIGURE 21.6 Physiological changes of the dominant bacteria, *Leptospirillum* group II, in an acid mine drainage biofilm as the biofilm matures. Figure depicts significant changes in *Leptospirillum* group II proteins involved in (A) general metabolism, (B) cellular processes and (C) environmental sensing. Proteins with yellow fill and pathway headings in yellow font (e.g., "Fhs" and "Reverse glycine cleavage system") were significantly more abundant in early and intermediate growth stages, and proteins with blue fill and pathway headings in blue font (e.g., "Cyt_{572}" and "Pentose phosphate pathway") were significantly more abundant in late growth stage samples. Proteins labeled in white were detected by proteomics, but did not demonstrate a biologically relevant abundance pattern. Proteins filled with a gray-checked pattern were not detected or are unknown. From Mueller *et al.* (2011).

interpret them. A variety of analysis platforms have been developed to accommodate this need, many of which are open-source and/or freeware packages (Table 21.1). Examples of these include standalone tool sets like MOTHUR (Schloss *et al.*, 2009) and QIIME (Caporaso *et al.*, 2010). Others are web-based portals like the Ribosomal Database Project Pyrosequencing Pipeline (Cole *et al.*, 2009), VAMPS (http://vamps.mbl.edu/), the Genboree Microbiome Toolset (Riehle *et al.*, 2012) and PlutoF (Abarenkov *et al.*, 2010). Many of the web-based portals feature the functions of MOTHUR and QIIME, some utilize custom algorithms, and most feature additional platform-specific analysis modules. One of the biggest advantages of web-based platforms is that they link the features of popular analysis packages with the power of larger, institutional servers. Their main disadvantage, however, is that by being shared resources, they can be subject to high demand, and one may sometimes have to wait longer than anticipated for results to be processed.

Marker gene analysis platforms tend to revolve around a core set of functions. These include: (1) the conversion of raw sequence data (i.e., sff or fastq files) into FASTA format; (2) quality filtering of sequences; (3) separation of pooled sequences into their originating samples on the basis of barcode tags; (4) data "reduction" to allow for increased computational efficiency; and (5) detection of potentially chimeric reads. Beyond these features, many platforms also offer algorithms that: (1) attempt to

TABLE 21.1 Common Platforms for Sequence Analysis and Their Capabilities

Platform/Package	Website	Features	Data Types Analyzed	Reference
Marker gene analysis				
QIIME	http://qiime.org	Quality filtering, separation of sequence by barcode, OTUs, taxonomic identities, diversity analyses, between community comparisons	Marker gene sequences; developed for 16S but can be used with 18S or ITS sequence	Caporaso *et al.*, 2010
MOTHUR	http://www.mothur.org	Quality filtering, separation of sequence by barcode, OTUs, taxonomic identities, diversity analyses, between community comparisons	Marker gene sequences; developed for 16S but can be used with ITS or other marker gene sequences	Schloss *et al.*, 2009
Ribosomal Database Project	http://rdp.cme.msu.edu	Archive submission portal; quality filtering; taxonomic identities; calculation of some diversity indices	Largely developed to support 16S analysis; includes 28S database for fungi	Cole *et al.*, 2009
VAMPS	http://vamps.mbl.edu	Wraps features of QIIME and MOTHUR; includes links to data from large projects like the Human Microbiome Project and the Microbiome of the Built Environment	16S rRNA gene sequences	Huse *et al.*, 2010
Genboree Microbiome Toolset	http://genboree.org	Web-based platform for QIIME; offers additional custom analysis modules	16S rRNA gene sequences	Riehle *et al.*, 2012
PlutoF	http://unite.ut.ee/workbench.php	Quality filtering, separation of sequence by barcode, OTUs, taxonomic identities	ITS sequences	Abarenkov *et al.*, 2010
(Meta)genome analysis				
IMG and IMG/M	http://img.jgi.doe.gov	Quality filtering; genome and metagenome assembly and annotation; comparative analysis of genomes or metagenomes	Shotgun genomes and metagenomes	Markowitz *et al.*, 2010
MG-RAST	http://metagenomics.anl.gov	Quality filtering; taxonomic and functional annotation; no assembly provided	Shotgun metagenomes, marker gene surveys	Meyer *et al.*, 2008
CAMERA	http://camera.calit2.net	Quality filtering; metagenome assembly and annotation; viral diversity analyses	Shotgun metagenomes, marker gene surveys for Bacteria, Archaea and viruses	Sun *et al.*, 2011
EBI Metagenomics	https://www.ebi.ac.uk/metagenomics	Sequence archiving; quality filtering; taxonomic analysis of 16S reads; functional annotation	Shotgun genomes, metagenomes, marker gene surveys	Hunter *et al.*, 2011

minimize errors as a result of sequencing "noise"; (2) cluster sequences into operational taxonomic units (OTUs) on the basis of similarity; (3) assign identities to each sequence through comparison to reference databases; and (4) perform additional analyses including the calculation of diversity indices, evaluation of sample-to-sample similarities and differences, and detection of features that distinguish one community from another.

21.7.1.3 Quality Criteria

The sequencing process is inherently prone to error (i.e., the incorporation of incorrect base calls during sequencing). Such errors include substitutions made by DNA polymerases, chimeric sequence formation and the difficulties entailed in reliably reproducing homopolymeric regions of sequence (Schloss *et al.*, 2011). Although these error rates vary among sequencing platforms and tend to be relatively low, their cumulative effects on marker gene survey data can alter our perception of microbial community diversity. As a result, it is common to employ a series of quality filters to the sequence data prior to analysis. These include:

- **The removal of low quality sequences**
  As each base is incorporated during a sequencing reaction, a score indicating the quality of each base call is also generated and recorded into the sequencing record. The greater the number of errors in a stretch of sequence, the lower the quality score tends to be. Sequences can be trimmed according to quality scores, and this can be done in one of two ways. The first involves trimming away low-scoring regions of sequence from each read and retaining what remains. The second removes entire sequences from a data set on the basis of average read quality. Typically, sequences with an average quality score lower than 20 are removed.

  From time to time, a base position cannot be called with certainty. These are known as ambiguous base calls, and they are indicated in a stretch of sequence by the letter N (e.g., ATCCN). Sequences containing ambiguous base calls are indicators of poor sequence quality (Huse *et al.*, 2007), and are typically removed from analysis.
- **The removal of sequences that are too long or too short**
  Sequences that are very short or very long, relative to the expected sequence length for a given sequencing platform, tend to be of lower quality and contain large numbers of errors (Huse *et al.*, 2007). As a result, users typically filter out these sequences. For example, it is common to remove sequences that are shorter than 200 bp or longer than 1000 bp from sequence runs generated on the Roche 454 platform, which average 450 bp in length.
- **The removal of sequences containing exceptionally long homopolymers**
  Homopolymeric runs are regions of sequence in which the same base call is incorporated multiple times in a row. The sequence ACGGGGGGGTC, for example, contains a homopolymer of seven guanine residues. Although homopolymers do exist in nature, they can occur erroneously during the sequencing process (Huse *et al.*, 2007). Some sequencing platforms (the Roche 454 platform, in particular) have difficulty reproducing homopolymeric sequences correctly. As a precaution against spurious homopolymers, most analysis platforms allow users to define an acceptable homopolymer length (e.g., a homopolymer limit of 6 is commonly utilized), and filter out sequences containing longer homopolymeric spans.
- **Barcode and primer trimming and the removal of sequences containing mismatches to their barcode or primer sequences**
  High-throughput sequencing platforms offer the ability to multiplex samples for sequencing. Multiplexing allows pools of DNA amplicons originating from multiple samples to be mixed together and sequenced simultaneously. The incorporation of barcodes into the amplicon sequences permits them to be sorted bioinformatically and attributed back to their sample of origin. Barcodes, also known as tags, are typically short (i.e., 8–12 bp in length) sequences that can be ligated onto PCR products after they are produced or incorporated into the sequencing primer.

  Although barcodes provide a means for assigning reads to their sample of origin, they also represent an additional opportunity for quality control. Typically, sequences that contain errors (i.e., incorrect base calls) in their barcode sequence are considered to be of low quality and are removed from analysis, although some protocols will accept one or two mismatches (Caporaso *et al.*, 2010; Schloss *et al.*, 2011). This is also true of primer sequences. Once sequences have been evaluated for barcode and primer mismatches and pooled by sample of origin, the barcode and primer sequences are trimmed away.

21.7.1.4 Removal of Chimeras

In Greek mythology, the chimera was described as a monster that was part lion, part goat and part snake. During the PCR process, it is possible for DNA polymerase to begin copying one target, become disrupted and finish its amplification cycle by picking up copying a second target. The resulting product is a hybrid of the two original templates and is commonly referred to as a chimera, or chimeric sequence (Figure 21.7). It is estimated that chimeric reads may account for 5% or more of sequence libraries (Ashelford *et al.*, 2005), and the risk for chimera production is potentially problematic

when one is trying to characterize the composition and diversity of a mixed microbial community.

The detection of chimeras typically involves the comparison of each individual read to all others within a sequence library or a reference database. Those that appear to have strong similarities to two different and divergent "parent" sequences are typically flagged as potential chimeras. Multiple software packages for the detection of chimeras are available. The earliest ones were developed for the analysis of small sequence libraries and are generally not capable of analyzing large, high-throughput sequence libraries [e.g. Pintail (Ashelford *et al.*, 2005), Chimera Check (Cole *et al.*, 2007), Bellerophon (Huber *et al.*, 2004)]. Newer packages like ChimeraSlayer (Haas *et al.*, 2011), UChime (Edgar *et al.*, 2011) and B2C2 (Gontcharova *et al.*, 2010) are more frequently used for this purpose. Regardless of the chimera detection package that one chooses, users are cautioned to consider that the output generated only identifies *potential* chimeras. The results should be reviewed in greater detail, when possible, as "true" (i.e., nonchimeric) sequences can be flagged incorrectly.

21.7.1.5 The Operational Taxonomic Unit (OTU) Concept

The concept of a bacterial species can be difficult to define. Revisions to existing taxonomies are published on a regular basis with phylogenetic relationships constantly being redefined on the basis of new molecular information (Information Box 21.1). Horizontal gene transfer between individual bacteria obscures relationships that are defined on the basis of function, and it is widely acknowledged by microbiologists that we have only just begun to characterize and classify the extensive diversity of microbial species.

With all of this as a background, sequence-based surveys emerged as a means of characterizing individual bacteria and microbial communities at large. As a means of grappling with the questions of how to quickly distinguish one species from another when many species are present in a given sample, the concept of the operational taxonomic unit (OTU) emerged. DNA–DNA hybridization studies have long been a gold standard for defining species similarity, but scientists noticed that bacteria that share high levels of similarity via DNA–DNA hybridization also shared a high degree of similarity between their 16S rRNA gene sequences (Stackenbrandt and Goebel, 1994). This concept has also been applied to fungi (O'Brien *et al.*, 2005; Amend *et al.*, 2010), although the ITS region is typically utilized instead of the small ribosomal subunit.

The OTU is a computational construct that is used to represent species, and it is heavily utilized in the field of microbial ecology. OTUs are defined on the basis of sequence similarity, and typically a 97% sequence similarity cutoff is employed. That is, if two sequences have 97% of their base calls in common over the entire length of both sequences, they are considered to belong to the same OTU. OTUs are convenient in that they represent an entity that can be counted and used as the basis for diversity estimates (Schloss and Handelsman, 2005), and they are not tied to known biological diversity (i.e., they

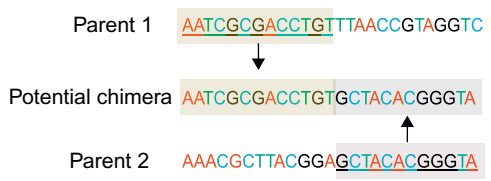


FIGURE 21.7 A chimeric sequence may be generated during the PCR process when DNA polymerase begins replicating one strand of DNA and finishes on another. The resulting chimera contains sequence from one parent template at the 5′ end and the other parent template at the 3′ end. The detection of chimeric sequences often involves BLAST-like searches of reference databases or the other reads produced in the same sequence library, in an attempt to identify reads that share a high degree of similarity with multiple "parent" sequences.

Information Box 21.1 The Evolving Taxonomy of Microorganisms

One of the challenges for phylogenetic classification of microbial communities and interpretation of these data is the dramatic evolution of microbial taxonomy, especially over the past few decades. For example, one of the most-studied 2,4-D-degrading bacteria was originally named *Alcaligenes eutrophus* JMP134 after its isolation from soil (Don and Pemberton, 1981); however, a search of the literature will find that this bacterium has been referred to by multiple names over the past three decades including:

***Alcaligenes eutrophus* JMP134**
↓
***Ralstonia eutropha* JMP134**
↓
***Wautersia eutropha* JMP134**
↓
***Cupriavidus necator* JMP134**
↓
***Cupriavidus pinatubonensis* JMP134**

These changes have occurred as the bacterium and related organisms have been reclassified in light of new information for a variety of properties including: lipid; composition; 16S rRNA gene sequence, DNA–DNA hybridization; and phenotype. Although these continual changes are improving the taxonomic classification of microorganisms, they can make it even more difficult to draw functional inferences for environmental microorganisms based solely upon comparison of sequence data for phylogenetic marker genes (e.g., 16S rRNA) to previously classified organisms, whose characterization may have been published under a different name(s) in the literature.

can be used to quantify previously undescribed or uncharacterized organisms). They also allow large, complex collections of sequence data to be summarized quickly in text format. However, one of the major downfalls of OTUs is that without additional characterization, they lack the ability to convey information about phylogenetic relationships, or the degree of similarity, shared with other OTUs. Although all of the sequences that belong to an OTU are, by definition, closely related to one another (i.e., 97% sequence similarity is often used as the cutoff for all sequences within an OTU), the ability to discern whether "OTU A" and "OTU B" are similar to one another can quickly become lost.

21.7.1.6 Diversity Analyses

What is diversity? Biological and ecological diversity are concepts that deal with richness, variability and variety within the context of an environmental system (i.e., a defined unit) (see also Chapter 19). This may be genetic diversity, organism diversity or ecological diversity (Magurran, 2004). In the context of microbial communities, we typically consider aspects of all three. Genetic diversity, often in the form of marker gene sequences, is used as a proxy to describe organism diversity (i.e., OTUs or species), and communities of microorganisms are compared with one another in an attempt to describe the richness and variation that exists within and between communities.

Two key concepts contribute to our understanding of diversity. For the sake of discussion, we will use the terms "species" and "communities" here, but other entities (e.g., genes, taxonomic families) could be used in their place. The first of these concepts is richness, or the number of different types of species that exist within a community. This is a relatively easy concept to define, but in practice it is often difficult to quantify with 100% certainty, because it is extremely difficult to sample microbial communities exhaustively.

The second concept that contributes to our understanding of diversity is evenness, a term which describes the variability of species abundances. An extremely "even" community is one in which all species are present in similar proportions. As an example of an even community, consider an assemblage that contains four species, each of which accounts for 25% of the individuals (or biomass) in the community. In contrast, an "uneven" community is one in which large disparities exist with respect to the relative abundances of its members. Like the example provided above, an uneven community could also contain four species, but in this case, one species accounts for 60% of the community, the second accounts for 30% of the community, the third accounts for 7% of the community and the last accounts for the remaining 3%.

As a means of communicating information about diversity, the concepts of richness and evenness are often communicated as a single value, known as a diversity index. Multiple diversity indices have been developed (Information Box 21.2), and each has strengths, weaknesses and biases (Magurran, 2004). A full discussion of these is beyond the scope of this chapter, but some of the most commonly utilized indices and their applications will be described here.

Alpha diversity refers to the diversity of a defined unit, sample, assemblage or habitat (Rosenzweig, 1995), and it is often described in terms of species or OTU richness, the Shannon (or Shannon–Weiner) index and/or the Simpson index. Because indices like Shannon and Simpson can be biased by disparities in sampling effort or sample size (Magurran, 2004), it is common to subsample sequence (or OTU) libraries to an even depth before calculating diversity index values in order to facilitate head-to-head comparisons between one's samples. Typically, this is accomplished by randomly selecting an equal number of sequences from each sample in one's study prior to the calculation of diversity values.

The Shannon index (H′) (Shannon and Weaver, 1949) is based on information theory and attempts to quantify the uncertainty surrounding one's ability to predict, in advance, the identity of an organism sampled at random from a dataset (or community). It is based on the idea that both the number of species in a community and their relative abundances contribute to the "complexity" of a community, and thus the likelihood of being able to correctly predict the identity of an organism randomly sampled from the community. The Shannon index is calculated as:

$$H' = -\Sigma \rho_i \ln \rho_i$$

where ρ_i is the proportion of the i^{th} species in the community. This could be species "A," "B," "C," etc. The value of the proportion of each species in the community multiplied by the log of that value is calculated for every species in the community, and then summed to generate the Shannon index score. Natural log, log 2 or log 10 can be used, but natural log is commonly employed. Although larger Shannon index values are generally considered to represent greater levels of diversity, the means by which the index is calculated make it difficult to interpret whether changes to the statistic are a result of changes to richness, evenness or both. Despite this, the Shannon index is commonly utilized to describe microbial diversity.

Like the Shannon index, Simpson's index (Simpson, 1949) deals with probabilities. More specifically, it attempts to define the probability of any two organisms being drawn from the same community belonging to the same species. Although defining these probabilities inherently deals with defining the number of species in a community (i.e., its richness), the Simpson index tends to have a greater focus on species dominance (i.e., evenness)

Information Box 21.2 Diversity Indices

Although diversity can be characterized at multiple levels (Chapter 19), diversity indices are most frequently used to describe alpha-diversity and beta-diversity. An introduction to commonly used diversity indices is provided below, but many others have been developed and may be encountered in the literature.

Commonly used alpha-diversity indices include:

Species richness (observed species)—a count of the number of unique species that occur in a sample or community

Shannon (H′)—the Shannon (or Shannon–Wiener) index considers both the number of unique species and their relative abundances within a sample (Shannon and Weaver, 1949). Larger values reflect communities with greater species richness and evenness, while lower numbers reflect communities with fewer species and/or a very uneven distribution among them (e.g., one species may account for a very large percentage of the community).

Simpson (D)—the Simpson index evaluates the relative abundances of all species in a community, and attempts to define the probability that any two organisms drawn from the same community will be of the same species (Simpson, 1949; Magurran, 2004). Small values of the Simpson index tend to reflect communities with high richness and low dominance, and high values reflect communities with (potentially) lower richness and high dominance (i.e., most of the community belongs to one or a few species). The Simpson index is often presented in inverse form (1/D or 1 − D) so that large numbers represent increasing evenness and smaller numbers represent increasing dominance.

Chao I—the Chao I index is a correction factor for observed richness. It evaluates the number of species that occur once (singletons) versus those that occur twice (doubletons), and attempts to estimate the number of species that would be captured if the entire community could be sampled exhaustively (Chao, 1984).

Rarefaction—Rarefaction is not an index, but rather a technique used to assess species richness. It involves plotting the number of unique species detected versus the number of organisms sampled. The shape of the resulting curve is used to indicate the "completeness" of a survey. A curve that flattens and reaches a clear asymptote suggests that the majority of the diversity in a community has been captured, while one that maintains a steep slope indicates that more sampling is needed.

Phylogenetic diversity—also known as Faith's diversity (Faith, 1992), this index quantifies the total length of the branches needed to account for a set of taxa on a phylogenetic tree. Increasing values of the index reflect increasing levels of diversity within the community being described.

Commonly used beta-diversity indices include:

Sørensen index—evaluates the degree of similarity between two communities by quantifying the number of species shared in common, relative to the total number of species held in both communities (Sørensen, 1948). This metric can be used with presence/absence (i.e., binary) data.

Jaccard index—evaluates the degree of similarity between two communities by quantifying the number of species shared in common relative to the sum of the number of species uniquely held by each community (Jaccard, 1908). This metric can be used with presence/absence data.

Bray–Curtis dissimilarity—Bray–Curtis dissimilarity (Bray and Curtis, 1957) is an extension of the Sørensen index. Calculated the same way, it is allows for quantitative values (i.e., counts or relative abundances) to be used instead of binary data.

Unifrac distance—a measurement that reflects the amount of branch length shared by two or more communities when their members are placed on a common phylogenetic tree (Lozupone and Knight, 2005). The Unifrac distance is equivalent to "1 minus the fraction of shared branch length."

than it does on species richness. The Simpson index (D) is calculated as:

$$D = \Sigma \rho_i^2$$

where dominance (D) is calculated as the sum of the squared proportions of all species in a given community. Large values of D are typically interpreted to represent high dominance and low diversity, whereas small values of D tend to represent lower dominance, higher diversity communities. Because the interpretation of these values is not necessarily intuitive, ecologists commonly calculate inverse Simpson (1/D) or subtract Simpson from 1(1 − D) to obtain a value that is more easily interpreted.

Once one has described the diversity within a community, it is common to want to compare diversity among communities along gradients (Whittaker, 1960) or separated by space and time. This is also known as beta diversity. A typical first step in describing beta diversity is to calculate the degree of similarity shared between communities in terms of their species composition, species distribution or both. The terms similarity and distance are frequently used to describe the degree to which two communities resemble one another, and they are the inverse of one another (i.e., a similarity of 30% is equivalent to a dissimilarity or distance of 70%).

Multiple approaches to calculating community similarity exist. As mentioned above, some indices, like the

Sörensen or Jaccard indices, consider only the presence or absence of species between two samples. Others, like the Bray–Curtis distance, Spearman distance, Hellinger distance, consider species' presence/absence and relative abundances. The Unifrac distance (Lozupone and Knight, 2005), a third type of measure, attempts to place community similarities and distances in a phylogenetic context, and quantifies the amount of phylogeny (i.e., branch length on a common phylogenetic tree) shared between two communities. Once similarity or dissimilarity values have been calculated among a set of communities, they are commonly communicated using ordination plots. Nonmetric multidimensional scaling (NMDS) and principal coordinates analysis (PCoA) plots are frequently used for this purpose (Figure 19.5).

21.7.1.7 Phylogenetic Analyses

Phylogenetic analysis is another route for analyzing marker gene sequences, especially in the case of 16S rRNA gene data. The use of the Unifrac metric helps to place communities in a phylogenetic context by first building a large phylogenetic tree, and then calculating the amount of the tree that is shared between two or more communities. The amount of phylogenetic diversity within a single sample can also be calculated this way.

From the perspective of single sequences or OTUs, phylogenetic analysis more typically involves trying to place a sequence or OTU of interest into phylogenetic context by comparing it with sequences of known origin. This process is similar to that which is used to generate OTUs: (1) a collection of sequences is gathered; (2) all sequences are compared with one another to determine the amount of sequence similarity that they share with one another; (3) these distances are interpreted and used to identify "nearest neighbors" (i.e., closest relatives), and may be used to construct a phylogenetic tree. This approach is commonly used to help describe the identity of a sequence or OTU whose best match in a public database is an uncultured or unclassified bacterium (or archaeaon or fungus). It has also been used to identify highly novel organisms and provide justification for the addition of new phyla (Hugenholtz *et al.*, 1998), and potentially even taxonomic domains (Wu *et al.*, 2011).

21.7.2 Bioinformatics and Analysis of Genomic/Metagenomic Data

Common first steps in the analysis of genomic or metagenomic data are an assessment of sequence quality and the removal of low-quality reads. Depending on the downstream analyses that will be performed, this can be a very important step. As each base is incorporated during a sequencing reaction, a score indicating the quality of each base call is also generated and recorded into the sequencing record. Often, base calls at the 5′ and 3′ ends of a sequence read are of lower quality than those that are incorporated in the middle. Likewise, overly long or extremely short reads also tend to be of lower quality, especially those produced on the 454 platform (Huse *et al.*, 2007). Quality scores can be used by some assembly algorithms, but many commonly used assemblers do not take them into consideration (Mende *et al.*, 2012). As a result, trimming and quality filtering of raw sequence reads is often advised, and tends to lead to more accurate genome and metagenome assemblies (DiGuistini *et al.*, 2009; Mende *et al.*, 2012).

21.7.2.1 Assembly-Based Approaches

Once a genome or metagenome has been sequenced, it is much like a jigsaw puzzle (or a collection of many jigsaw puzzles). It represents a large collection of pieces, some of which are informative on their own, and others of which yield better information and a more complete picture once they are assembled and placed in context with other fragments. Also like a jigsaw puzzle, genome and metagenomic sequence data may contain pieces (i.e., sequence fragments) that are duplicated, misshapen (i.e., contain errors) or missing. These add to the challenge of sequence assembly and interpretation, but they do not preclude it completely.

During the assembly process, fragments of sequence that originated from the same parent sequence are identified, and ordered relative to one another to build a larger, contiguous strand of sequence, also known as a contig. Contigs are typically constructed by identifying regions of common, overlapping sequence that are shared between the two smaller sequence fragments. Depending on the sequencing approach used, spatial information (i.e., known distances between fragments) may also be available to aid in the assembly process, and provide a degree of quality control. For example, if it is known that the ends of two different fragments should be oriented 1000 bp apart from one another, the distance can be used as a placeholder, which helps to constrain (i.e., control) the addition of new sequences and contigs. As multiple contigs are joined into longer and longer sequences, scaffolds are formed. Scaffolds are not necessarily contiguous runs of sequence, but can include gaps of known length. Depending on the complexity of the sample and the depth to which it is sequenced, assembly from metagenomic sequencing can yield high-quality draft, or even complete genome sequences.

Multiple approaches and software packages have been developed for the purpose of sequence assembly. The earliest assemblers were designed to piece together single genomes with fragments of relatively long read length. As the high-throughput sequencing of shorter gene fragments

and the sequencing of mixed communities (i.e., metagenomes) became more common, newer assemblers designed to handle greater levels of complexity were developed [e.g., Velvet (Zerbino and Birney, 2008), SOAP (Li *et al.*, 2010)].

Regardless of the sequence data or assembly algorithm used, the assembly process can be quite computationally intensive. As a result, preprocessing algorithms have emerged to help reduce the complexity and redundancy of the input data, and reduce the computational load required to complete the assembly (Pell *et al.*, 2012). This is particularly important in complex and highly diverse communities, such as those found in soil, where large amounts of sequence data must be generated in order to provide adequate coverage of the community.

21.7.2.2 Mapping to Reference Genomes

A reference genome, also known as a reference assembly, is a collection of nucleic acid sequence and annotation information describing the gene content of an organism. The nucleic acid sequences may be assembled (i.e., pieced together from smaller sequence fragments) into contigs, scaffolds or complete chromosomes. Often, open reading frames (ORFs) for individual genes will be identified, and attempts will be made to annotate or assign an identity and/or function to each of the ORFs.

Reference genomes play an important role in shaping our interpretation of new genomes and metagenomic data. Just as the "reference" portion of their name implies, reference genomes can serve as a framework for describing the gene content—both in terms of taxonomic origin and potential function—of new genomes and metagenomes. A common step in analyzing shotgun sequence datasets, like metagenomes, is to "map" the unassembled reads to a collection of reference genomes. This can be done using BLAST searches, but fast, memory-efficient alignment algorithms, such as bowtie (Langmead *et al.*, 2009) or BWA (Li and Durbin, 2010), are more commonly used for this purpose. The mapping algorithms search for regions of homology (i.e., similarity) between the reference genome and the sequence of interest. The amount of similarity that they share, the degree of coverage (i.e., amount of the genome that generates matches within your pool of shotgun sequences) and the depth of coverage (i.e., the number of copies of each gene or genome found within the shotgun sequence pool) influence the quality of the amount of information that can be derived from the "map."

The mapping of reads to reference genomes can be used to identify and remove host-derived reads if the community of interest has come from a plant, animal or insect host. By mapping metagenomic reads from various body sites sampled during the Human Microbiome Project to a reference (human) genome, it was discovered that human-derived reads accounted for approximately 1% of the sequences generated from stool, but 80% of more of the sequences generated from samples of saliva, the anterior nares (nostril) and vagina (Human Microbiome Project Consortium, 2012). While the identification and removal of host "contamination" represents an important application of reference genome mapping, the technique can also be used to evaluate the potential origins of your reads. For example, by mapping metagenomic sequence reads to reference genomes, researchers studying a mixed-community, cellulosic bioreactor system were able to determine that their reactor harbored a variety of cellulose- and xylose-degrading bacteria, including *Clostridium thermocellum*, *Thermoanaerobacterium thermosaccharolyticum* and *Moorella thermoacetica* (Hollister *et al.*, 2012). They also learned that their reactor-housed bacteria shared some similarity with previously sequenced *Bacillus* spp.; however, the degree of similarity was low enough and the maps sparse enough to suggest that they had encountered novel, or at least unsequenced, species.

Historically, collections of reference genomes have been biased toward the inclusion of model organisms, pathogens and other organisms of economic or biotechnological importance, but in recent years, large-scale sequencing projects like the Human Microbiome Project (HMP) (Nelson *et al.*, 2010) and the Genomic Encyclopedia of Bacteria and Archaea (GEBA) (Wu *et al.*, 2009) have increased the scope and size of reference genome collections, systematically generating new genome sequences in the attempt to fill out the underrepresented portions of the microbial tree of life. They have utilized innovative isolation and culture techniques (Pope *et al.*, 2011), single cell sequencing (Rinke *et al.*, 2013) and in some cases assembly from metagenome sequences (Hess *et al.*, 2011).

Since the first bacterial genome (*Haemophilus influenza*) was sequenced in the mid-1990s (Fleischmann *et al.*, 1995), the collection of publicly available reference genomes has grown to include >6000 high-quality draft or completed bacterial and archaeal reference genomes and >300 eukaryotic reference genomes (Genomes Online Database, http://www.genomesonline.org, July 2013). A recent evaluation of the publically available reference genome collection found that the addition of new genomes as a result of the Human Microbiome Project reference genome sequencing initiative resulted in a 20−40% improvement in read recruitment from human metagenome samples than would have been possible previously (Nelson *et al.*, 2010). Likewise, the recent release of >200 genomes generated by single cell sequencing is estimated to have increased the phylogenetic coverage of publically available reference genomes by $>11\%$ (Rinke *et al.*, 2013). Although this growth is impressive, much more work remains to be done (Fodor *et al.*, 2012). Our understanding and appreciation of the microbial world is fundamentally linked to the information contained in reference

genome collections, and it is anticipated that continued efforts to expand these collections will provide new insight into microbial structure, function and evolution.

21.7.2.3 Databases

As the ability to generate genome and metagenome sequence data has grown, so too has the need to analyze, store and share it. Even as improved algorithms for sequence assembly and annotation are developed, the archiving, analysis and dispersal of genome and metagenomic sequence data is no trivial task. Powerful servers with large storage capacity are typically required to handle the data associated with these large and ever-growing projects. These resource requirements are often greater than individual academic laboratories can support, but centralized databases and similar repositories also serve a valuable purpose in their ability to facilitate the sharing of data within the scientific community.

A variety of databases have been developed with these needs in mind. Some, like the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) or the European Nucleotide Archive (ENA) at the European Molecular Biology Laboratory, house compressed versions of the raw (or sometimes assembled) sequence data and associated metadata from genome and metagenomic sequencing projects. Others, like the Integrated Microbial Genomes and Metagenomes (IMG) system (Markowitz *et al*., 2010), MG-RAST (Meyer *et al*., 2008), CAMERA (Sun *et al*., 2011) and the EBI Metagenomics service (https://www.ebi.ac.uk/metagenomics/), will house data, but also allow users to upload genome or metagenomic sequence data and analyze it. Common options offered by these platforms often include sequence assembly, annotation and the ability to carry out comparative analyses.

In addition to the archiving and analysis of genome and metagenome data, the rapid growth of genomic and metagenomic sequencing projects had led to the need to track and catalogue them. Despite the fact that the costs associated with generating sequence have declined, storage and dissemination of data still remain a challenge, and preventing the duplication of projects can help to reduce these burdens. The Genomes OnLine Database (GOLD) (Pagani *et al*., 2012), first established in 1997, has emerged to fill this need. GOLD serves as a central repository for information about sequencing projects, including genomes and metagenomes, as well as genome resequencing projects, single cell sequencing projects and (meta)transcriptomes. Information catalogued in GOLD includes project type, sequencing status (e.g., targeted, in progress, complete), project metadata, organism phylogeny and contact information for the scientist or research group leading the project efforts.

21.7.3 Integration of 'Omics Data

The ability to generate multiple 'omics datasets from the same system, at the same point in time, has the potential to provide a highly detailed picture of the system's biology and ecology. Efforts to integrate multiple'omics technologies with one another are still relatively few, especially in mixed microbial communities, and they often rely on the layering of 'omics data onto reference pathways or the correlation of one 'omics data set with another.

21.7.3.1 Layering of -Omics Data Using Reference Databases

Reference databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al*., 2012), BioCyc and MetaCyc (Caspi *et al*., 2012), provide curated, and often experimentally verified, information regarding metabolic pathways, and the enzymes, reactions, compounds and genes that allow them to function. Although many metabolic pathways occur commonly across the Tree of Life, these reference databases also include information regarding the (known) taxonomic distribution of particular genes and pathways.

Reference databases like KEGG and MetaCyc can serve as a platform to support the exploration of 'omics data, and new software tools are emerging to allow these databases to serve as a platform for 'omics integration. The KEGG database can be accessed directly through a web-based interface (http://www.genome.jp/kegg/), allowing users to explore genes, compounds and pathways, or to input information about differentially detected genes and compounds to assess which functional pathways may be affected. Likewise, the interactive Pathways Explorer (iPath, http://pathways.embl.de/) is a web-based tool that allows for the visualization, analysis and customization of pathway information, and the Pathview package (Luo and Brouwer, 2013) is a standalone package for multi-'omics integration.

The value of multiple-'omics datasets is often realized in the ability of one data type to confirm or refute the results of another. For example, following the *Deepwater Horizon* oil spill in the Gulf of Mexico in 2010, Mason *et al*. (2012) used a combination of metagenomic, metatranscriptomic and single-cell genomic sequencing to characterize microbial community responses. The shotgun metagenomic sequencing results revealed that, relative to the microbial communities inhabiting uncontaminated seawater, the hydrocarbon-exposed communities were significantly enriched in genes related to motility, chemotaxis and aliphatic hydrocarbon degradation, as well as the degradation of more recalcitrant compounds, including benzene, toluene and polycyclic aromatic hydrocarbons (see also Chapter 31). Analysis of the transcriptome of

these communities confirmed that the expression of the chemotaxis, motility and aliphatic hydrocarbon degradation genes was significantly enhanced. Surprisingly, though, the transcriptomic analysis found that the expression of genes related to the degradation of the more recalcitrant compounds had not changed, demonstrating that although differences in gene abundance profiles can provide strong clues about how a system works, the addition of transcriptomic and/or metabolomic data can be important to pinpoint which genes, compounds and metabolites are actually being used.

21.7.3.2 Correlation and Network-Based Approaches

The layering of 'omics data onto reference pathways allows for the exploration of these datasets in the context of known, well-characterized reactions and pathways. In the case of multi-'omics studies, it allows functionally related data types (e.g., genes and compounds involved in the same reaction) to be considered in the context of known biology and biochemical reactions. While layering approaches can be very useful, they are not necessarily designed to convey global patterns and relationships within and between 'omics datasets.

In contrast, correlation and network-based approaches can be employed to identify co-occurrence and/or co-abundance patterns within and between 'omics datasets. Correlation and network-based approaches are relatively simple, and are often naïve to the errors inherent in 'omics measurements, and the biology that they are being used to describe. These approaches often employ Pearson or Spearman correlations and can be as simple as asking, "Does gene 'A' occur with similar abundances as genes 'B' and 'C' or metabolite 'D' across all of my samples, or do some bacteria always (or never) occur together in my samples?" While these may seem like simple questions, the potential does exist for correlations to identify artifacts of the data rather than true biological relationships (Friedman and Alm, 2012). As such, more sophisticated methods for examining correlations have been proposed. These include partial least squares regression (Pir *et al*., 2006), sparse correlations for compositional data (Friedman and Alm, 2012) and generalized boosted linear models (Faust *et al*., 2012).

Despite the potential pitfalls of correlation-based analyses, they have been used with success in many studies, and have the potential to reveal new insights into the biology of a system of interest. For example, within-'omic (i.e., analyses using a single-'omic technology) correlations have been used to assign functional context to genes of unknown identity (Wang *et al*., 2012; Buttigieg *et al*., 2013), identify genes, metabolites or bacteria that are associated with, or characteristic of, ecological or environmental subtypes (Bhavnani *et al*., 2011; Barberan *et al*., 2012; Greenblum *et al*., 2012), and provide insight into the culture of previously uncultivable organisms (Duran-Pinedo *et al*., 2011). Examples of cross-'omic (i.e., multi-'omic) correlations are fewer in number, but recent attempts to integrate transcriptome and metabolite datasets from laboratory chemostats (Pir *et al*., 2006) and metabolites and community composition in the human gut (McHardy *et al*., 2013) have been described in the literature. The integration of mixed 'omics datasets is considered to be at the forefront of science and has tremendous potential for characterizing environmental microorganisms; however, it represents a technological challenge that remains to be fully resolved, especially for the study of complex microbial communities.

QUESTIONS AND PROBLEMS

1. Discuss the potential advantages and disadvantages of the various 'omics approaches for characterizing environmental microorganisms.
2. What is the value of reference genomes?
3. What kind(s) of information can be learned from 16S rRNA gene sequences? From genomic or metagenomic sequencing?
4. Which 'omics approach provides the most direct indication of microbial activity.
5. Discuss the major quality criteria used for processing DNA sequence data.
6. What is microbial diversity? How can it be determined using 'omics -based approaches?
7. Discuss the major limitations of 'omics approaches for studying microbial community diversity.

REFERENCES AND RECOMMENDED READING

Abarenkov, K., Tedersoo, L., Nilsson, R. H., Vellak, K., Saar, I., Veldre, V., *et al*. (2010) PlutoF—a web based workbench for ecological and taxonomic research, with an online implementation for fungal ITS sequences. *Evol. Bioinform. Online* **6**, 189–196.

Amend, A. S., Seifert, K. A., Samson, R., and Bruns, T. D. (2010) Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 13748–13753.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* **71**, 7724–7736.

Bailly, J., Fraissinet-Tachet, L., Verner, M. C., Debaud, J. C., Lemaire, M., Wesolowski-Louvel, M., *et al*. (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. *ISME J.* **1**, 632–642.

Barberan, A., Bates, S. T., Casamayor, E. O., and Fierer, N. (2012) Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* **6**, 343–351.

Bhavnani, S. K., Victor, S., Calhoun, W. J., Busse, W. W., Bleecker, E., Castro, M., *et al*. (2011) How cytokines co-occur across asthma patients: from bipartite network analysis to a molecular-based classification. *J. Biomed. Inform.* **44**, S24–S30.

Bray, J. R., and Curtis, J. T. (1957) An ordination of the upland forest communities in southern Wisconsin. *Ecol. Monographs* **27**, 325–349.

Bundy, J. G., Davey, M. P., and Viant, M. R. (2009) Environmental metabolomics: a critical review and future perspectives. *Metabolomics* **5**, 3–21.

Buttigieg, P. L., Hankeln, W., Kostadinov, I., Kottmann, R., Yilmaz, P., Duhaime, M. B., *et al*. (2013) Ecogenomic perspectives on domains of unknown function: correlation-based exploration of marine metagenomes. *PLoS One* **8**, e50869.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336.

Carvalhais, L. C., Dennis, P. G., Tyson, G. W., and Schenk, P. M. (2012) Application of metatranscriptomics to soil environments. *J. Microbiol. Meth.* **91**, 246–251.

Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., *et al*. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **40**, D742–D753.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scand. J. Statist.* **11**, 265–270.

Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., *et al*. (2007) The Ribosomal Database Project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.* **35**, D169–D172.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., *et al*. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141–D145.

de Menezes, A., Clipson, N., and Doyle, E. (2012) Comparative metatranscriptomics reveals widespread community responses during phenanthrene degradation in soil. *Environ. Microbiol.* **14**, 2577–2588.

DiGuistini, S., Liao, N., Platt, D., Robertson, G., Seidel, M., Chan, S. K., *et al*. (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol.* **10**, R94.

Don, R. H., and Pemberton, J. M. (1981) Properties of six pesticide degradation plasmids isolated from *Alcaligenes paradoxus* and *Alcaligenes eutrophus*. *Appl. Environ. Microbiol.* **145**, 681–686.

Dunn, W. B. (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian, and plant metabolomes. *Phys. Biol.* **5**, 1–24.

Duran-Pinedo, A. E., Paster, B., Teles, R., and Frias-Lopez, J. (2011) Correlation network analysis applied to complex biofilm communities. *PLoS One* **6**, e28438.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200.

Faith, D. P. (1992) Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., *et al*. (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, e1002606.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Fodor, A. A., DeSantis, T. Z., Wylie, K. M., Badger, J. H., Ye, Y., Hepburn, T., *et al*. (2012) The "Most Wanted" taxa from the human microbiome for whole genome sequencing. *PLoS One* **7**, e41294.

Friedman, J., and Alm, E. J. (2012) Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687.

Gontcharova, V., Youn, E., Wolcott, R. D., Hollister, E. B., Gentry, T. J., and Dowd, S. E. (2010) Black Box Chimera Check (B2C2): a windows-based software for batch depletion of chimeras from bacterial 16S rRNA gene datasets. *Open Microbiol. J.* **4**, 47–52.

González, J. M., and Robb, F. T. (2000) Genetic analysis of *Carboxydothermus hydrogenoformans* carbon monoxide dehydrogenase genes cooF and cooS. *FEMS Microbiol. Lett.* **191**, 243–247.

Gosalbes, M. J., Durban, A., Pignatelli, M., Abellan, J. J., Jimenez-Hernandez, N., Perez-Cobas, A. E., *et al*. (2011) Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS One* **6**, 1–9.

Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012) Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 594–599.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., *et al*. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21**, 494–504.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R248–R249.

Hemme, C. L., Deng, Y., Gentry, T. J., Fields, M. W., Wu, L., Barua, S., *et al*. (2010) Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J.* **4**, 660–672.

Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., *et al*. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–467.

Hettich, R. L., Pan, C., Chorney, K., and Giannone, R. J. (2013) Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communites. *Anal. Chem.* **85**, 4203–4214.

Hollister, E. B., Forrest, A. K., Wilkinson, H. H., Ebbole, D. J., Tringe, S. G., Malfatti, S. A., *et al*. (2012) Mesophilic and thermophilic conditions select for unique but highly parallel microbial communities to perform carboxylate platform biomass conversion. *PLoS One* **7**, e39689.

Huber, T., Faulkner, G., and Hugenholtz, P. (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317–2319.

Hugenholtz, P., Goebel, B. M., and Pace, N. R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 4765–4774.

Human Microbiome Project Consortium (2012) A framework for human microbiome research. *Nature* **486**, 215–221.

Hunter, C., Cochrane, G., Apweiler, R., and Hunter, S. (2011) Chapter 38: the EBI Metagenomics Archive, integration and analysis resource. In "Handbook of Molecular Microbial Ecology, Volume I: Metagenomics and Complementary Approaches" (F. J. de Bruijn, ed.), Wiley-Blackwell, Hoboken, New Jersey, U.S.A, pp. 333–340.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* **7**, 1889–1898.

Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., *et al*. (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science* **316**, 593–597.

Ito, T., Tanaka, M., Shinkawa, H., Nakada, T., Ano, Y., Kurano, N., *et al*. (2013) Metabolic and morphological changes of an oil accumulating trebouxiophycean alga in nitrogen-deficient conditions. *Metabolomics* **9**, S178–S187.

Jaccard, P. (1908) Nouvelles recherches sur la distribution florale. *Bulletin de la Societé Vaudoise des Sciences Naturelles* **44**, 223–270.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114.

Kim, S. I., Kim, J. Y., Yun, S. H., Kim, J. H., Lee, S. H., and Lee, C. (2004) Proteome analysis of *Pseudomonas* sp. K82 biodegradation pathways. *Proteomics* **4**, 3610–3621.

Kyle, J. L., Parker, C. T., Goudeau, D., and Brandl, M. T. (2010) Transcriptome analysis of *Escherichia coli* O157:H7 exposed to lysastes of lettuce leaves. *Appl. Environ. Microbiol.* **76**, 1375–1387.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.

Laserson, J., Jojic, V., and Koller, D. (2011) Genovo: de novo assembly for metagenomes. *J. Comput. Biol.* **18**, 429–443.

Laskin, R. S. (2012) Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10**, 631–640.

Li, H., and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., *et al*. (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272.

Liebeke, M., Brozel, V. S., Hecker, M., and Lalk, M. (2009) Chemical characterization of soil extract as growth media for the ecophysiological study of bacteria. *Appl. Microbiol. Biotechnol.* **83**, 161–173.

Lozupone, C., and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235.

Luo, W., and Brouwer, C. (2013) Pathview: an R/bioconductor package for pathway-based data integration and visualization. *Bioinformatics* **29**, 1830–1831.

Magurran, A. E. (2004) "Measuring Biological Diversity," Blackwell Publishing, Maldan, MA.

Markowitz, V. M., Chen, I. M, Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., *et al*. (2010) The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res.* **38**, D382–D390.

Mason, O. U., Hazen, T. C., Borglin, S., Chain, P. S. G., Dubinsky, E. A., Fortney, J. L., *et al*. (2012) Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J.* **6**, 1715–1727.

McHardy, I., Goudarzi, M., Tong, M., Ruegger, P., Schwager, E., Weger, J., *et al*. (2013) Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* **1**, 17.

Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., *et al*. (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One* **7**, e31386.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., *et al*. (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386.

Mueller, R. S., Dill, B. D., Pan, C., Belnap, C. P., Thomas, B. C., VerBerkmoes, N. C., *et al*. (2011) Proteome changes in the initial bacterial colonist during ecological succession in an acid mine drainage biofilm community. *Environ. Microbiol.* **13**, 2279–2292.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., *et al*. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329.

Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., Wortman, J. R., *et al*. (2010) A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999.

O'Brien, H. E., Parrent, J. L., Jackson, J. A., Moncalvo, J. M., and Vilgalys, R. (2005) Fungal community analysis by large-scale sequencing of environmental samples. *Appl. Environ. Microbiol.* **71**, 5544–5550.

Pagani, I., Liolios, K., Jansson, J., Chen, I. M., Smirnova, T., Nosrat, B., *et al*. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40**, D571–D579.

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., and Brown, C. T. (2012) Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13272–13277.

Pir, P., Kirdar, B., Hayes, A., Onsan, Z. I., Ulgen, K., and Oliver, S. (2006) Integrative investigation of metabolic and transcriptomic data. *BMC Bioinf.* **7**, 203.

Podar, M., Abulencia, C. B., Walcher, M., Hutchison, D., Zengler, K., Garcia, J. A., *et al*. (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73**, 3205–3214.

Pope, P. B., Smith, W., Denman, S. E., Tringe, S. G., Barry, K., Hugenholtz, P., *et al*. (2011) Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. *Science* **333**, 646–648.

Ram, R. J., VerBerkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., *et al*. (2005) Community proteomics of a natural microbial biofilm. *Science* **308**, 1915–1920.

Riehle, K., Coarfa, C., Jackson, A., Ma, J., Tandon, A., Paithankar, S., *et al*. (2012) The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences. *BMC Bioinformatics* **13**, S11.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J-F., *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437.

Rosenzweig, M. L. (1995) "Species Diversity in Space and Time," Cambridge University Press, Cambridge, UK.

Schloss, P. D., and Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71**, 1501–1506.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., *et al.* (2009) Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541.

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**, e27310.

Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M., and Rosenow, C. (2003) Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.* **13**, 216–223.

Shannon, C. E., and Weaver, W. (1949) "The Mathematical Theory of Communication," University of Illinois Press, Urbana, IL.

Simpson, E. H. (1949) Measurement of diversity. *Nature* **163**, 688.

Stackenbrandt, E., and Goebel, B. M. (1994) Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Bacteriol.* **44**, 846–849.

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., *et al*. (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.* **39**, D546–D551.

Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab* **5**, 1–34.

Thomas, T., Gilbert, J., and Meyer, F. (2012) Metagenomics—a guide from sampling to data analysis. *Microb. Inform. Exp.* **2**, 3.

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

VerBerkmoes, N. C., Denef, V. J., Hettich, R. L., and Banfield, J. F. (2009) Functional analysis of natural microbial consortia using community proteomics. *Nat. Rev. Microbiol.* **7**, 196–205.

Wang, P. I., Hwang, S., Kincaid, R. P., Sullivan, C. S., Lee, I., and Marcotte, E. M. (2012) RIDDLE: reflective diffusion and local extension reveal functional associations for unannotated gene sets via proximity in a gene network. *Genome Biol.* **13**, R125.

Wessel, A. K., Hmelo, L., Parsek, M. R., and Whiteley, M. (2013) Going local: technologies for exploring bacterial microenvironments. *Nat. Rev. Microbiol.* **11**, 337–348.

Westermeier, R., and Naven, T. (2002) Proteomics technology. In "Proteomics in Practice: A Laboratory Manual of Proteome Analysis" (J. Adams, ed.), Wiley-VCH, Darmstadt, Germany, pp. 1–160.

Whittaker, R. H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol. Monographs* **30**, 279–338.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., *et al*. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056–1060.

Wu, D., Wu, M., Halpern, A., Rusch, D. B., Yooseph, S., Frazier, M., *et al*. (2011) Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. *PLoS One* **6**, e18011.

Wu, M., Ren, Q., Durkin, A. S., Daugherty, S. C., Brinkac, L. M., Dodson, R. J., *et al*. (2005) Life in hot carbon monoxide: the complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901. *PLoS Genet.* **1**, 563–574.

Zengler, K., Walcher, M., Clark, G., Haller, I., Toledo, G., Holland, T., *et al*. (2005) High-throughput cultivation of microorganisms using microcapsules. *Methods Enzymol.* **397**, 124–130.

Zerbino, D. R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829.

Zhang, W., Li, F., and Nie, L. (2010) Integrating multiple 'omics analysis for microbial biology: applications and methodologies. *Microbiology* **156**, 287–301.