

# Marine Environmental Genomics: New Secrets from a Mysterious Ocean

## AUTHORS

Karla B. Heidelberg  
J. Craig Venter Institute

John F. Heidelberg  
Institute for Genomic Research and  
J. Craig Venter Institute

Today we are witnessing an explosion in interest in marine microbial communities. Microscopic marine organisms, including communities of phytoplankton, Bacteria, Archaea, protists and viruses, play critical roles in the biosphere by participating in virtually all of the Earth's biogeochemical cycles and thereby affecting geology, hydrology, and even possibly global climate change. These prolific and diverse microbial communities, while not seen by casual observers, by some estimates account for more than 90 percent of the ocean's biomass and 98 percent of the primary production in the marine environment (e.g. Whitman et al., 1998). Therefore, for humans to truly begin to understand the impact of our activities on our environment, and to being to develop new tools to combat negative effects, it is important for us to understand both diversity and function of the marine microbial assemblages.

Traditional 'black box' microbial ecology set the foundation for understanding ocean function and taught us that unseen biology is of critical importance in the ocean. Through these studies we learned that microbes were at the core of virtually all the biogeochemical carbon (e.g. Carlson et al., 2001) and nutrient (e.g. Redfield, 1958) cycles. For example, the global dissolved organic carbon pool is estimated to be approximately 700 Pg C, a value comparable to the mass of inorganic C in the atmosphere. Interaction of microbes within the dissolved organic pool could strongly impact the balance between oceanic and atmospheric carbon dioxide. However,

## ABSTRACT

Ocean microbes play critical roles in ecosystem dynamics and biogeochemical cycles. For a number of reasons, these organisms have been hard to study; among other characteristics, they are difficult or maybe impossible to culture. The recent application of cultivation-independent genomic techniques to study bacterial communities has begun to fundamentally change our views of microbial ecology and function. These approaches are providing more comprehensive insights into the structure and function of natural assemblages of microbial populations. Genomics-based technologies are revealing previously unknown groups of microorganisms and novel metabolic pathways, leading to a deeper appreciation of the fundamental genetic and potential functional diversity of ocean microbes. When evaluated in the context of observed ecosystem functions, we can begin to understand the complex interactions of individuals and populations with their physical and chemical environment. The continuing study and interpretation of the community genomic data will require a close synergy among interdisciplinary researchers working throughout the world's ocean basins. Community genomic approaches have and will become a powerful tool for understanding marine microbial ecology in the future.

powerful as these studies are, they do not provide information on which microbes are responsible for specific biogeochemical processes.

Additional studies were designed to evaluate ocean function with the use of culturing technologies. However, a significant challenge to our ability to study and understand these microorganisms is that the vast majority are not easily cultured on typical growth media, the traditional approach in much of microbiology dating back to the early days of medical microbiology and Koch's principles. Culturing methods were thought to be the necessary first step towards evaluating microbial diversity and function, but we now know that studying only microbial populations that can be grown in culture provides very little information on natural diversity or environmental function (e.g. Beja et al., 2002a). Consequently, over the past decade, scientists have developed molecular tools to explore environmental diversity without relying on culturing technologies. The most commonly used culture-independent method relies on comparisons of homologous genes between organisms by using techniques such as polymerase chain reaction (PCR) products typically targeting

phylogenetic genes. Amplification and sequencing of the conserved regions of the 16S ribosomal RNA (rRNA) gene (part of the protein synthesis machinery found in all living cells), can provide the identity of the specific organisms in the sample. Initial application of this methodology resulted in an explosion of information, providing for the first time a mechanism to evaluate microbial diversity for the > 99% of the microbial population that could not be grown in the laboratory (e.g. DeLong et al., 1989; Giovannoni et al.; 1990; Pace, 1996, 1997).

Despite the impact of these rRNA based surveys, a phylogenetic identification of a microorganism based solely on an rRNA sequence does not allow inference of physiology, biochemistry, or ecological significance. Therefore, the specific biological properties of abundant uncultured microorganisms remain almost entirely unknown. Another limitation of the 16S rRNA sequencing technique is that it does not distinguish between strains of the same species that may fill significantly different biological niches and habitats. For example, *Escherichia coli*, which persist in the human gut in a mutualistic relationship, may not be dis-

tinguished from a closely related strain that has acquired various toxins and is a deadly human pathogen.

Recent advances in genomic science have helped scientists to study cultured microorganisms with much greater precision than ever before. Through the merging of DNA sequencing technologies and the use of newly developed computational methodologies a complete set of genetic blueprints of the organism (the genomes) can now be ascertained in a very short period of time. Since the first bacterial genome was sequenced a decade ago, (Fleischmann et al., 1995), the number of fully sequenced deposited prokaryotic genomes into public databases has grown tremendously (Figure 1). Genome sequencing of microorganisms was initially prioritized to study cultured pathogens, but recently there has been an expansion and focus on environmentally important marine microbes. As expected, having the completed manually curated genomes of key ocean organism genomes—e.g. a diatom, *Thalassiosira pseudonana* (Armbrust et al., 2004), two het-

erotrophic bacteria (Moran et al., 2004; Giovannoni et al., 2005), and several cyanobacteria (e.g. Dufresne et al., 2003; Palenik et al., 2003; Rocap et al., 2003)—are providing exciting information on the detailed physiological and genetic controls of photosynthesis, and the cycling of carbon and nitrogen in the world's oceans. With recent efforts funded by the Gordon and Betty Moore Foundation's Marine Microbial sequencing program (see Marine Microbial Initiative, <http://www.moore.org/>), the possible addition of up to approximately 150 genomes from cultured marine prokaryotes will greatly expand the existing marine genome reference library.

However, perhaps the real potential of genomics may be in the newly developed methods applied to study uncultured organisms—both prokaryotes and viruses. Several key studies in 2004 (Tysen et al., Venter et al., Tringe et al.) show that much could be learned about entire communities of uncultured organisms by using high-throughput, shotgun DNA sequencing technologies. Rather than

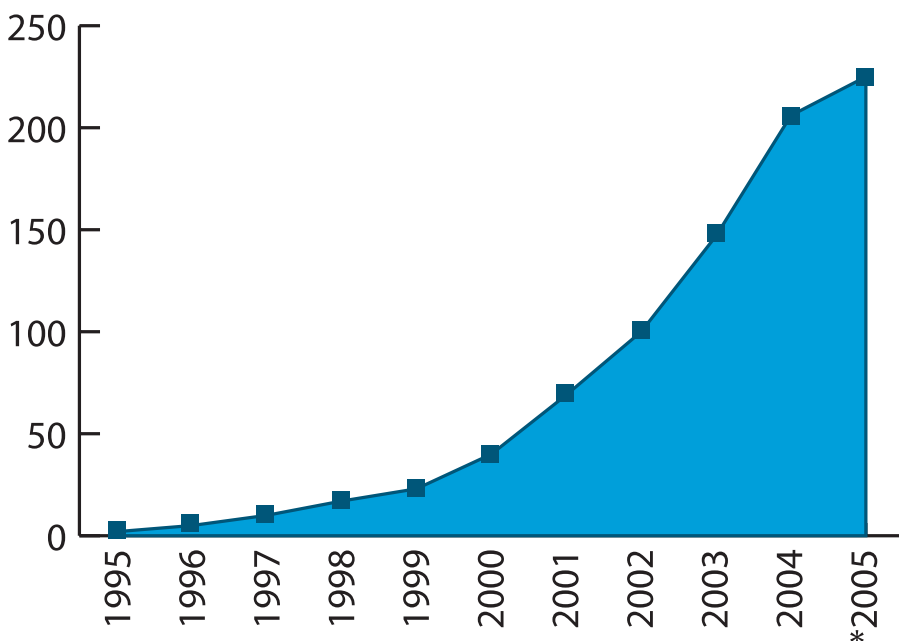
trying to isolate a single organism, the total community DNA is isolated *en mass*, the DNA is extracted, and this community of DNA serves as the template for sequencing. Once extracted, the DNA can be shotgun sequenced as short insert libraries or analyzed with targeted large insert libraries (Figure 2). These are powerful meta- or community genomics approaches and allow for the gene complement of the entire community to be sequenced to yield information on the potential biological functions of the genes to be catalogued at a very rapid pace. Analysis of small insert clones from the Sargasso Sea (Venter et al., 2004), for example, yielded over one million previously unknown genes, including almost 800 rhodopsins (a light absorbing molecule). In addition to entire genome sequencing, studies using large insert libraries to evaluate detailed parts of organism's genomes have helped make important new discoveries, such as finding of a new form of phototrophy (Béjà et al., 2002b; reviewed by Karl, 2002). It is clear that microbial communities have an extraordinary diverse suite of mechanisms and pathways that are still yet to be discovered.

As often is the case with new technological advances, genomic technologies are being used by scientists based, for the most part, in the developed world. It is tempting, for reasons of convenience, to maintain research programs in well-studied waters of one's own country. Ocean communities, however, function irrespective of political boundaries, and globally, these new technologies have so much potential for informing marine management, sustainability, and monitoring programs. However, scientists conducting basic science research internationally face challenging obstacles due to disparate access and permitting requirements, even with well established mutually beneficial collaborations with in-country scientists. Researchers proposing to work internationally need to be aware of International Conventions such as the United National Convention on Biological Diversity, CBD (and related agreements; <http://www.biodiv.org/default.shtml>) and the United Nations Law of the Sea section governing marine science research. The CBD, in particular, stipulates that each nation has sovereign rights to its genetic and biochemical resources, and strives

### Figure 1

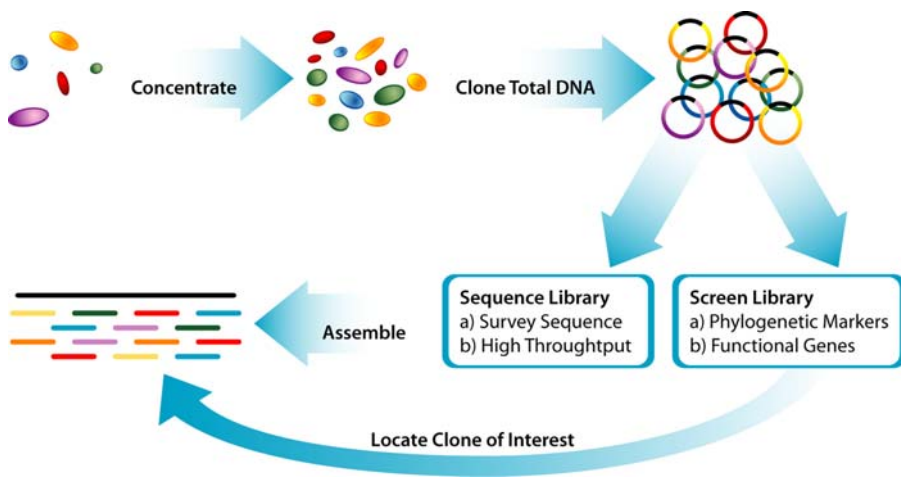
Publicly available completed genomes. Data reported by publication date, or if not published, the date that genome data was deposited into the Comprehensive Microbial Database (CMR; <http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi>). Data for 2005 as of May. Data presented represent a total of 226 genomes (199 Bacteria and 20 Archaea and 3 Viruses).

## Completed Prokaryotic Genomes



## Figure 2

Schematic representation of a typical marine metagenomic study. The total microbial population is concentrated (often after pre-filtration to remove larger eukaryotic organisms), and the total DNA extracted. The DNA can then be cloned into large insert vectors (Fosmids or BACs) and screened for specific genes prior to sequencing or the DNA can be cloned into small insert vectors and subjected to high throughput sequencing for more generalized exploration.



to ensure that there is prior informed consent, equitable benefit sharing and promotion of sustainable sampling programs. Many countries are currently developing or have in place additional legislation governing access to genetic resources as well. The challenge for scientists conducting basic research is that most regulations have been primarily developed for for-profit research programs—where benefit sharing arrangements can be clearly defined. Further, approaches to benefit sharing components such as education, capacity building and open collaborations are relatively new and have yet to be rigorously evaluated for their success with respect to the quality of science and the mutual gain for the host country. While accepting the requirements for working internationally, the research community should continue to assess other yet undefined mechanisms that may improve the existing systems for promoting basic science research to ensure that descriptions of ocean processes are representative of many biomes of the globe.

## Prospectus for the Future

Programs such as the Sloan Foundation's International Census of marine microbes (<http://icomm.mbl.edu>) are bringing international focus on marine microbial assemblages. Further, marine microbial genomic programs

are providing once unimaginable detail about communities of organisms—and the communities of genes existing together. The exploding databases hold enormous promise for refining concepts of species biodiversity, coevolution of life and ocean function in the coming years. Global microbial diversity far exceeds our previous estimations and, further, the concepts we have of 'microbial species' may be inadequate to describe this diversity.

However, a more complete understanding of microbial diversity and the environmental processes that microbes control will require much more than a biotic inventory of genes or species. Metagenome sequence data is only the first step in advancing our understanding of the microbial-mediated processes in the oceans. There remains a strong need for the merging of the traditional methodologies in microbial ecology, evolution, biogeochemistry, Earth and remote sensing sciences with the development of newer technologies to evaluate microbial function under differing environmental conditions (e.g. Zehr and Ward, 2002; Steward et al., 2004; DeLong, 2005).

The challenge of understanding the complex microbial processes on the scale now possible demands the coordinated application of many, often previously disparate disciplines to develop new concepts and a new generation of computational, laboratory and *in situ* method-

ologies. For example, one may have once hoped that metagenomes would have enabled scientists to paint a reasonably clear picture of the evolutionary past and biogeochemical potential of microorganisms. However, it is more the case that such datasets are still revealing more about how much we still do not know, uncovering unexpected levels of prokaryotic and viral biocomplexity, and raising more questions than answers for the process of gene expression. Each individual sequence is no longer just a piece of a genome. It is a piece of an entire functioning biological community.

Through future efforts of structural and functional annotation, gene names and ontology classification can be assigned to sampled proteins, and it is this information that will be valuable to experimental researchers studying specific metabolic and chemical processes that are microbially mediated in the oceans. Environmental metagenomics data provides the information needed for detailed analysis of individual processes or for designing new approaches to examine the potential interplay of perhaps thousands of species present and functioning at a point in space and time.

Intensive microbial genetic/biodiversity surveys and *in situ* studies, covering the full range of environmental conditions and geological and evolutionary histories, will help to determine the roles and functions of genes and organisms. New cultivation technologies for marine microbes (e.g. Connon and Giovannoni, 2002; Zengler et al., 2002) are providing key organisms for experimentation and tools for interpreting the results of metagenomic investigations. The future also brings a rapidly expanding genomic focus on the viral fraction of marine systems and their effects on microbial populations (reviewed by Edwards and Rohwer, 2005).

Just as high-throughput DNA sequencing has brought big changes in genomics, other high-throughput technologies will do the same for areas such as gene expression (microarrays), metabolomics (NMR, mass spectrometry), environmental proteomics (mass spectrometry, protein chips) and species characterizations (flow cytometry). Mathematical models will be needed to capture, simulate, and understand the cooperative behaviors that arise from complex processes of biology within

the environment. These types of complimentary studies will help to provide a context for the individual cellular components in time and space and will help us decipher the structure and function of these complex, non-linear systems.

Rapid increases in information-rich metagenomic studies will also require new approaches to data storage, analysis, synthesis, and access. Existing depositories for genomic data have little ability to accommodate locational, physical and biological environmental metadata to better identify sources and important environmental parameters associated with the data. There are also conceptual challenges to be overcome as the environmental metagenomic sequence data are very different from what has so far been used in the field of genomics. We are quickly amassing an enormous amount of detail on microbial biodiversity and gene complement of both cultured and uncultured microorganisms existing together in environmental assemblages. The data is often comprised of large numbers of short DNA fragments for which the species of origin is unknown. New methods are needed for predicting from which species (or clade) a DNA fragment originates as well as for comparing entire ecosystems rather than individual genomes. Since relatively few scientists have the capacity and/or computational power to utilize the massive datasets, new systems and user friendly Web-based tools need to be developed to accommodate the rapidly expanding databases (sequences, genes, gene families, genomes and associated environmental data) to allow for better utilization of deposited data.

Although it is difficult to predict the new discoveries in the immediate future, one thing is certain: the recent advances in genome related sciences will change forever how we view and study our oceans. These tools, when integrated into existing bodies and knowledge and when used in collaborative studies, will continue to provide innumerable opportunities for both basic and applied research well into the future. At the turn of the century, we had no idea that we would uncover so much information about the microbial biodiversity and gene complement or achieve so many novel insights. As the datasets grow, new scales of

variability will be described. How public domain databases will deal with the new environmental data remains to be seen. This is indeed an exciting and challenging time for studying our oceans.

## References

- Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., Hellsten, U., Hildebrand, M., Jenkins, B. D., Jurka, J., Kapitonov, V. V., Kröger, N., Lau, W. Y., Lane, T. W., Larimer, F. W., Lippmeier, J. C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M. S., Palenik, B., Pazour, G. J., Richardson, P. M., Rynearson, T. A., Saito, M. A., Schwartz, D. C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F. P. and Rokhsar, D. S.** 2004. The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution and metabolism. Science 306(5693):79-86.
- Béjà, O., Suzuki, M. T., Heidelberg, J. F., Nelson, W. C., Preston, C. M., Hamada, T., Eisen, J. A., Fraser, C. M. and DeLong, E. F.** 2002a. Unsuspected diversity among marine aerobic anoxygenic phototrophs. Nature 415(6872):630-633.
- Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., Jovanovich, S. B., Gates, C. M., Feldman, R. A., Spudich, J. L., Spudich, E. N. and DeLong, E. F.** 2002b. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science. 289(5486):1902-1906.
- Carlson, C.A., Bates, N.R., Hansell, D.A. and Steinberg, D.K.** 2001. Nutrient Cycling: The Carbon Cycle. In: Encyclopedia of Ocean Science. eds. J. Steele, S. Thorpe, and K. Turekian. pp. 390-400. London: Academic Press.
- Connon, S. A. and Giovannoni, S. J.** 2002. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. Appl Environm Microbiol. 68(8):3878-3885.
- DeLong, E. F.** 2005. Microbial community genomics in the ocean. Nature. 3:459-469.
- DeLong, E. F., Wickham, G. S. and Pace, N. R.** 1989. Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. Science. 243:1360-1363.
- Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I.M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I.B., Scanlan, D.J., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., Wolf, Y. I., Hess, W. R.** 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci. USA. 100(17):10020-10025.
- Edwards, R. A. and Rohwer F.** 2005. Viral Metagenomics. Nature Reviews in Microbiology. 3(6):504-510.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocyne, J. D., Scott, J., Shirley, R., Liu, L-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald L. A., Small, K. V., Fraser, C. M., Smith, H. O. and Venter, J. C.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223):496-512
- Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. and Field, K. G.** 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature. 345:60-63.
- Giovannoni, S. J., Tripp, H. J., Givan S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappé, M. S., Short, J. M., Carrington, J. C. and Mathur, E. J.** 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science. 309(5738):1242-1245.

- Karl, D. M.** 2002. Hidden in a Sea of Microbes. Nature. 415(6872):590-591.
- Moran, M. A., Buchan, A., Gonzalez, J. M., Heidelberg, J. F., Whitman, W. B., Kiene R. P., Henriksen, J. R., King, G. M., Belas, R., Fuqua, C., Brinkac, L., Lewis, M., Johri, S., Weaver, B., Pai, G., Eisen, J.A., Rahe, E., Sheldon, W. M., Ye, W., Miller, T. R., Carlton, J., Rasko, D. A., Paulsen, I. T., Ren, Q., Daugherty, S. C., Deboy, R.T., Dodson, R. J., Durkin, A. S., Madupu, R., Nelson, W. C., Sullivan, S. A., Rosovitz, M. J., Haft, D. H., Selengut, J. and Ward, N.** 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature. 432(7019):910-913.
- Pace, N. R.** 1996. New perspective on the natural microbial world: molecular microbial ecology. Amer Soc Microbial News. 62:463-467.
- Pace, N. R.** 1997. A molecular view of microbial diversity and the biosphere. Science. 276:734-740.
- Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., McCarren, J., Paulsen, I., Dufresne, A., Partensky, F., Webb, E. A. and Waterbury, J.** 2003. The genome of a motile marine *Synechococcus*. Nature. 424(6952):1037-42.
- Redfield, A. C.** 1958. The biological control of chemical factors in the environment. American Scientist. 46:205-221.
- Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. E., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Tolonen, A., Webb, E. A., Zinser, E. and S.W. Chisholm.** 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424(6952):1042-1047.
- Steward, G. F., Jenkins, B. D., Ward, B.B. and Zehr, J. P.** 2004. Development and testing of a DNA microarray to assess nitrogen (nifH) gene diversity. Appl Environ Microbiol. 70(3):1455-1465.
- Tringe, S. G., von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., Detter, J. C., Bork, P., Hugenholtz, P. and Rubin, E. M.** 2005. Comparative metagenomics of microbial communities. Science. 308(5721):554-557.
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. and Banfield, J. F.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 428(6978):37-43. .
- Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., Smith, H. O.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 304(5667):66-74.
- Whitman, W. B., Coleman, D. C. and Wiebe, W. J.** 1998. Prokaryotes: the unseen majority. Proc Natl Acad Sci. USA. 95(12):6578-83.
- Zehr, J. P. and Ward, B. B.** 2002. Nitrogen cycling in the ocean: new perspectives on process paradigms. Appl Environ Microbiol. 68(3):1015-1024.
- Zengler, K., Toledo, T., Rappé, M., Elkins, J., Mathur, E. J., Short, J. M. and Keller, M.** 2002. Cultivating the uncultured. Proc Natl Acad Sci. 99(24):15681-15686.