1 Introduction

1.1 Exploring the Microbial World

Microorganisms shaped the geochemical evolution of our planet throughout its history, and they continue to play a key role in the modern world. In deep time they oxygenated Earth's atmosphere and set the stage for life as we know it. Today, microbes mediate global biogeochemical cycles, influence the speciation and fate of pollutants, and modulate climate change through production and consumption of greenhouse gases. The field of geomicrobiology and microbial geochemistry (GMG), which studies the interplay between microbes and the Earth system, has roots in the 19th century (Druschel & Kappler 2015; Druschel et al. 2014). However, only recently has the breadth of microbial geomicrobiological processes and extent to which they shape geological, geochemical, and environmental processes become clear. Many methods and concepts central to GMG are also relevant to environmental engineering (e.g., drinking water and wastewater treatment) and medicine (e.g., human microbiome), including the omics approaches that are the focus of this book.

How to study this microbial world? Inherent challenges abound; microorganisms are small. Their cellular morphology is typically not informative of their phylogeny, physiology, or role in biogeochemical or ecological processes. Microbes often live in highly diverse microbial communities where it is hard to decipher the activities of different microorganisms or to trace specific microbial processes. Traditional microbiological approaches revolve around the cultivation of bacteria and archaea, which enables powerful laboratory-based methods of dissecting microbial physiology, biochemistry, and genetics as they relate to geochemical processes (Newman et al. 2012). Yet most microorganisms in nature are resistant to cultivation owing to symbiotic lifestyles or unknown nutritional requirements (Staley & Konopka 1985). Further, it can

Genomic Approaches in Earth and Environmental Sciences, First Edition. Gregory Dick.

be impractical to grow pure cultures due to the extremely slow growth of many microorganisms, which in the environment is perhaps more akin to stationary phase than to growing cultures (Roy et al. 2012). Comprehensive culturing is also impractical because of the stunning complexity of natural microbial communities (thousands of species). Finally, the results from pure cultures may not be representative of *in situ* processes (Madsen 2005).

Traditional geochemical methods of measuring process rates and products and using biological poisons or inhibitors of specific microbial enzymes offer critical quantitative data and some mechanistic insights (Madsen 2005; Oremland et al. 2005). However, these approaches provide little information with regard to the identity or nature of the microorganisms that underpin processes of interest. Exciting advances in microscopy and spectroscopy that provide opportunities to link microorganisms to biogeochemical processes are described and reviewed elsewhere (Behrens et al. 2012; Newman et al. 2012; Wagner 2009).

Recent advances in DNA sequencing technologies open up entirely new avenues to study geomicrobiology by circumventing the cultivation step and providing extensive information on microorganisms as they exist in natural settings. This data comes from the sequence of macromolecules (Box 1.1)

Box 1.1 Definitions of key macromolecules studied by omics approaches

Deoxyribonucleic acid (DNA): DNA consists of four nucleotide bases – guanine (G), adenine (A), thymine (T), and cytosine (C) – that are joined together in a sequence to form genes.

Gene: a unit of genetic information encoding protein, tRNA, or ribosomal RNA. Genes are about 1000 bases long, on average.

Genome: the genome is the collection of all genetic information in an organism, including the genes as well as elements between genes that are involved in regulating gene expression. Microbial genomes range in size from approximately 400 000 to 10 million bases and from 400 to 10 000 genes.

Ribonucleic acid (RNA): There are several major forms of RNA, including messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal rRNA (rRNA). mRNA is an intermediate between DNA and protein (see Fig. 1.1); rRNA is a structural and catalytic component of ribosomes, the machinery that translates mRNA into protein. tRNA are small molecules that recognize the three-base code of mRNA and translate it into amino acids during protein synthesis.

Protein: proteins are polymers (long chains) of amino acids. The two main roles of proteins are (1) to provide structure or scaffolding, e.g., in cell wall or protein synthesis; (2) to catalyze biochemical reactions in the cell, including those required for energy metabolism, biosynthesis of macromolecules, transport of elements into and out of the cell, and generation of biogenic minerals ("biomineralization"). Proteins can also "sense" the environment and transduce signals that elicit cellular responses.

Lipids: hydrocarbons, often with polar head groups, that are the primary constituents of cell membranes. In some cases, specific lipids are diagnostic of specific microbial groups or metabolisms. Unlike other biological macromolecules, lipids may be preserved in sediments over geological time (millions to billions of years), so they have great value in potentially providing information on ancient ecosystems. Like other macromolecules, the synthesis of lipids is conducted by proteins that are encoded by genes. Hence, the "lipidome" can theoretically be predicted from the genome.

Carbohydrates: macromoleucles consisting of carbon, hydrogen. Carbohydrates decorate the cell surface and are an important interface between the cells and their environment. Because they are often negatively charged, they can play important roles in binding cations and influencing biomineralization.

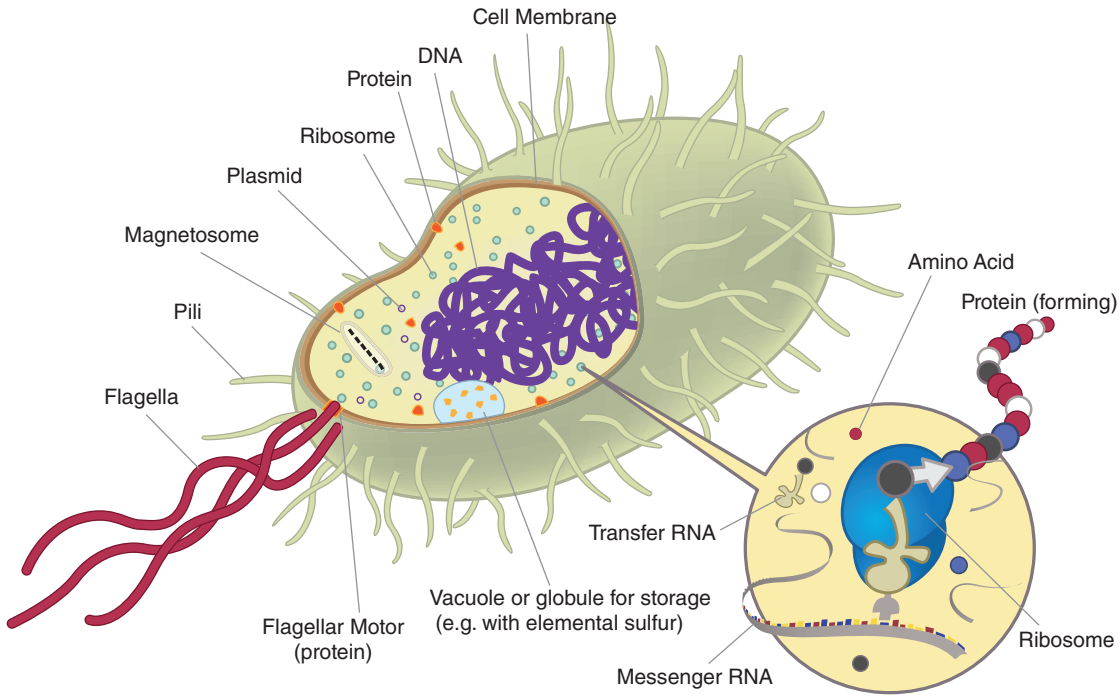

Figure 1.1 Generalized structure of a bacterial or archaeal cell. Inset details translation and protein synthesis. Source: Druschel and Kappler (2015), p. 390, Fig. 1. Reproduced with permission from the Mineralogical Society of America.

that constitute microbial cells (Fig. 1.1). This book focuses on DNA, RNA, and protein, and also touches on lipids and the pool of small molecules within a cell (metabolites). The collection of genes that encode an organism is known as the genome. Genes are transcribed as messenger RNAs, or transcripts, the total pool of which is called the transcriptome. Transcripts are then translated into protein, which actually performs the structural and biochemical functions of the cell. The total protein content of a cell is known as the proteome. The total content of small molecules within a cell is referred to as the metabolome. These small molecules include metabolites, the substrates, intermediates, and products of biochemical reactions catalyzed by enzymes. The study of the whole collection of each of these molecules in a pure culture is referred to as genomics, transcriptomics, proteomics, and metabolomics. When such information is derived from a whole community of microorganisms, we say "community genomics" or "metagenomics" (or metatranscriptomics, metaproteomics). Collectively, these approaches, whether applied to a single organism or a community of organisms, are referred to in shorthand as "omics."

Whereas genomes encode all the proteins that could possibly be made in a given cell, a genome does not give any information about which proteins and RNA are actually being produced at any given time, or about the quantities in which they are produced. Transcriptomics and proteomics provide

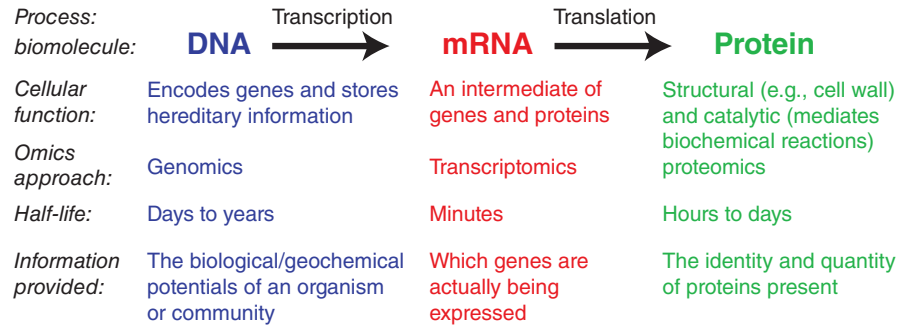

Figure 1.2 Macromolecules that serve as the basis for the three main omics approaches. Source: Dick and Lam (2015), p. 404, Fig. 1, with permission from the Mineralogical Society of America.

this information. DNA, RNA, and protein have different lifetimes based on the stability of the molecules and the biochemical mechanisms that degrade them. Thus these molecules provide information at different time scales (Fig. 1.2). Genomes also provide a "molecular fossil record" of how genes and organisms have evolved over the billions of years of life on Earth (David & Alm 2011; Macalady and Banfield 2003; Zerkle et al. 2005).

1.2 The DNA Sequencing Revolution: Historical Perspectives

The "meta-omics" revolution has its roots in the pioneering work of Carl Woese and colleagues, who sequenced microbial rRNA genes in order to uncover their phylogenetic relationships (Woese & Fox 1977). This work recognized that, because rRNA genes serve critical functions, they are present in every organism and are highly conserved at the sequence level. Thus, they hold invaluable information about the evolutionary relationships of microorganisms. Through painstaking labor, the sequence of rRNA genes from a wide range of organisms was deciphered, leading to an astonishing discovery: methane-producing microorganisms previously assumed to be bacteria were actually a new and completely separate domain of life – the archaea (Sapp & Fox 2013). This transformed our understanding of the tree of life by revealing that it is composed of three domains: bacteria, archaea, and eukarya (Woese & Fox 1977). The advent of rRNA gene sequencing also provided a practical and objective tool for classifying microorganisms, a task which had been declared impossible previously (Woese & Goldenfeld 2009).

Soon after, Pace and colleagues applied sequencing to rRNA genes purified directly from uncultured communities of microorganisms (Stahl et al. 1984). Subsequent application of polymerase chain reaction (PCR) to the amplification of rRNA genes (with an explicit focus on one of these genes,

known as 16S rRNA) directly from the environment increased the throughput of this approach and revealed startling insights into the microbial world in seawater and other environments (DeLong 1992; Fuhrman et al. 1992). Spurred by rapidly advancing technologies and the ever declining costs and increasing throughput of DNA sequencing technologies (Loman et al. 2012), the culture-independent approach quickly revealed the staggering diversity of the microbial world (Pace 2009). This work revealed that only a tiny fraction of microbial groups have been studied in culture (Baker & Dick 2013; Pace 2009).

In parallel with the explosion of 16S rRNA gene sequencing, faster, cheaper DNA sequencing also enabled a new era of sequencing whole microbial genomes (Land et al. 2015). Information on the complete gene content theoretically provides a picture of the metabolic and physiological potential of microorganisms (however, see the caveats and challenges discussed in Chapter 3). The first bacterial genomes were published in 1995 (Fleischmann et al. 1995; Fraser et al. 1995), and the number of microbial genomes sequenced has expanded exponentially ever since (Fournier et al. 2013).

A major initial finding of these sequencing efforts was that microbial genomes have startling variability of gene content (Tettelin et al. 2005; Welch et al. 2002). This led to concepts of the pangenome, core genome, and flexible genome (Cordero & Polz 2014) (see Chapter 2). Genome sequences from cultured organisms are valuable because they enable studies of the links between genotype and phenotype and represent taxonomic and functional anchors in the tree of life for interpreting metagenomic data. Particularly valuable are genomes from type strains that have been validly described and named, which are estimated to account for a substantial portion (~15%) of phylogenetic diversity (Kyrpides et al. 2014). However, despite the microbial genome sequencing revolution, less than 3% of these type strains have had their genomes sequenced (Kyrpides et al. 2014). Thus, even the genomic coverage of *cultured* microbial life remains woefully inadequate, and of course, the cultured portion is just a small fraction of the total microbial world. The Microbial Earth Project (www.microbial-earth.org/) was recently launched to track the inventory of type strains of bacteria and archaea and their genome sequencing projects.

At the confluence of environmental 16S rRNA gene sequencing of microbial communities and whole genome sequencing of cultured microbes is the direct retrieval of genomes from uncultured microbial communities. Early metagenomic approaches used cloning of environmental DNA followed by sequencing and/or screening of expressed products for functions of interest (Riesenfeld et al. 2004; Stein et al. 1996). The term "metagenomics" was first coined in 1998, in the context of accessing natural products (e.g., antibiotics) from uncultured soil microorganisms (Handelsman et al. 1998). The power of the functional metagenomics approach lies in the direct connection of sequence to function and was illustrated beautifully by the discovery of bacterial light-driven proton pumps as a new form of phototrophy in the

oceans (Béjà et al. 2000). This method can also provide valuable insights by directly linking phylogenetic marker genes to function (Pham et al. 2008), which is particularly valuable when the cloned fragments are large, as in BAC or fosmid libraries. However, because of the cost and labor involved in constructing and screening such clone libraries, this approach was not readily scalable. The "functional metagenomic" approach also faces practical challenges such as genetic and biochemical incompatibility between environmental genes and hosts (e.g., differences in codon bias, required cofactors). Some of these issues can be overcome by recent synthetic genomic approaches, but they still limit the throughput of exploratory, discovery-driven functional screening.

Shotgun metagenomics, in which community DNA is randomly fragmented and sequenced, was then demonstrated as a viable and valuable approach (Tyson et al. 2004; Venter et al. 2004) and quickly emerged as the dominant method used in metagenomics studies. For the first time, whole genomes of uncultured organisms could be reconstructed from microbial communities, revealing their metabolic potential (Tyson et al. 2004) and evolutionary processes (Allen & Banfield 2005). Several spectacular discoveries, including the linking of ammonia oxidation to archaea (Venter et al. 2004), demonstrated the power and promise of metagenomics. A vision for the potential advances that metagenomics could bring to science and society was beginning to come into view (National Research Council 2007). Hugenholtz and Tyson (2008) recount a brief history and highlights of these early stages and different approaches of metagenomics. For a more in-depth historical account see Handelsman (2004) and Gilbert and Dupont (2011). The rapid decrease in costs and increase in throughput of DNA sequencing has enabled shotgun sequencing of more complex microbial communities (Fig. 1.3). Recent papers report the reconstruction of thousands of genomes from metagenomes (Anantharaman et al. 2016).

While the genomic sequence provides information on the metabolic and physiological *potential* of microorganisms, it does not indicate whether those functions are being carried out at a particular point in time or space. To address this question, characterizing the expression of mRNA or protein is required. Metatranscriptomics was applied with great success to surface seawater microbial communities, revealing that flexible genes are highly expressed in the environment (Frias-Lopez et al. 2008). Critically, this paper also used qPCR to independently evaluate the accuracy of RNA amplification, which is required to obtain sufficient cDNA for many sequencing applications (see Chapter 9).

Whereas the DNA- and RNA-based analyses described above rely on the sequencing of nucleotides, proteomic tools use mass spectrometry to accurately measure the masses of small peptide fragments and even individual amino acids. The matching of these measured masses with calculated peptide masses derived from genomic information enables the identification of protein fragments. Metaproteomics is challenging because analytical methods for translating MS/MS spectra into protein sequence are complex

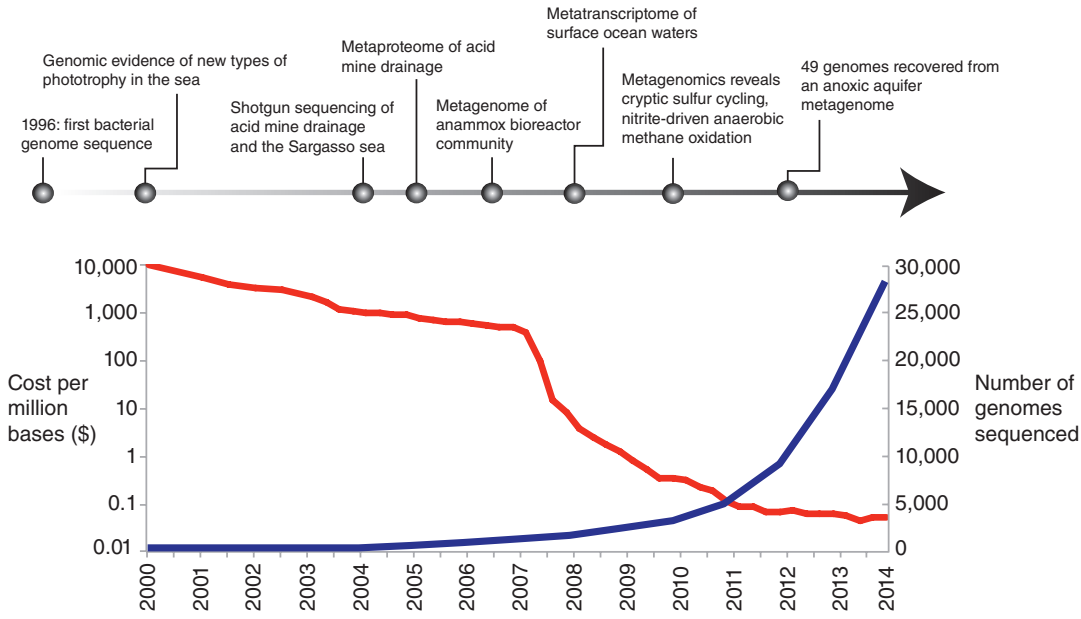

Figure 1.3 Major milestones in microbial community omics (*top*) and the decreasing cost and increasing throughput of DNA sequencing (*bottom*).

Source: Modified from Dick and Lam (2015), p. 406, Fig. 3, permission from the Mineralogical Society of America.

and largely reliant on having the corresponding genomic sequence for interpretation. Similarly, the recovery of total protein from many environmental samples is more challenging than the extraction of DNA and RNA. Not surprisingly, initial progress on application of proteomics to microbial communities was accomplished in low-diversity communities for which genomic sequence was available (Ram et al. 2005; Verberkmoes et al. 2009). Indeed, with sufficient genomic information, protein expression from very closely related strains can be differentiated (Lo et al. 2007). These studies yielded insights into the biochemical mechanisms of iron oxidation, a central process sustaining primary production and pyrite dissolution in acid mind drainage, and showed that among the most highly expressed proteins are "hypothetical" and "conserved hypothetical" proteins (Ram et al. 2005). With growing databases of genomic sequence and improving algorithms for interpreting MS/MS spectra, metaproteomics is now a viable approach for studying more complex microbial communities (see Chapter 10).

References

Allen, E. E. & Banfield, J. F. (2005) Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology*, **3**, 489–498.

Anantharaman, K., Brown, C. T., Hug, L. A., et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, **7**, 13219.

Baker, B. J. & Dick, G. J. (2013) Omic approaches in microbial ecology: charting the unknown. *Microbe*, **8**, 353–360.

Behrens, S., Kappler, A. & Obst, M. (2012) Linking environmental processes to the in situ functioning of microorganisms by high-resolution secondary ion mass spectrometry (NanoSIMS) and scanning transmission X-ray microscopy (STXM). *Environmental Microbiology*, **14**, 2851–2869.

Béjà, O., Aravind, L., Koonin, E. V., et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, **289**, 1902–1906.

Cordero, O. X. & Polz, M. F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology*, **12**, 263–273.

David, L. A. & Alm, E. J. (2011) Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, **469**, 93–96.

Delong, E. F. (1992) Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 5685–5689.

Dick, G. J. & Lam, P. (2015) Omics approaches to microbial geochemistry. *Elements*, **11**, 403–408.

Druschel, G. K. & Kappler, A. (2015) Geomicrobiology and microbial geochemistry. *Elements*, **11**, 389–394.

Druschel, G. K., Dick, G. J. & Boyd, E. S. (2014) Geomicrobiology and Microbial Geochemistry 2014 Workshop Report. Available at: https://dx.doi.org/10.6084/m9.figshare.3083524.v1 (accessed 25 October 2017).

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

Fournier, P. E., Drancourt, M., Colson, P., Rolain, J. M., La Scola, B. & Raoult, D. (2013) Modern clinical microbiology: new challenges and solutions. *Nature Reviews Microbiology*, **11**, 574–585.

Fraser, C. M., Gocayne, J. D., White, O., et al. (1995) The minimal gene complement of Mycoplasma–Genitalium. *Science*, **270**, 397–403.

Frias-Lopez, J., Shi, Y., Tyson, G. W., et al. (2008) Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 3805–3810.

Fuhrman, J. A., Mccallum, K. & Davis, A. A. (1992) Novel major archaebacterial group from marine plankton. *Nature*, **356**, 148–149.

Gilbert, J. A. & Dupont, C. L. (2011) Microbial metagenomics: beyond the genome. *Annual Review of Marine Science*, **3**, 347–371.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, **68**, 669–685.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry and Biology*, **5**, R245–259.

Hugenholtz, P. & Tyson, G. W. (2008) Metagenomics. *Nature*, **455**, 481–483.

Kyrpides, N. C., Hugenholtz, P., Eisen, J. A., et al. (2014) Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *Plos Biology*, **12**, e1001920.

Land, M., Hauser, L., Jun, S. R., et al. (2015) Insights from 20 years of bacterial genome sequencing. *Functional and Integrative Genomics*, **15**, 141–161.

Lo, I., Denef, V. J., Verberkmoes, N. C., et al. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature*, **446**, 537–541.

Loman, N. J., Constantinidou, C., Chan, J. Z. M., et al. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology*, **10**, 599–606.

Macalady, J. & Banfield, J. F. (2003) Molecular geomicrobiology: genes and geochemical cycling. *Earth and Planetary Science Letters*, **209**, 1–17.

Madsen, E. L. (2005) Identifying microorganisms responsible for ecologically significant biogeochemical processes. *Nature Reviews Microbiology*, **3**, 439–446.

National Research Council (2007) *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. National Academies Press, Washington DC.

Newman, D. K., Orphan, V. J. & Reysenbach, A. L. (2012) Molecular biology's contributions to geobiology. In: A.H. Knoll & K.O. Konhauser (eds), *Fundamentals of Geobiology*. Blackwell Publishing, Chichester.

Oremland, R. S., Capone, D. G., Stolz, J. F. & Fuhrman, J. (2005) Whither or wither geomicrobiology in the era of 'community metagenomics'? *Nature*, **3**, 572–578.

Pace, N. R. (2009) Mapping the tree of life: progress and prospects. *Microbiology and Molecular Biology Reviews*, **73**, 565–576.

Pham, V. D., Konstantinidis, K. T., Palden, T. & Delong, E. F. (2008) Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical profile in the North Pacific Subtropical Gyre. *Environmental Microbiology*, **10**, 2313–2330.

Ram, R. J., Verberkmoes, C., Thelen, M. P., et al. (2005) Community proteomics of a natural microbial biofilm. *Science*, **308**, 1915–1920.

Riesenfeld, C. S., Goodman, R. M. & Handelsman, J. (2004) Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental Microbiology*, **6**, 981–989.

Roy, H., Kallmeyer, J., Adhikari, R. R., Pockalny, R., Jorgensen, B. B. & D'hondt, S. (2012) Aerobic microbial respiration in 86-million-year-old deep-sea red clay. *Science*, **336**, 922–925.

Sapp, J. & Fox, G. E. (2013) The singular quest for a universal tree of life. *Microbiology and Molecular Biology Reviews*, **77**, 541–550.

Stahl, D. A., Lane, D. J., Olsen, G. J. & Pace, N. R. (1984) Analysis of hydrothermal vent–associated symbionts by ribosomal-RNA sequences. *Science*, **224**, 409–411.

Staley, J. T. & Konopka, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews of Microbiology*, **39**, 321–346.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & Delong, E. F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology*, **178**, 591–599.

Tettelin, H., Masignani, V., Cieslewicz, M. J., et al. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 16530–16530.

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

Venter, J. C., Remington, K., Heidelberg, J. F., et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Verberkmoes, N. C., Denef, V. J., Hettich, R. L. & Banfield, J. F. (2009) Systems biology functional analysis of natural microbial consortia using community proteomics. *Nature Reviews Microbiology*, **7**, 196–205.

Wagner, M. (2009) Single-cell ecophysiology of microbes as revealed by Raman microspectroscopy or secondary ion mass spectrometry imaging. *Annual Review of Microbiology*, **63**, 411–429.

Welch, R. A., Burland, V., Plunkett, G. 3rd, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 17020–17024.

Woese, C. R. & Fox, G. E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5088–5090.

Woese, C. R. & Goldenfeld, N. (2009) How the microbial world saved evolution from the Scylla of molecular biology and the Charybdis of the modern synthesis. *Microbiology and Molecular Biology Reviews*, **73**, 14–21.

Zerkle, A. L., House, C. H. & Brantley, S. L. (2005) Biogeochemical signatures through time as inferred from whole microbial genomes. *American Journal of Science*, **305**, 467–502.