Marine Environmental Genomics: Unlocking the Ocean's Secrets

BY ROBERT A. EDWARDS AND ELIZABETH A. DINSDALE

In 1944, Oswald Avery, Colin MacLeod, and Maclyn McCarty demonstrated that DNA was the chemical basis of heredity and the genetic cornerstone of life on Earth (Avery et al., 1944). Some 30 years later, Frederick Sanger, Steve Nicklen, and Alan Coulson developed the dideoxy termination sequencing reaction to allow accurate and rapid determination of the sequence of long stretches of DNA (Sanger et al., 1977). Another 30 years later, we find that automated techniques, novel sequencing approaches, and technological advancements are again transforming our vision of the distribution and diversity of organisms. We have sequenced a human genome, several other animal and plant genomes, and over 500 complete microbial genomes. Sequencing the environment was the next big challenge, and marine microbiologists rose to that challenge. Here we review the current state and future prospects for marine environmental genomics.

ROBERT A. EDWARDS *(raedwards@gmail.com) is Adjunct Assistant Professor, Department of Biology, San Diego State University, San Diego, CA, and Visiting Research Scientist, Mathematics and Computer Sciences Division, Argonne National Laboratory, Argonne, IL.*
ELIZABETH A. DINSDALE *is Adjunct Assistant Professor, Department of Biology, San Diego State University, San Diego, CA.*

THE FIRST OCEAN METAGENOMES

An early example of the way sequencing technology changed our view of marine microbial communities was the discovery and analysis of free-living archaea in the ocean's surface waters (DeLong, 1992; Fuhrman et al., 1992). Until ocean water was sampled using polymerase chain reaction (PCR) and fluorescent-based hybridization techniques, archaea had been considered specialists of extreme environments, including those with low pH, high temperature, high salinity, and limited or no oxygen. In the marine realm, archaea were thought to be restricted to the deep-sea vents, anoxic muds, and other limited locales. In contrast, the oxygenated, moderate-pH surface or deep waters were thought to harbor only bacteria (DeLong, 1992; Fuhrman et al., 1992). Because these free-living archaea remained recalcitrant to culturing, more genetic information was required to understand their functional contribution to the ecosystem. Large insert (approximately 40,000 base pairs or 40 kb) fosmid libraries were constructed and probed for the presence of archaea (Stein et al., 1996). The 16S rDNA gene, whose product is required for DNA transcription, was, and remains, the most authoritative determinant of the presence of bacteria or archaea in a sample. Because archaea accounted for less than 5% of the microbial cells in the oceans (though still numbering in the millions or more cells per milliliter of seawater), and because very few of the 40 kb fragments contained 16S rDNA genes, thousands of individual clones were screened before a single clone that contained an archaeal 16S rDNA gene was found (Stein et al., 1996). Approximately 2 kb portions of the 40-kb insert from the single archaeal clone were subcloned and sequenced, representing one of the first sequenced marine community genomes. These sequenced fragments revealed the true origin of this DNA fragment—from a *Crenarchaeota*—and provided insight into the evolution, ancestry, and metabolic potential of this organism (Stein et al., 1996).

Apart from identifying free-living archaea, subsequencing, fosmid-clone libraries can be used to determine the taxonomic extent of newly identified proteins, such as proteorhodopsins. A

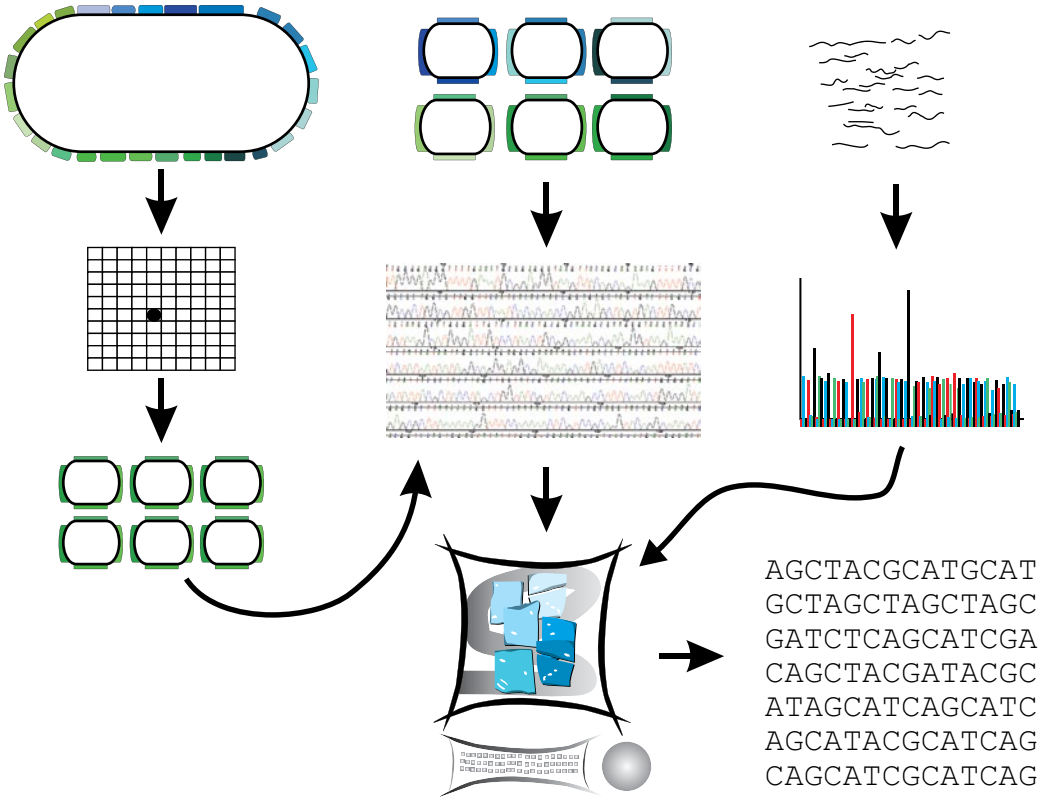

Figure 1. Metagenomes have been made in different ways. From left to right: First, large insert libraries were made, screened for a gene of interest, and that particular isolate was subcloned and sequenced. Second, random small insert libraries were sequenced using high-throughput Sanger Sequencing. Then, more recently, random uncloned fragments were sequenced using high-throughput pyrosequencing.

new type of rhodopsin, a purple pigment photoprotein that harvests biochemical energy from green light, was thought to be restricted to archaea. However, community sequencing approaches revealed this gene adjacent to a bacterial 16S rDNA gene (Béjà et al., 2000). Fosmid libraries have proved fruitful for identifying new sources of genetic information and remain the method of choice for isolating complete genes that perform biological functions of interest to the biotechnological community (Vergin et al., 1998; Robertson and Steer, 2004; Hårdeman and Sjöling, 2007). In addition, it has long been known that genes that perform related functions cluster together along the chromosome (Overbeek et al., 1999), and sequencing large contiguous DNA fragments contained on fosmids can yield complete pathways (DeLong et al., 2006). For example, this approach was used to identify the pathways whereby archaea use oxidized methane anaerobically (Hallam et al., 2004).

The application of high-throughput sequencing, exploiting technological advancements on Sanger's dideoxy sequencing approach (mainly through the industrialization of the processes), reduced the need to screen libraries for specific genes of interest before sequencing, and allowed for the direct sequencing of randomly chosen clones from environmental samples. Because viruses are ubiquitous, have much smaller genomes than other organisms, and are readily fractionated from bacteria, archaea, and eukarya, much of the pioneering work on random community genomics from marine environments was performed on phages—those viruses that infect bacteria (see Breitbart et al., this issue). The much heralded, and oft-debated Sargasso Sea random community genome publication represents a "line in the sand," demarking the entrance of marine microbiology into the high-throughput, post-genomic era (Venter et al., 2004). The Sargasso Sea study took an approach, now familiar from complete genome sequencing projects, that eschewed large insert libraries and selective subsequencing in favor of complete sequencing of smaller inserts (Figure 1). Theory held that with enough sequencing, whole genomes could be assembled from environments, and large inserts were no longer needed

for detailed understanding of the complexities of marine microbial life. Indeed, complete genomes were assembled from the individual reads—alas, they had a decidedly nonmarine origin (Falkowski and Vargas, 2004; DeLong, 2005; Mahenthiralingam et al., 2006). These data revolutionized our view of marine microbiology, altered our perception of how the data would be handled and analyzed in the future, and created a furor among biologists, as widely used databases overflowed with sequences simply labeled as hypothetical proteins from the Sargasso Sea (Tress et al., 2006).

An immediate observation from the Sargasso Sea sampling was an abundance of genes involved in photosynthesis. Rather than being from chlorophyll-based systems, many of these genes were rhodopsin-like photoreceptors (Venter et al., 2004). Furthermore, many of the photorhodopsin-like genes identified, 782 in total, were distinct from the proteins identified in earlier work (Béjà et al., 2000), suggesting that many more organisms in the ocean are capable of harvesting light than first imagined, and productivity estimates from satellite chlorophyll measurement may misjudge the amount of light being captured by marine microbes.

The Sargasso Sea environmental genome was a milestone in marine genome analyses, and it continues to be mined by researchers in a surprising number of areas, especially in biology and computer science. These data allow hypotheses to be generated, and tested (e.g., Rodriguez-Brito et al., 2006). However, the visage provided by the Sargasso Sea data set was dwarfed by the first release of the Global Ocean Sampling (GOS) expedition data set (Rusch et al., 2007; Yooseph et al., 2007), which contains the equivalent of about two human genomes, approximately 6.3×10^9 bp of sequences. The size of this data set hampers all but the most ardent computationalists from analyzing the data. However, several patterns are beginning to emerge from the initial publications and synthesis with other metagenomic data. First, in general, even with the deep sequence coverage provided by the data set, assembly of significantly long, contiguous regions of sequences (contigs) from the small fragments failed. Assembly of long contigs from previous environmental genome projects (apart from the contaminated Sargasso Sea sample) was aided by sampling low-complexity environments, such as acid mine drainage systems (Tyson et al., 2004), and using large insert libraries, like those described above (Béja, 2004). Second, similar species are found in widely geographically separated samples: bacteria similar to SAR11 and SAR86 were found in almost every sample (Rusch et al., 2007). This "marine-ness" of the samples has been reported elsewhere (Massana et al., 2000; Tringe et al., 2005; Angly et al., 2006). However, apart from the ubiquitous microbes, there were also differences in the samples from the pole to the equator, with tropical and temperate communities comprised of different organisms. Further, an increase in diversity towards the equator was found in the microbial communities present in both the GOS data (Rusch et al., 2007) and a survey of nine targeted oceanic regions (Pommier et al., 2007), which reflects trends seen in the ecology of macrobiota.

These two observations—sequences will not assemble but many bacteria are ubiquitous—appears contradictory. The third observation from the GOS data set hints at possible causes for the conflicting results—the high number of viral-like sequences within the bacterial samples. Large numbers of viral-like sequences were previously observed in marine bacterial/archaeal metagenomic libraries (DeLong et al., 2006). The viral signatures obviously provide flexibility in the genomes, as denoted by variation between sequence reads (Rusch et al., 2007), and also in

> The speed with which environmental genomics has impacted marine microbiology reinforces how far the field has come in increasing our understanding of the smallest, but perhaps most important, inhabitants of the ocean.

the large number of viral-like proteins (Yooseph et al., 2007). Previous studies have shown the dramatic amount of gene transfer that is likely to occur in the open ocean, approximately 100 transduction events per day per liter of water (Jiang and Paul, 1998). In addition to spreading genetic variation, microbial mortality by viruses may eliminate the most successful isolates as soon as they have reached appreciable numbers (the kill-the-winner hypothesis [Thingstad and Lignell, 1997]). Therefore, the viral component of the microbial community appears to hamper efforts of technologists to assemble complete microbial genomes from environmental samples (see Breitbart et al., this issue).

In terms of the distribution of microbes, is it true that "everything is everywhere and the environment selects" (Pommier et al., 2007, quoting Baas-Becking L.G.M. (1934) *Geobiologie of Inleiding Tot de Milieukunde*. W.P. Van Stockum & Zoon N.V., den Haag) or are there localized enrichments for specialist microbes? Environmental genomics studies on microbial use of organic matter indicate that generalist bacteria are capable of utilizing multiple carbon sources as they become available, suggesting that bacteria may exploit a wide range of environments (Mary Ann Moran, University of Georgia, *pers. comm.*, 2007). However, the Sargasso and GOS sequences were all collected from ocean surface waters, at depths ranging from 0.1 to 30 m. In contrast to covering a wide geographic range in one depth zone, a vertical transect (0–4000 m deep) taken at a single location reveals the changes in microbial communities that occur as light attenuates and pressure increases (DeLong et al., 2006). Low-light apparatus replaces the high-light photosynthetic apparatus before all photosynthesis is lost as light disappears, demonstrating environmental selection through specialization in light adaptation. Further adaptation appears to occur in the very deep-water samples, which contain large numbers of transposases, indicative of slow growth rates or the need to adapt to changing conditions, such as influxes of nutrients. However, deep waters are very geochemically and physically stable (DeLong et al., 2006) and, presumably, so are the microbial communities that inhabit them.

Metagenomic studies also target novel functions in different environments and, increasingly, statistical techniques are being used to discern the differences among environments (Tringe et al., 2005; Rodriguez-Brito et al., 2006). Whale carcasses that sink to the bottom of the ocean form ecological "islands" that undergo a prolonged breakdown (Smith and Baco, 2003). The environmental genomics study of three whale falls shows they harbor many fewer species than the open ocean. Although they might be considered an ideal location for finding enzymes that break down fats in cold water (e.g., cold-water esterases used in laundry detergents), none were reported in these metagenomes. Nonetheless, environmental genomics will likely be widely used in future gene discovery applications to harness natural biological processes for industrial applications (Li and Qin, 2005). Identifying changes in metabolic potential of microbial communities within various environments will help identify important areas of biogeochemical activity.

Technological shifts are yet again upending our view of marine microbiology. Industrialization of an alternative method of sequencing, called pyrosequencing, which does not rely on Sanger's dideoxy terminators, has emerged as a contender in environmental genomics studies (Margulies et al., 2005; Angly et al., 2006; Edwards et al., 2006; Prosser et al., 2007). The advantage of pyrosequencing is in the adaptations that enable hundreds of thousands of sequences to be interrogated simultaneously and cheaply. Recent studies on marine samples suggest that there are several orders of magnitude more species than previously imagined in the ocean (Sogin et al., 2006), and they are complemented by major groups that were thought not to occur in the ocean, such as single-stranded viral sequences (Angly et al., 2006). Therefore, in addition to the commonly sampled organisms that are found in the 16S rDNA libraries and shotgun sequences, it is becoming increasingly apparent that the ocean harbors a "rare biosphere" that may be the source of the genetic material for the variation observed.

> ...it is clear that the more we learn, the more we realize how much we don't yet know.

The advent of cheap, fast sequencing through pryosequencing offers the ability to use metagenomics to answer important questions in marine ecology and geochemistry rather than just provide generalized observations. For example, disease has been steadily increasing in marine environments, and many important commercial species, such as oysters and mussels, are being affected and lost for commercial purposes (Barber, 2004; Harvell et al., 2004). Coral reefs are particularly vulnerable to disease, resulting in both the loss of individual species and altered community structure and function (see Rosenberg et al., this issue). Recent studies of microbial communities on coral reefs using metagenomics found dramatic effects caused by adjacent human populations. As the influence of human activity increases, the microbial communities shift from a balanced heterotrophic/autotrophic mix toward an overwhelmingly heterotrophic population that besieges the corals, according to recent work of author Dinsdale and 13 colleagues. This group conducted one of the first studies to measure contributions of each trophic level, from viruses through microbes to corals, algae, fishes, and sharks, to the ecosystem (additional data from Stuart Sandin, Scripps Institution of Oceanography, *pers. comm.*, 2007).

The coral reef studies highlight an outstanding problem in environmental genomics: the association of "metadata" with genomic data. It is critical to identify not only where these sequences are from but also what is happening around them (i.e., obtain a richer set of metadata). These data are essential for truly understanding the role of microbes in the environment and will lead to new ways of exploring genome sequences (Lombardot et al., 2006; Field et al., in press).

The last 60-plus years gave us the identification of DNA as the genetic material, the means to sequence that material, and now the technological leaps to sequence DNA cheaply and efficiently. The speed with which environmental genomics has impacted marine microbiology reinforces how far the field has come in increasing our understanding of the smallest, but perhaps most important, inhabitants of the ocean. However, it is clear that the more we learn, the more we realize how much we don't yet know. Technological advances over the next few years will include single-cell sequencing, long reads from single DNA molecules, and an explosion of synthetic DNA approaches to reconstructing sequences in the ocean. Together, these tools will help answer some of the remaining questions (see Box 1).

> Technological advances over the next few years...will help answer some of the remaining questions.

ACKNOWLEDGEMENTS

We thank Mya Breitbart, Ed DeLong, and Mary Ann Moran for critical comments on this manuscript.

REFERENCES

Angly, F.E., B. Felts, M. Breitbart, P. Salamon, R.A. Edwards, C. Carlson, A.M. Chan, M. Haynes, S. Kelley, H. Liu, and others. 2006. The marine viromes of four oceanic regions. *PLoS Biology* 4.

Avery, O.T., C.M. MacLeod, and M. McCarty.

BOX 1. QUESTIONS FOR THE FUTURE

THE SCALE OF VARIATION

- Temporal variation
- Spatial variation

GROUND TRUTH

- Identifying contamination
- Identifying junk sequence

METADATA

- What to collect, where to collect it.

ASSEMBLY

- Is it a realistic goal?

EFFECTS OF FUTURE TECHNOLOGICAL SHIFTS

- Single-organism sequencing (e.g., with amplification)
- Long-read sequencing (e.g., 500 kb per read)

1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine* 79:137–158.

Barber, B.J. 2004. Neoplastic diseases of commercially important marine bivalves. *Aquatic Living Resources* 17:449–466.

Béjà, O. 2004. To BAC or not to BAC: Marine ecogenomics. *Current Opinion in Biotechnology* 15:187–190.

Béjà, O., L. Aravind, E.V. Koonin, M.T. Suzuki, A. Hadd, L.P. Nguyen, S.B. Jovanovich, C.M. Gates, R.A. Feldman, J.L. Spudich, and others. 2000. Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* 289:1,902–1,906.

DeLong, E.F. 1992. Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences of the United States of America* 89:5,685–5,689.

DeLong, E.F. 2005. Microbial community genomics in the ocean. *Nature Reviews Microbiology* 3:459–469.

DeLong, E.F., C.M. Preston, T. Mincer, V. Rich, S.J. Hallam, N.U. Frigaard, A. Martinez, M.B. Sullivan, R. Edwards, B.R. Brito, and others. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311:496–503.

Edwards, R.A., B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D.M. Peterson, M.O. Saar, S. Alexander, E.C. Alexander Jr., and F. Rohwer. 2006. Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics* 7:57.

Falkowski, P.G., and C. Vargas. 2004. Shotgun sequencing in the sea: A blast from the past? *Science* 304:58–60.

Field, D., G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M.J. Allen, M. Ashburner, and others. In press. Towards a richer description of our complete collection of genomes and metagenomes: The "Minimum Information about a Genome Sequence" (MIGS) specification. *Nature Biotechnology*.

Fuhrman, J.A., K. McCallum, and A.A. Davis. 1992. Novel major archaebacterial group from marine plankton. *Nature* 356:148–149.

Hallam, S.J., N. Putnam, C.M. Preston, J.C. Detter, D. Rokhsar, P.M. Richardson, and E.F. DeLong. 2004. Reverse methanogenesis: Testing the hypothesis with environmental genomics. *Science* 305:1,457–1,462.

Hårdeman, F., and S. Sjöling. 2007. Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. *FEMS Microbiology Ecology* 59:524–534.

Harvell, D., R. Aronson, N. Baron, J. Connell, A. Dobson, S. Ellner, L. Gerber, K. Kim, A. Kuris, H. McCallum, and others. 2004. The rising tide of ocean diseases: Unsolved problems and research priorities. *Frontiers in Ecology and the Environment* 2:375–382.

Jiang, S.C., and J.H. Paul. 1998. Gene transfer by transduction in the marine environment. *Applied and Environmental Microbiology* 64:2,780–2,787.

Li, X., and L. Qin. 2005. Metagenomics-based drug discovery and marine microbial diversity. *Trends in Biotechnology* 23:539–543.

Lombardot, T., R. Kottmann, H. Pfeffer, M. Richter, H. Teeling, C. Quast, and F.O. Glöckner. 2006. Megx.net—database resources for marine ecological genomics. *Nucleic Acids Research* 34: D390-393, doi: 10.1093/nar/gkj070.

Mahenthiralingam, E., A. Baldwin, P. Drevinek, E. Vanlaere, P. Vandamme, J.J. Lipuma, and C.G. Dowson. 2006. Multilocus sequence typing breathes life into a microbial metagenome. *PLoS ONE* 1:e17.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, and others. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380

Massana, R., E.F. DeLong, and C. Pedros-Alio. 2000. A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. *Applied and Environmental Microbiology* 66:1,777–1,787.

Overbeek, R., M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev. 1999. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* 96:2,896–2,901.

Pommier, T., B. Canback, L. Riemann, K.H. Bostrom, K. Simu, P. Lundberg, A. Tunlid, and A. Hagstrom. 2007. Global patterns of diversity and community structure in marine bacterioplankton. *Molecular Ecology* 16:867–880.

Prosser, J.I., B.J. Bohannan, T.P. Curtis, R.J. Ellis, M.K. Firestone, R.P. Freckleton, J.L. Green, L.E. Green, K. Killham, J.J. Lennon, and others. 2007. The role of ecological theory in microbial ecology. *Nature Reviews Microbiology* 5:384–392.

Robertson, D.E., and B.A. Steer. 2004. Recent progress in biocatalyst discovery and optimization. *Current Opinion in Chemical Biology* 8:141–149.

Rodriguez-Brito, B., F. Rohwer, and R. Edwards. 2006. An application of statistics to comparative metagenomics. *BMC Bioinformatics* 7:162.

Rusch, D.B., A.L. Halpern, G. Sutton, K.B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J.A. Eisen, J.M. Hoffman, K. Remington, and others. 2007. The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* 5:e77.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74:5,463–5,467.

Smith, C.R., and A.R. Baco. 2003. Ecology of whale falls at the deep-sea floor. *Oceanography and Marine Biology: An Annual Review* 41:311–354.

Sogin, M.L., H.G. Morrison, J.A. Huber, D.M. Welch, S.M. Huse, P.R. Neal, J.M. Arrieta, and G.J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proceedings of the National Academy of Sciences of the United States of America* 103:12,115–12,120.

Stein, J.L., T.L. Marsh, K.Y. Wu, H. Shizuya, and E.F. DeLong. 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.

Thingstad, T.F., and R. Lignell. 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology* 13:19–27.

Tress, M.L., D. Cozzetto, A. Tramontano, and A. Valencia. 2006. An analysis of the Sargasso Sea resource and the consequences for database composition. *BMC Bioinformatics* 7:213.

Tringe, S.G., C. von Mering, A. Kobayashi, A.A. Salamov, K. Chen, H.W. Chang, M. Podar, J.M. Short, E.J. Mathur, J.C. Detter, P. Bork, P. Hugenholtz, and E.M. Rubin. 2005. Comparative metagenomics of microbial communities. *Science* 308:554–557.

Tyson, G.W., J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, and others. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74.

Vergin, K.L., E. Urbach, J.L. Stein, E.F. DeLong, B.D. Lanoil, and S.J. Giovannoni. 1998. Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. *Applied and Environmental Microbiology* 64:3,075–3,078.

Yooseph, S., G. Sutton, D.B. Rusch, A.L. Halpern, S.J. Williamson, K. Remington, J.A. Eisen, K.B. Heidelberg, G. Manning, W. Li, and others. 2007. The *Sorcerer II* Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biology* 5:e16.