Eco-DAS VIII
Symposium Proceedings

Eco-DAS VIII Chapter 2, 2010, 17-29

A hitchhiker's guide to the new molecular toolbox for ecologists

Chris L. Dupont[1], Dreux Chappell[2], Ramiro Logares[3], and Maria Vila-Costa[4]*
[1]Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, CA 92121
[2]MIT-WHOI Joint Program in Oceanography and Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543
[3]Evolutionary Biology Centre, Limnology, Uppsala University, Uppsala, Sweden
[4]Department of Marine Sciences, University of Georgia, Athens, Georgia 30602, USA and Group of Limnology-Department of Continental Ecology. Centre d'Estudis Avançats de Blanes-CSIC. Accés Cala Sant Francesc, 14. 17300 Blanes, Catalunya,Spain

Abstract

Thirty years ago, marine microbes were described by crude morphology and the ability to grow on different carbon sources. Our understanding of their ecological role in aquatic environments was murky at best. Since then, the development of new molecular methods facilitated by DNA sequencing resulted in a revolution in the field of microbial molecular ecology and evolution. Plummeting sequencing costs and the resulting massive flux of data introduced novel challenges to marine microbiologists, in particular, and the infrastructure of science in general. In a cycle of positive feedbacks, a wide array of novel molecular and bioinformatic tools have been developed, addressing these challenges and allowing microbiologists to investigate subjects that previously stymied the field. Additionally, these advances fostered new connections between previously disparate disciplines. Here we provide a summary of the challenges of the new molecular toolkit, a history of the molecular revolution in microbial ecology, and a glimpse into the future. Finally, for the interested hitchhikers, we present a theoretical approach to integrating the new molecular toolkit into any ecological research program.

"But the reason I call myself by my childhood name is to remind myself that a scientist must also be absolutely like a child. If he sees a thing, he must say that he sees it, whether it was what he thought he was going to see or not. See first, think later, then test. But always see first. Otherwise you will only see what you were expecting... So, the other reason I call myself Wonko the Sane is so that people will think I am a fool. That allows me to say what I see when I see it. You can't possibly be a scientist if you mind people thinking that you are a fool."
—Wonko the Sane

"Let's be straight here. If we (scientists) find something we can't understand we like to call it something you can't understand, or indeed pronounce.
—Wonko's science colleague
(Douglas Adams, *So Long, and Thanks for All the Fish*)

Glossary

Assembly. The computational process of taking individual shotgun sequencing reads and, through overlapping alignments, putting them back together to create the original genome that was the source.

Barcoding. In this case, barcoding refers to the practice of adding known short ligator sequences to a pool of DNA before sequencing. This allows multiple samples to be sequenced in multiplex fashion with bioinformatic separation of the samples afterward.

Bioinformatics. the study of the nature and organization of biological information, incorporating fundamental biology, mathematics, statistics, and computer programming

Cloning. Technique of producing clones (identical organisms) that contain pieces of foreign DNA. Clones are obtained by fragmentation of the DNA, insertion of all fragments into a suitable vector, usually carried into *E. coli* and propagated by growing the bacteria. The common vectors used are artificial plasmids (naturally occurring, circular, extrachromosomal DNA molecules) or phage (a bacterial virus). DNA inserts are usually few kilobase pairs in size.

DNA clone library (or genomic library). A set of clones that collectively contain all or a fraction of the DNA from a single organism or community.

Expressed Sequence Tag. A portion of a cDNA sequence that

*Corresponding author: E-mail: cdupont@jcvi.org, 858 200 1886

Publication was supported by NSF award OCE0812838 to P.F. Kemp
ISBN: 978-0-9845591-1-4, DOI: 10.4319/ecodas.2010.978-0-9845591-1-4.17

was transcribed from mRNA. Sequencing of ESTs is one of the fundamental methods for transcriptomics and metatranscriptomics.

Functional genomics. Addresses global issues of transcription, translation, post transcriptional regulation, and post translational regulation. An example incorporating transcription and post-transcriptional regulation would involve examining the abundance of mRNAs and small RNAs that are activated during major metabolic shifts (as from growth under aerobic to growth under anaerobic conditions) or during embryogenesis and development of organisms.

Genome. The hereditary information of an organism encoded in its DNA (or, for some viruses, RNA). A Metagenome is the collective set of genomes within a community.

Genomics. The study of the genomes and their contents.

Gbp, Mbp, and Kbp: Giga, Mega, and Kilo base pair. Unit of measurement for DNA, equal to 1 billion, 1 million, and 1 thousand complimentary pairs of nucleotides, respectively.

Metabolomics. The study of all low-molecular-weight cellular constituents (such as metabolic intermediates, hormones and other signaling molecules, and secondary metabolites) of a tissue, organism, or community.

mRNA, Messenger RNA. These are products of gene transcription, and in Eukaryotes, splicing at intron-exon boundaries. mRNA is the central step in the central dogma (Box 1).

Multiplex sequencing. Refers to the process of pooling multiple samples into one optical lane of a sequencing run. The samples are then separated according the presence of barcode sequences added during library preparation.

Operon. A piece of DNA that includes a set of adjacent genes that are transcribed polycistronically,that is as one piece of mRNA. Often the coded proteins all operate in one cellular process.

Paired end sequencing. Also known as pairwise end sequencing or double-barreled shotgun sequencings. Here, a sample of DNA is sheered and size selected. Then, through methods that are specific to each sequencing technology, the ends of both the forward and reverse strand are sequenced. With knowledge of the relative distance between each read, paired end sequencing greatly eases the assembly of sequencing data.

PCR (Polymerase Chain Reaction). Amplification of a piece of DNA to millions of copies.

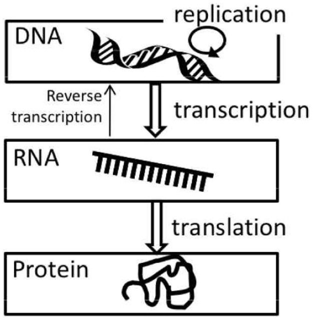
Proteome. All of the proteins translated in an organism or tissue for a differentiated organism at a point in time. Proteomics involve a survey via mass spectrometry or protein microarrays of all the proteins expressed by a certain cell or tissue under specified conditions. By extension, metaproteomics is a survey of the proteins synthesized by a community.

Shotgun sequencing. A process where DNA is sheered into smaller fragments, randomly sequenced, and then assembled using computational methods. The first complete microbial genomes were shotgun sequenced, and the method was integral to the completion of the human genome.

Transcriptome. The entire set of mRNA transcribed from DNA in a particular cell or tissue under defined conditions. The goal of Transcriptomics and Metatranscriptomics is to quantify the abundance of different mRNAs via sequencing or microarrays.

Box 1. Central dogma in molecular biology: DNA makes RNA makes protein.

The "central dogma of molecular biology," enunciated by Francis Crick in 1958, outline the residue-by-residue transfer of genetic information from DNA to proteins.



DNA contains the codes of genetic information. Replication of itself, base by base, ensures that the genetic code stays intact in the living cells (DNA replication). In addition, the expression of this genetic material is transmitted to an intermediate molecule, messenger RNA (transcription) that is translated into a sequence of amino acids (translation), which ultimately makes up a protein. Only a small percentage of the genes undergo transcription at a given time responding to the metabolic conditions of the cell. Much of DNA that does not encode proteins is now known to encode various types of functional RNA. Gene regulation is a tight process in both prokaryotic and eukaryotic cells, and it can occur at each step of the transfer chain.

The tsunami of DNA sequencing and the new challenges

During the last 20 years, the field of molecular ecology and evolution has exploded, with much of the credit due to the falling costs of DNA sequencing. Several DNA sequencing techniques, including numerous "next generation technologies" are summarized in Table 1, and the reader is encouraged to explore further. Previously, sequencing costs were prohibitively expensive for most molecular laboratories and sequencing-based initiatives were limited to large collaborations or institutions, yet costs have been dropping exponentially. In 1990, Sanger sequencing cost roughly $10 per base pair. Now, the same basic technique costs roughly one tenth of a cent for a base pair. New high-throughput sequencing technologies (e.g., Life Sciences 454, ABI Solid, Illumina) generate sequencing data at the cost of 1/1000th of a cent per base pair (Table 1). Further, the establishment of university run sequencing centers and commercial sequencing companies (e.g., Macrogen) remove the technical difficulties of actually running the sequencers and performing the preliminary quality measurements. This allows individual laboratories around the world to obtain massive amounts of sequence data quickly with modest investment. With the impediments to acquiring DNA sequence data removed, new obstacles emerged.

One challenge involves unfamiliarity with new sequencing technologies; all sequencing methods have an inherent error rate and a set of associated quality controls. Many molecular biologists are familiar with the processing of Sanger sequencing data and the manual examination of the raw trace files used to be mantra for first year graduates. With falling Sanger costs and the advent of high-throughput sequencing, researchers turned to using quality control scores to weed out spurious or error prone reads. Similar methods are required for all of the next-generation technologies as well. For example, pyrosequencing (e.g., 454) originally had a relatively high error rate (4 of 100 base pairs), which necessitated a set of quality control steps, including the removal of reads with unresolved bases, abnormal length, or errors in the barcode or ligator sequences (Margulies et al. 2005; Huse et al. 2007). Subsequently, Kunin et al. (2010) suggested even more stringent controls and refuted the biodiversity estimates made by Sogin et al. (2006). To some extent, this is a red herring, as the diversity estimates from Sogin et al. (2006) appear to be only moderately modified (Huse et al. 2010). Most published studies used older 454 chemistries, whereas the new Life Technologies Titanium chemistry has a greatly reduced error rate along with a longer read length. The Huse and Kunin studies examined the influence of the error rate on amplifications of internal regions of the 16S gene, a homopolymer rich molecular that lends itself to increased error rates (Margulies et al. 2005). As many functional genes do not contain long homopolymer regions, the issue is less critical for non-targeted metagenomic studies. Similarly, the ABI Solid platform produces sequences in color-space, and the conversions to sequence can be very error prone (Dupont, personal experience). Scientists using high-throughput sequencing as a method must familiarize themselves with the multiple rapidly evolving technologies and the benefits and drawbacks of each method.

Another challenge is the sheer mass of data. Researchers accustomed to modest clone libraries are now receiving 1 million sequences from a single 454 run. The traditional methods of manual alignments, phylogenetic analysis, and visualization are no longer possible. Manually performing interative BLAST searches with 1 million sequences and collating the results is completely untenable. The ability to manipulate massive, text-based datasets, apply the desired analyses, and generate quantitative summaries represents a major bottleneck. As a real example, one of the authors witnessed how in less than a year (2008-09), his whole lab (a medium-sized lab in Europe specializing in microbial ecology and evolution) moved almost entirely into 454 FLX sequencing, leaving behind older, but well-established techniques, like tRFLP. Most of the scientists are desperate to learn the bioinformatic tools and techniques needed to process and depurate that data. Naturally, the resident bioinformatics expert (bioinformaticist, bioinformatician, or bioinformagician, if you will) in the group attained a status similar to that of a tribal shaman with mystical powers.

Finally, envision a virtually optimal situation: a molecular

Table 1. DNA sequencing technologies. The revolutionary new generation technologies to sequence DNA are dropping off the cost of DNA sequencing. Here we summarize the past (Sanger), the present (454, Illumina, Solid), and future (the rest) of automated DNA sequencing technologies (data modified from [Shendure and Hanlee 2008] and amended based on technological developments and costs at the J. Craig Venter Institute).

	Sequencing by synthesis	Read length	Seqs per run	Total sequence	$/Mbp
Sanger	Dye-terminator sequencing	700-1000 bp	96	0.07-0.1 Mb	$1000
454 Titanium	Polymerase (pyrosequencing)	400-500bp	1,000,000	500 Mb	$30
Illumina	Polymerase(reversible terminators)	100-125 bp	500,000,000	50 Gb	$0.9
ABI Solid	Ligase (octamers with 2-base coding)	50-75bp	200,000,000	17 Gb	$1.7
Polonator	Ligase (nonamers)	13 bp*	350,000,000*	4.5 Gb*	1$*
HeliScope	Polymerase (asynchrnous extensions)	30bp*	450,000,000*	14 Gb*	1$*

*Early estimates.

biologist, trained in programming and possessing a clear processing schema/pipeline. During most of her research, she managed to run most analyses in her personal computer or a cluster of 10 Linux computers interconnected at her institute. However, when trying to work with her six new 500 Mbp datasets, she finds that even simple analyses take long time, and that the most complex ones saturate the storage capacity of the system. Complicating matters, other researchers are demanding storage space and computational power. Computing power and data storage represents potentially one of the most significant challenges to the future of molecular science.

Advancing beyond the challenges—The first two challenges are the responsibility of the researcher in question, whereas to a greater degree the third challenge applies to the community as a whole. A scientist must have a working knowledge of the pros and cons of the various sequencing methods, as well as the diversity of approaches within each technology. For example, a relatively simple microbial community, such as a collection of cells collected by flow cytometry based upon fluorescence characteristics, can be analyzed quite elegantly with 454 sequencing of paired end libraries (see glossary) followed by assembly (Tripp et al. 2010). In contrast, a comparison of the composition of multiple communities might better be accomplished through barcoding (see glossary) of 16S rRNA amplicons from each community with subsequent multiplex sequencing.

The difficulty in managing large sequence datasets requires learning programming or collaborating closely with bioinformatics experts. In medium to large labs, a standard approach is to contract programmers as a support for the researchers or recruit students and post-doctoral researchers interested in applying bioinformatics to environmental issues. Today, several labs are moving into this direction, and it is becoming very familiar to meet computer scientists in labs traditionally steeped in molecular biology. In addition, many universities and research institutes have developed informatics departments, providing in-house collaborators to small and large labs alike. Even in these scenarios, the environmental scientist must be familiar and conversant enough in the methodologies to best communicate their desires and goals.

Each of the previous challenges will be addressed with the incorporation of sequencing and bioinformatics into the education system at all levels. However, a major requirement for the future of science is the investment in computational capacity and refinement of public databases and cyberinfrastructure. Genbank (http://www.ncbi.nlm.nih.gov/Genbank/) and EMBL (http://www.ebi.ac.uk/embl/) are pioneer examples of public access databases that have facilitated the advance of molecular biology, ecology, and evolution. However, the deposition of the Global Ocean Survey dataset, which dramatically increased the raw amount of sequence information available for query, challenged both of these databases. In response, other promising cyberinfrastructure projects appeared, including CAMERA (http://camera.calit2.net/), SEED (http://www.theseed.org/wiki/ Home_of_the_SEED), STRING (http://string.embl.de/) and Greengenes (http://greengenes.lbl.gov/) (Overbeek et al. 2005; Desantis et al. 2006; Seshadri et al. 2007; Jensen et al. 2009). Other projects, like the Norwegian project Bioportal (http://www.bioportal.uio.no/), specialize in providing free access to high-throughput multipurpose computing capacity. It is vital for the community to acknowledge the importance of these existing projects, which are regularly audited to consider whether they warrant funding, and to generate new cyberinfrastructure.

The effect of trickle down Reagan economics in science—The requirement for public deposition of sequencing data upon publication incited advances in diverse fields like ecology, evolution, biodiversity, biogeography, and biochemistry. Projects like the Global Oceanic Sampling (GOS) expedition by the J. Craig Venter Institute, which provided the community with a large metagenomic dataset used extensively by groups not involved in the original sampling and data analysis, exemplify the benefits of large scale sequencing. Similarly, microbial genomes sequenced and released to the community with the support of the Gordon and Betty Moore Foundation and the Joint Genome Institute are also a community resource. Overall, the generation of large DNA datasets by sequencing facilities with the financial support of governments and foundations has given extraordinary results, promoting the growth in several other fields through the direct injection of new data.

A more subtle but equally important influence is the growth of fields involved in handling and analyzing the sheer flood of data. Graduate degrees in bioinformatics, a melding of programming, mathematical theory, biology, and statistics, are now part of the curriculum at academic institutions around the world. Similarly, the need for individuals and teams trained in information technology grows with each new sequencing advance; an ABI Solid Sequencing run delivers massive amounts of image data, which can quickly saturate a dedicated server. Even after conversion of the color space data to sequence space, each run generates 10 Gbp, at 8 bits per base pair and 32 bits per label character. Further, for the largest sequencing institutions, regular memory backups are *de rigeur,* yet this becomes a substantial issue when the database grows by Gbp per day. The raw amount of sequence data generated daily approaches 500 Gbp and is expected to increase; novel and creative ways of managing, protecting, and searching this data will be essential. The molecular revolution is creating a wealth of careers for college graduates beyond the obvious.

A previously unfathomable feast for microbial ecologists and evolutionary biologists—Despite the difficulties stated above, we want to stress that the sequencing tsunami is a boon for microbial ecologists and evolutionists. The availability and possibility to produce huge amounts of data allows for the exploration of questions that were seemingly intractable. For example, next generation sequencing can be linked to evolutionary diversification experiments, a detailed revisitation of Lenski's classic experiments studying microbial evolution in real-time (e.g., Cooper et al. 2003). In addition, large amounts of sequencing data can provide a deeper insight to microbial population genetic, biogeochemical interactions of different microbial taxa, and the nature of microbial species (if there is such a thing). Additionally, with the proper collection of environmental data, it may be possible to investigate the power of natural selection in structuring microbial communities at local scales. In this sense, a strict adherence to the Minimum Information of Metagenome Sequences (MIMS) standards will greatly facilitate comparative metagenomics (Sterk et al. 2010). However, the discovery that different DNA extraction methods yield contrasting pictures of in vitro model communities suggests that great care must be used when comparing datasets collected by different research groups (Morgan et al. 2010)

Microbial biodiversity and biogeography has greatly benefited already from molecular data and will continue to in the future. In particular, we will most likely acquire a much clearer picture of the extent of microbial diversity on Earth and its spatiotemporal distribution patterns. Long-standing biogeographic hypotheses about microbes like “everything is everywhere, but the environment selects” (Baas Becking 1934) will be tested in depth using new sequence data. Associated hypotheses that most microbial ecosystems are constituted by a few common and abundant species, and many rare taxa (Pedros-Alio 2006) have also been tested and preliminarily verified (Yooseph et al. In press). Last but not least, we will be able to investigate whether ecological rules developed for macroorganisms apply to microbes or not. For example, the long-lived paradigm of a latitudinal trend in diversity has been recently been confirmed for marine planktonic bacteria (Fuhrman et al. 2008).

The dilemma for a post-molecular era researcher—At least during the last five decades, much of the research in microbial ecology consisted of an approach similar to the industrial vertical integration practiced by Carnegie steel and many oil companies. Sample collection, lab work, data analysis, and publication of the results were all handled by the same research group. Let us call this the “generalist approach” (GA). The GA seems to be still omnipresent in the working methodology of several labs, most likely as a heritage rather than as the most efficient working methodology. Overall, the rapid evolution of microbiological research makes it increasingly difficult for a single researcher or a small group to continue using the GA. A researcher or group using a generalist approach to any given problem must master, in addition to the complexity and science behind the problem: (a) the ability to produce molecular libraries and sequence them, (b) bioinformatics theory and programming, and (c) the usage of high-throughput computing facilities. A group of GA researchers may spend a long time, potentially more than the typical grant lifetime, to deliver the products of their research. With the inherent costs of sequencing, data management, and the salaries of a large laboratory, the GA has become increasingly expensive. A few laboratories, notably those funded by the Gordon and Betty Moore Foundation (www.moore.org), have been able to pursue the GA while embracing the sequencing revolution. This sort of integration has resulted in remarkable scientific breakthroughs and also provided holistic training for a host of graduate students and post-docs, though admittedly the non-peer reviewed process of Moore Foundation funding has created a culture of haves and have-nots.

In contrast, other research groups using a “specialist approach” (SA) along with well-planned collaborations may deliver the products of similar research in a much shorter time. By subcontracting companies to do the sequencing, costs and data delivery time are minimized. Collaborations with computer scientists for data handling and analysis facilitate a focus on creatively integrating these new technologies and datasets into their field. Naturally, the GA still seems to be the default option when a new research project is started, particularly for young investigators who are under institutional pressure to prove their independence. However, even for the young researcher, it is a good idea to contemplate the possibility that using the SA approach may give better and faster results for some projects, allowing for an unprecedented diversity in research. Admittedly, the SA does require a substantial input of intellectual energy for the ecologist, as the sequencing and bioinformatic approaches must be tailored to the system and question, necessitating knowledge of the various pros and cons. Further, the logistical difficulties in implementing a project can increase disproportionately with the number of collaborators.

A historical case study: Marine microbial ecologists improbably jump from past to future

We laid out the problem that researchers are facing—a flood of molecular data— and began to suggest how to deal with and capitalize on this flood of data. But, to truly understand where we need to go, it is helpful to examine the path to the modern era. Before DNA sequencing became commonplace, macroecologists seemed to have all the fun. Life was broken down into two kingdoms: plants and animals, both endowed with a glittering and visible phenotypic diversity. As microscopes became more powerful, researchers discovered the cryptic world of microbiology. With the description of bacteria and fungi, the kingdoms of life grew from two to four (Copeland 1938) to five (Whittaker 1969). Though marine viruses were isolated and known to exist (Spencer 1955), their role in the marine food web and phylogenetic standing was largely unknown. No single kingdom structure seemed to be able to adequately define life and more kingdoms were added every few years, to the annoyance of biology teachers and delight of textbook manufacturers. Microbiologists were limited to gross morphological distinctions of organisms, endless plating, and that unique headache that comes from a day of squinting into a microscope in a dark room.

In 1977, Carl Woese and George Fox changed the way that microbial ecologists and indeed all biologists look at the world by sequencing the 16S ribosomal RNA. The sequence of this molecule, present in all living life (a tacit exclusion of viruses), was presumed to provide information about the relatedness of organisms. By using each nucleotide position of the 16S sequences instead of morphological characteristics as the phylogenetic characters, Woese and Fox (1977) reclassified life into three superkingdoms or domains: Prokaryote, Eukaryote, and Archaea. Take a moment to think about this: an entire new branch of life was discovered. Naturally, this new system of classification was initially met with skepticism, yet is now widely accepted. Woese and Fox's (1977) research altered the tree of life and phylogenetic classification, but also provided a new tool for classifying newly discovered organisms. At the time of this discovery and the years immediately following it, sequencing was prohibitively expensive (over $10 per base pair) and complicated, limiting the usage by ecologist. As technological advances made sequencing more accessible, 16S sequencing became a staple approach for microbial phylogenetics.

In the late 1980s and early 1990s, sequencing-based techniques led to the discovery of a massive uncultured microbial diversity. In 1988, one of the most abundant photosynthetic organisms in the oligotrophic open ocean, *Prochlorococcus*, was discovered and distinguished using flow-cytometry and pigment analyses (Chisholm et al. 1988). Hard to cultivate (Chisholm et al. 1992), the sequencing of the molecular markers *rpoC* and ITS revealed the phylogeny and relationship to other marine cyanobacteria (Palenik and Haselkorn 1992; Urbach et al. 1992). Another major bacterioplankton group, SAR11, was completely unknown until cloning and sequencing efforts revealed its abundance in the oligotrophic ocean (Giovannoni et al. 1990). Whereas it was 13 years before a representative of this major group of oceanic bacterioplankton was cultivated in the laboratory (Rappe et al. 2002), 16S surveys of different ocean regimes revealed that SAR11 is ubiquitous and abundant in the oceans. In 1992 came the startling discovery of that the branch of life thought to be limited to "extreme" environments, the Archaea, is both present and abundant in seawater (Delong 1992; Fuhrman et al. 1992). Thus the tools that Woese and Fox developed in 1977 were put to great use by marine microbiologists in the 1990s. As a result of DNA sequencing, we were aware of the "The Uncultivated Microbial Majority" as we entered the new millennium (Rappe and Giovannoni 2003). However, most of these diversity studies used non-metabolic genes for markers and many of the observed organisms were not in culture, therefore stymieing the elucidation of the ecological function of a given organism, such as the newly discovered Crenarcheota.

Early in the 21st century, a number of additional technological and theoretical breakthroughs prompted further advances in marine microbiology. Perhaps equally important was a return to a Darwinian style of science; early metagenomic experiments are probably best described as "discovery" experiments, a direct contrast to "hypothesis"-based experimental designs. Essentially, in the "discovery" approach, researchers search raw environmental DNA sequence data for new metabolisms and diversity (Table 2). Whereas this type of approach may seem anathematic to a student in analytical chemistry, the results of the "discovery" approach have been instrumental to hypothesis generation for more subsequent studies. While there are many examples, we will focus on two. The sequencing and bioinformatics curation of a bacterial artificial chromosome (BAC) library (Table 2) made from community in the seawater of Monterey Bay (Beja et al. 2000b) revealed a bacterial rhodopsin in the genome of marine bacterioplankton that was able to act as a light-driven proton pump (Beja et al. 2000a). Further experiments using degenerate PCR primers found that the genes for this novel mechanism

Table 2. DNA sequencing methods. For sequencing genomes of cultured microorganisms (pure-culture genomics) or DNA of mixed microorganisms retrieved from the environment (metagenomica), two main strategies can be used. The main difference between them is the size of the DNA fragment resulting of fragmentation of the original DNA. After sequencing, DNA fragments are assembled in silico to find the original nucleotide sequence. Data modified from Moran 2008.

Sequencing strategy	Metagenomic technique	Size of DNA fragment	Cloning	Reference	Information about
Hierarchical sequencing (Large fragments)	Bacterial Artificial Chromosome (BAC)	100 Kb	Y	(Beja et al. 2000b)	Adjacent genes Operon structure Metabolic pathways
	Fosmids	40 Kb	Y	(Delong et al. 2006)	Gene discovery Gene abundance and diversity
Whole-genome shotgun sequencing (short fragments)	WGS	0.8 Kb	Y	(Rusch et al. 2007)	High-throughput sampling of many samples Higher amount of sequences
	Pyrosequencing	0.1–0.4 Kb	N	(Edwards et al. 2006)	More representative of a metacommunity Gene discovery Gene abundance, diversity and distribution

Note that the size of one gene is approx. 1 Kb. So, as example, a DNA fragment of 100 Kb can contain around 100 continuous genes.

existed in surface waters from around the globe and that the absorption of this rhodopsin is tuned to the spectral light quality of the water column (Beja et al. 2001). Our second example was revealed by shotgun (*see* "Glossary") Sanger sequencing of the Sargasso Sea microbial community 0.1 µm-0.8 µm in size and the subsequent bioinformatic assembly (*see* "Glossary," Table 2) of the greater than 1 million reads into bits of microbial genomes (Venter et al. 2004). One assembly contained several genes that were clearly of Archaeal origin but also contained the gene coding for ammonia monooxygenase, which heretofore had been believed to the provenance of the Bacterial superkingdom. As with proteorhodopsin, subsequent experiments verified the presence of the gene in the genomes of cultivated Archaea (Konneke et al. 2005) and the metagenomes of microbial communities from many regions of the oceans, particularly in the bathypelagic (Francis et al. 2005). Essentially, "discovery"-style metagenomic experiments can yield very exciting results, but sometimes finding those one or two novel findings can seem like a very daunting task at the outset.

These initial experiments showed that valuable ecological data could be acquired through the methods of fosmid library construction and shotgun sequencing. Subsequently, these methods have been used in a more traditional experimental setup. Again, while there have been many studies of this kind, we will focus on two. DeLong and colleagues (2006) constructed fosmid libraries from multiple depths at the Hawaii-Ocean-Time-Series station Aloha and end sequenced them using Sanger sequencing. Among other results, this study revealed distinctive vertical gradients in protein families, a scenario analogous to the biogeography of species (Delong et al. 2006). Mou et al. (2008) manipulated coastal seawater by adding different carbon sources along with the atypical nucleotide BrdU. Organisms that replicated during the incubation time would incorporate BrdU into their DNA, allowing for the subsequent enrichment through immunoprecipitation. The immunopreciptated DNA was then shotgun sequenced using 454 pyrosequencing. As expected, only select members of the community responded to the new carbon sources, but there was little difference in the community response to different carbon compounds. This suggests that coastal microbial communities contain a "generalist" population equipped to use temporally variant inputs of a broad variety of carbon sources.

Interdisciplinary collaboration has become a fundamental requirement to both reveal and understand the interactions between the marine microbial world and the biogeochemistry of the planet. One example is the cycling of dimethylsulfoniopropionate (DMSP). This sulfur-containing compound is synthesized by different taxa of marine phytoplankton, for whom it may serve multiple roles including osmotic balance and oxidative stress response (Sunda et al. 2002). Upon phytoplankton death or grazing, DMSP, which contains reduced sulfur and carbon, is released into seawater where it can serve as a source of sulfur and carbon for the ubiquitous bacterium SAR11 among other organisms (Tripp et al. 2008). Through several alternative pathways microorganisms can degrade DMSP to produce dimethylsulfide (DMS), potentially as a way to attract phytoplankton grazers, which in turn, increases the flow of available carbon. DMS, a gas responsible for the "smell of the sea," can ventilate to the atmosphere where its oxidation products act as cloud condensation nuclei, scattering incoming solar radiation back to the space and promoting a hypothetical cooling effect of the planet (Charlson et al. 1987). While the relative fraction of DMSP released to seawater that is assimilated versus converted to DMS has been studied, there is little understanding of how these pathways are regulated by microbial community composition and physiology. Recently, the sequencing of microbial genomes facilitated the discovery of key genes encoding assimilatory and dissassimilatory DMSP degradation, which when coupled with the available metagenomic datasets allowed a first pass approximation for the importance of the different pathways in a broad array of aquatic environs (Howard et al. 2006; Todd et al. 2007, 2009; Curson et al. 2008; Vila-Costa et al. 2010). However, as elucidated below, metagenomics only presents a fraction of the information. Further studies on expression of these genes and proteins in environmental samples (metagenomics, metatranscriptomics, etc.) will be needed to elucidate the role of bacteria in the oceanic sulfur cycle. For example, many of the genomes of organisms that possess the assimilatory DMSP degradation pathway also contain the disassimilatory degradation pathway.

The role of model organisms in microbial ecology and metagenomics—The original genome sequences of marine microbes, both prokaryotic and eukaryotic, stimulated a suite of ecological hypotheses that have been subsequently tested and verified. The genome sequences of marine *Prochlorococcus* and *Synechcoccus* revealed the presence of regions characterized by the watermarks of horizontal gene transfer, and it was hypothesized that these regions might encode proteins involved in niche differentiation (Palenik et al. 2003; Rocap et al. 2003). Subsequent metagenomic and genomic sequencing revealed that these genomic "islands" are rarely conserved in natural populations, suggesting that they are important in fine-tuning or customizing the ecological function of these organisms at a sub-species level (Coleman et al. 2006). Specifically, they appear to be important in changing nitrogen and phosphate assimilation (Martiny et al. 2006, 2009a, 2009b), and possibly the interaction with phages and grazers (Strom 2008). However, even for a fully closed genomic sequence, one where every base pair is known, a substantial portion of the open reading frames code for proteins with little or no sequence similarity to functionally characterized proteins. This portion dismayingly falls into a rapidly growing population of "hypothetical" proteins. This is the case for many of the encoded proteins in the cyanobacterial genomic "islands."

Fortunately, a full genome sequence for a model organism, or even a series of partial expressed sequence tags (ESTs, *see* "Glossary"), does facilitate analysis of physiology and the wealth of uncharacterized proteins (Table 3). For example, DNA microarrays can be constructed for the entire genome or just the open reading frames. The entire complement of expressed RNA, following conversion to cDNA and labeling with fluorescent tags, is hybridized to a glass slide tiled or spotted with the complimentary sequences. In this fashion, the transcription of a fraction or the whole of a genome can be measured for different growth conditions or stresses, providing functional information. Even at the most basic level, the expression of a hypothetical gene provides strong evidence that it indeed has a function, allowing a revised and improved annotation of the genome (Allen et al. 2008; Mock et al. 2008; Tetu et al. 2009).

A well-annotated genome can be a godsend for analyzing metagenomic datasets. The recruitment of metagenomic sequence reads to a genome allows them to be tied to specific organisms, and in some cases, a function (Rusch et al. 2007). However, a comparison of the genomes of over 200 marine prokaryotes and the Global Ocean Survey sequencing revealed that a substantial proportion of the marine microbial community is unrepresented by the organisms in culture (Yooseph et al. 2010). Therefore, a major challenge for metagenomics lies in being able to work with sequences that have no known representative in culture. One bioinformatics approach that may have great promise involves the bioinformatic assembly of metagenomic sequences into contiguous DNA sequences. This approach led to the discovery of Archaeal ammonium oxidation (Venter et al. 2004). Recently, the process of "aggressive" assembly resulted in near complete genomes of two unique phylotypes of marine *Prochlorococcus*. A comparative analysis of these assemblies with those of other *Prochlorococcus* suggests that these phylotypes are physiologically adapted to the pervasive high nutrient, low chlorophyll, low Fe regions of the ocean, which is consistent with the observed biogeography (Rusch et al. 2010). However, only the genomes of the predominant organisms within a sample are likely to be amenable to assembly, and most metagenomic experiments only sequence a very small fraction of the genomic material in each sample. Therefore, an approach coupling cell sorting followed by multiple displacement amplification to isolate individual genomes (Ishoey et al. 2008) might be the future for assembly based metagenomics.

This diversity of uncultured organisms highlights the need for cross discipline collaborations between researchers working in metagenomics and physiologists talented at isolating and culturing marine microbes. The development of the High Throughput Culture Collection involved a reinvention of the typical culturing approach of using nutrient rich media (Giovannoni et al. 2007). Instead, diverse mixtures of nutrient poor media and exceptionally high dilution rates led to the

Table 3. Sequencing enabled tools. Here we list some of the tools available and the information that DNA sequencing based techniques offer (cDNA = complimentary DNA, but specifically here it is reverse transcribed mRNA; RTqPCR = real time quantitative PCR; LC = Liquid chromatography).

Material	Technology	Synopsis	Application	Reference (example)
DNA	metagenomics	Sequencing DNA directly retrieved from the environment	Gene discovery, diversity and abundance	(Venter et al. 2004)
	Molecular phylogenetics	16S rRNA clone libraries (V6, V3)	Organisms evolutionary relationships Taxonomy, biodiversity	Sogin, Gilbert, Brown
RNA	meta-transcriptomics	Sequencing mRNA extracted and amplified from the environment profiles	Direct analysis of gene expression	(Gilbert et al. 2008)
	RNA chips (microarray)	Synthetic oligonucleotides in a array that hybridize with cDNA	Detection and comparison of gene expression profiles	(Parro et al. 2007)
	RTqPCR	Amplification of mRNA using specific primers for a functional gene	Detection and quantification of specific expressed gene	Suzuki et al. 2001
	EST library (Expressed sequence tag)	Collection of clones containing reverse-transcribed mRNAs. (cDNA library)	Analysis of gene expression Identification of new genes Gene sequence determination	(Nagaraj et al. 2006)
Proteins	Antibody generation	Screening of a protein extract with specific antibodies	Protein activity in nature	
	Antibody chip	As with antibodies, but multiplexed	Detection and quantification of multiple proteins	(Fan et al. 2008)
	MS/MS (tandem mass spectrometer) LC/MS/MS	Separation of mixed proteins by chromatography and estimation of masses of peptides	Identification of relevant proteins Study protein fingerprinting of organism	(Wilmes and Bond 2009)

isolation of a host of marine microbes previously only observed in clone libraries and other metagenomic datasets. In reverse fashion, well-assembled metagenomic datasets might provide insight to the physiology of the uncultured portion of the marine microbial community, allowing for directed culturing efforts. Essentially, a concerted approach between scientists of disparate disciplines will be required to further unveil the uncultured and numerically significant portion of the marine microbial community.

Advancing beyond the base of the central dogma in the environment: transcriptomics and proteomics—Whereas DNA sequencing opened the cryptic world of marine microbial diversity and metabolism, only part of the central dogma is observed with metagenomic and genomic sequencing (*See* "Box 1"). An organism will only express a fraction of its genome at any given time, and the translation of these RNA transcripts to proteins is also regulated. Finally, even after translation, protein activity is modulated by allosteric interactions and protein-protein interactions. Essentially, the presence of a DNA sequence for a unique metabolism or microbe does not indicate how the metabolism or organism behaves in the environment. Therefore, researchers have begun to study the downstream products of genomes and metagenomes, RNA and proteins (Table 3).

Transcriptomic studies attempt to characterize and quantify the diversity and abundance of RNA transcripts within an organism or ecosystem. Whereas the exact process varies to some degree, generally RNA is isolated and converted to cDNA via the viral enzyme reverse transcriptase. The cDNA is amenable to the typical library construction procedures required for shotgun sequencing (Poretsky et al. 2005, 2009; Gilbert et al. 2008). Proteomics involves the isolation of proteins, digestion at specific bonds, separation by high performance liquid chromatography, and mass spectrometric identification of the protein fragments. An alternative method is separation by two-dimensional electrophoresis and subsequent mass spectrometry. The result is a series of peptide fragment masses (Wilmes and Bond 2009).

In each of these techniques, the availability of a DNA sequence dataset for reference is necessary to different degrees. For transcriptomics, the cDNA sequence and translated amino acid sequence can be searched against genomic and metagenomic sequence datasets. This can allow the attribution of a transcript to both an organism and a function, yet the efficiency of this step hinges upon the relatedness of the cDNA and DNA datasets. For example, the relative paucity of genomes for eukaryotic phytoplankton may limit the interpretation of a transcriptomic library built from a natural phytoplankton assemblage. Even given homology to a genomic sequence, the large proportion of individual genes that code for proteins with unknown function confounds a holistic connection of environmental transcripts to a physiological and ecological function. Conversely, transcriptomics may provide a method for elucidating the functions of the unknown proteins through the examination of expression in response to a variety of environmental stimuli (Gilbert et al. 2008). This obviously does not provide an absolute function for the unknown protein but does allow the reclassification from "hypothetical" to "X-regulated protein."

Proteomics depends absolutely upon a sequence library for reference. Only through the matching of peptide masses to a protein sequence library can the presence and putative function of specific proteins be confidently determined. However, despite this difficulty, proteomics is not subject to the caveats of transcriptional or posttranscriptional regulation. Essentially, if detected, the protein is expressed. As with transcriptomics, with a complementing genomic or metagenomic database, proteomics allows for a reinterpretation of genomic data. Hypothetical proteins that are detected by proteomics can be confidently removed from the "hypothetical" pool, and peptide libraries generated under different conditions can provide functional information. Finally, the use of stable isotope feeding experiments can provide information to the turnover times of specific proteins (Li 2010), though this has yet to be used in aquatic systems.

Is the new molecular toolkit limited to just microbiologists?—Naturally, the answer to this question is no. Indeed, a molecular approach can yield incredible insights in non-microbial systems and the dismissal of such a toolkit, despite the apparent hurdles, is foolhardy. For example, the expression of specific stress response proteins provides a much more sensitive assay to the physiological state of an organism exposed to competition than growth rate or a variety of other parameters. The caveat is the relative lack of complete genome sequences, particularly for the multicellular Eukaryotes that are keystone species in many aquatic ecosystems. Due to the size of Eukaryotic genomes and the requisite cost of sequencing an entire genome, this barrier might seem intractable for researchers working with limited budgets.

There are indeed increased numbers of eukaryotic genomes representative of diverse phyla (*see* http://www.genomesonline.org for a most recent list) that should facilitate the development of a molecular toolkit for nearly any organism at a minimum of cost. At the base of our suggested approaches is the construction of a series of EST libraries. ESTs only target expressed parts of an organism's genome while introns and other "junk DNA" are avoided, maximizing the generation of usable functional gene models for each unit of sequencing. By building EST libraries from an organism exposed to a series of stressors, one can sequence a broad diversity of transcripts. Further, certain genes will be expressed in greater abundance under conditions of stress, allowing for the preliminary identification of specific genes and encoded proteins that might be indicative of the physiological state of an organism.

Following bioinformatic curation and assembly, the high throughput sequencing of EST libraries will provide a large array of high quality gene models. With these, several low cost targeted assays become available to researchers interested in ecological questions. In one scenario, a researcher can examine the transcription of a single gene or set of genes using quantitative reverse transcriptase PCR. Alternatively, given the low cost of microarray printing, a researcher can assay the transcription of entire complements of genes. A complete gene model also allows for the production of a genomic antibody, which targets specific amino acid sequences. The use of antibodies to study "keystone" proteins was used quite effectively in marine microbiology before the birth of the genomic era. The relative abundance of the photosystem electron transfer proteins flavodoxin and ferredoxin provided insight to the extent that a natural population of marine phytoplankton are starved for iron (Laroche et al. 1996). Critically, the generation of a series of robust gene models dramatically increases the diversity of potential targets and circumvents the conventional and difficult task of isolating an individual protein. These antibodies can also be printed onto glass slides, creating microarrays for the detection of numerous proteins at once (Fan et al. 2008) (Table 3), though this has yet to be done in an environmental setting.

If developed and tested in laboratory conditions, both of these highly sensitive methods can be used to precisely query an organism about its physiological state. These "biochemical interrogations" are both relatively low in cost and much more sensitive that many traditional physiological measurements. The exact interrogations are only limited by the diversity of gene models available, an understanding of biochemistry, and the ecologist's imagination. The collection of samples to study protein abundance or gene transcription is also astoundingly simple, and with the ability to literally freeze a sample in near native state, relatively indicative of an organism's physiological state at an exact time. In contrast, many traditional measurements of stress or physiology involve extensive and invasive sample handling and manipulation.

For those interested in population dynamics, the development of EST libraries of different organisms within a population, or from different populations of an organism, will almost certainly result in a wealth of single nucleotide polymorphisms (SNPs). SNPs have a history of use in ecology (Morin et al. 2004), and through the construction of EST libraries, many new SNPs for an organism will be discovered even without a reference sequence (Ratan et al. 2010). Some of these might code for single amino acid polymorphisms (SAPS), which are often associated with specific phenotypes for an organism. Again, the subsequent assays are inexpensive.

Let us consider a hypothetical generation of an EST library using Titanium 454 sequencing. Tissue samples would be collected from numerous individual organisms across biological, physical, and chemical gradients. The extraction of RNA and conversion to cDNA for 24 samples will cost roughly $2000.00 in raw laboratory materials and require 40-80 h of labor (Dupont, personal experience). These 24 samples can be barcoded with different ligation end sequences and analyzed with a single Titanium 454 run which costs $16,000. At the time of writing, the user is provided with 400-500 million base pairs of raw data with an average read length of 400 bp; essentially the user receives over 40,000 ESTs each for 24 unique samples for under $30,000. For a comparison, the typical environmental EST libraries constructed using Sanger sequencing normally generated less than 10,000 ESTs.

The expertise to generate or perform the bioinformatic curation of EST libraries should not be an impediment (Nagaraj et al. 2006). The increasing ranks of professionals trained in molecular biology and bioinformatics improves the likelihood of finding a willing collaborator. Further, as mentioned previously, sequencing can often be outsourced, leaving the researcher with the responsibility for understanding the question and envisioning how to best use the newly available tools. Curiously enough, the aforementioned trends of positive feedback also apply in collaboration. The ecologist gains access to a sensitive and ultimately adaptable molecular toolkit, yet the raw sequence data from an organism without a genome will almost always provide a dataset of interest to evolutionary biologists and phylogeneticists. Essentially, both parties in the suggested collaboration are provided with data that they are uniquely adapted to analyze and disseminate to the greater community.

References

Allen, A. E., and others. 2008. Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. Proc. Nat. Acad. Sci. 105:10438-10443.

Baas Becking, L. G. M. 1934. Geobiologie of inleiding tot de milieukunde. W.P. Van Stockum & Zoon.

Beja, O., and others. 2000a. Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. Science 289:1902-1906 [doi:10.1126/science.289.5486.1902].

———, and others. 2000b. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ. Microbiol. 2:516-529 [doi:10.1046/j.1462-2920.2000.00133.x].

———, E. N. Spudich, J. L. Spudich, M. Leclerc, and E. F. Delong. 2001. Proteorhodopsin phototrophy in the ocean. Nature 411:786-789 [doi:10.1038/35081051].

Charlson, R. J., J. E. Lovelock, M. O. Andrea, and S. G. Warren. 1987. Oceanic phytoplankton, atmospheric sulfur, cloud albedo, and climate. Nature 326:655-661 [doi:10.1038/326655a0].

Chisholm, S. W., R. J. Olson, E. R. Zettler, R. Goericke, J. B. Waterbury, and N. Welschmeyer. 1988. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. Nature 334:340-343 [doi:10.1038/334340a0].

———, S. L. Frankel, R. Goericke, R. J. Olson, B. Palenik, J. B. Waterbury, L. Westjohnsrud, and E. R. Zettler. 1992. Prochlorococcus marinus nov. gen. nov. sp.: An oxyphototrophic marine prokaryote containing divinyl chlorophyll a and chlorophyll b. Arch. Microbiol. 157:297-300 [doi:10.1007/BF00245165].

Coleman, M. L., M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. Delong, and S. W. Chisholm. 2006. Genomic islands and the ecology and evolution of Prochlorococcus. Science 311:1768-1770 [doi:10.1126/science.1122050].

Cooper, T. F., D. E. Rozen, and R. E. Lenski. 2003. Parallel changes in gene expression after 20.000 generations of evolution in *Escherichia coli*. Proc. Nat. Acad. Sci. 100:1072-1077 [doi:10.1073/pnas.0334340100].

Copeland, H. 1938. The kingdoms of organisms. Quart. Rev. Biol. 13:383-420 [doi:10.1086/394568].

Curson, A. R., R. Rogers, J. D. Todd, C. A. Brearley, and A. W. Johnston. 2008. Molecular genetic analysis of a dimethylsulfoniopropionate lyase that liberates the climate-changing gas dimethylsulfide in several a-proteobacteria and *Rhodobacter sphaeroides*. Environ. Microbiol. 10:757-767 [doi:10.1111/j.1462-2920.2007.01499.x].

Delong, E. F. 1992. Archaea in coastal marine environments. Proc. Nat. Acad. Sci. 89:5685-5689 [doi:10.1073/pnas.89.12.5685].

———, and others. 2006. Community genomics among stratified microbial assemblages in the ocean_s interior. Science 311:496-503 [doi:10.1126/science.1120250].

Desantis, T. Z., and others. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72:5069-5072 [doi:10.1128/AEM.03006-05].

Edwards, R. A., and others. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7:57 [doi:10.1186/1471-2164-7-57].

Fan, R., and others. 2008. Integrated barcode chips for rapid, multiplexed analysis of proteins in microliter quantities of blood. Nat. Biotechnol. 26:1373-1378 [doi:10.1038/nbt.1507].

Francis, C. A., K. J. Roberts, J. M. Beman, A. E. Santoro, and B. B. Oakley. 2005. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc. Nat. Acad. Sci. 102:14683-14688 [doi:10.1073/pnas.0506625102].

Fuhrman, J. A., K. Mccallum, and A. A. Davis. 1992. Novel major archaebacterial group from marine plankton. Nature 356:148-149 [doi:10.1038/356148a0].

———, J. A. Steele, I. Hewson, M. S. Schwalbach, M. V. Brown, J. L. Green, and J. H. Brown. 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc. Nat. Acad. Sci. 105:7774-7778 [doi:10.1073/pnas.0803070105].

Gilbert, J. A., D. Field, Y. Huang, R. Edwards, W. Li, P. Gilna, and I. Joint. 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS ONE 3:e3042 [doi:10.1371/journal.pone.0003042].

Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field. 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60-63 [doi:10.1038/345060a0].

———, R. A. Foster, M. S. Rappe, and S. Epstein. 2007. New cultivation strategies bring more microbial plankton species into the laboratory. Oceanography 20:62-69.

Howard, E. C., and others. 2006. Bacterial taxa that limit sulfur flux from the ocean. Science 314:649-652 [doi:10.1126/science.1130657].

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

———, D. M. Welch, H. G. Morrison, and M. L. Sogin. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. 12:1889-1898. [doi:10.1186/gb-2007-8-7-r143].

Ishoey, T., T. Woyke, R. Stepanauskas, M. Novotny, and R. S. Lasken. 2008. Genome sequencing of single microbial cells from environmental samples. Curr. Opin. Microbiol. 11:198-204 [doi:10.1016/j.mib.2008.05.006].

Jensen, L. J., and others. 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucl. Acids Res. 37:412-416 [doi:10.1093/nar/gkn760].

Konneke, M., A. E. Bernhard, J. R. De La Torre, C. B. Walker, J. B. Waterbury, and D. A. Stahl. 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437:543-546 [doi:10.1038/nature03911].

Kunin, V., A. Engelbrekston, H. Ochman, and P. Hugenholtz. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. 12:118-123 [doi:10.1111/j.1462-2920.2009.02051.x].

Laroche, J., P. W. Boyd, R. M. L. Mckay, and R. J. Geider. 1996. Flavodoxin as an in situ marker for iron stress in phytoplankton. Nature 382:802-805 [doi:10.1038/382802a0].

Li, Q. 2010. Advances in protein turnover analysis at the global level and biological insights. Mass Spectrom. Rev. 29:717-736 [doi:10.1002/mas.20261].

Margulies, M., and others. 2005. Genome sequencing in microfabricated high-density picolitre reactions. Nature 437:376-380.

Martiny, A. C., M. L. Coleman, and S. W. Chisholm. 2006. Phosphate acquisition genes in Prochlorococcus ecotypes: evidence for genome-wide adaptation. Proc. Nat. Acad. Sci. 103:12552-12557 [doi:10.1073/pnas.0601301103].

———, Y. Huang, and W. Li. 2009a. Occurrence of phosphate acquisition genes in Prochlorococcus cells from different ocean regions. Environ. Microbiol. 11:1340-1347 [doi:10.1111/j.1462-2920.2009.01860.x].

———, S. Kathuria, and P. M. Berube. 2009b. Widespread metabolic potential for nitrite and nitrate assimilation among Prochlorococcus ecotypes. Proc. Nat. Acad. Sci. 106:10787-10792 [doi:10.1073/pnas.0902532106].

Mock, T., and others. 2008. Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. Proc. Nat. Acad. Sci. 105:1579-1584 [doi:10.1073/pnas.0707946105].

Morgan, J. L., A. E. Darling, and J. A. Eisen. 2010. Metagenomic sequencing on an in-vitro-simulated microbial community. PloS One 5:e10209 [doi:10.1371/journal.pone.0010209].

Moran, M. A. 2008. Genomics and metagenomics of marine prokaryotes. *In*: Microbial ecology of the oceans. D.L. Kirchman.

Morin, P. A., G. Luikart, R. K. Wayne, and S. W. Group. 2004. SNPs in ecology, evolution, and conservation. Trends Ecol. Evol. 19:208-216 [doi:10.1016/j.tree.2004.01.009].

Mou, X., S. Sun, R. A. Edwards, R. E. Hodson, and M. A. Moran. 2008. Bacterial carbon processing by generalist spedies in the coastal ocean. Nature 451:708-711 [doi:10.1038/nature06513].

Nagaraj, S. H., R. B. Gasser, and S. Ranganathan. 2006. A hitchhiker's guide to expressed sequence tag analysis. Brief. Bioinform. 8:6-21 [doi:10.1093/bib/bbl015].

Overbeek, R., and others. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucl. Acids Res. 33:5691-5702 [doi:10.1093/nar/gki866].

Palenik, B., and R. Haselkorn. 1992. Multiple evolutionary origins of prochlorophytes, the chlorophyll b-containing prokaryotes. Nature 355:265-267 [doi:10.1038/355265a0].

———, and others. 2003. The genome of a motile marine Synechococcus. Nature 424:1037-1042 [doi:10.1038/nature01943].

Parro, V., M. Moreno, and E. Gonzalez-Toril. 2007. Analysis of environmental transcriptomes by DNA microarrays. Environ. Microbiol. 9:453-464 [doi:10.1111/j.1462-2920.2006.01162.x].

Pedros-Alio, C. 2006. Marine microbial diversity: can it be determined? Trends Microbiol. 14:257-263 [doi:10.1016/j.tim.2006.04.007].

Poretsky, R. S., N. Bano, A. Buchan, G. Lecleir, J. Kleikemper, M. Pickering, M. A. Moran, and J. T. Hollibaugh. 2005. Analysis of microbial gene transcripts in environmental samples. Appl. Environ. Microbiol. 71:4121-4126 [doi:10.1128/AEM.71.7.4121-4126.2005].

———, I. Hewson, S. L. Sun, A. E. Allen, J. P. Zehr, and M. A. Moran. 2009. Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. Environ. Microbiol. 11:1358-1375 [doi:10.1111/j.1462-2920.2008.01863.x].

Rappe, M. S., S. A. Connon, K. L. Vergin, and S. J. Giovannoni. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418:630-633 [doi:10.1038/nature00917].

———, and S. J. Giovannoni. 2003. The uncultured microbial majority. Ann. Rev. Microbiol. 57:369-394 [doi:10.1146/annurev.micro.57.030502.090759].

Ratan, A., Y. Zhang, V. M. Hayes, S. C. Schuster, and W. Miller. 2010. Calling SNPs without a reference sequence. BMC Bioinformatics 11:e130 [doi:10.1186/1471-2105-11-130].

Rocap, G., and others. 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature. 424:1042-1047.

Rusch, D. B., A. Martiny, C. L. Dupont, A. L. Halpern, and J. C. Venter. 2010. Characterization of *Prochlorococcus* clades from iron depleted oceanic regimes. Proc. Nat. Acad. Sci. 107(37):16184-16189 [doi:10.1073/pnas.1009513107].

———, and others. 2007. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. PloS Biol. 5:398-431 [doi:10.1371/journal.pbio.0050077].

Seshadri, R., S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier. 2007. CAMERA: A community resource for metagenomics. PloS Biology 5: e75 [doi:10.1371/journal.pbio.0050075].

Shendure, J., and J. Hanlee. 2008. Next generation DNA sequencing. Nat. Biotechnol. 26:1135-1145 [doi:10.1038/nbt1486].

Sogin, M. L., and others. 2006. Microbial diversity in the deep sea and the underexplored rare biosphere. Proc. Nat. Acad. Sci. 103:12115-12120 [doi:10.1073/pnas.0605127103].

Spencer, R. 1955. A marine bacteriophage. Nature 170:690-691 [doi:10.1038/175690a0].

Sterk, P., L. Hirschman, D. Field, and J. Wooley. 2010. Genomic Standards Consortium Workshop: Metagenomics, Metadata, and Metaanalysis (M3). *In*: Pacific Symposium on Biocomputing 15:481-484 [doi:10.1142/9789814295291_0050].

Strom, S. L. 2008. Microbial ecology of ocean biogeochemistry: A community perspective. Science 320:1043-1045 [doi:10.1126/science.1153527].

Sunda, W., D. J. Kieber, R. P. Kiene, and S. A. Huntsman. 2002. An antioxidant function for DMSP and DMS in marine algae. Nature 418:317-320 [doi:10.1038/nature00851].

Suzuki, M. T., C. M. Preston, F. P. Chavez and E. F. DeLong. 2001. Quantitative Mapping of Bacterioplankton Populations in Seawater: Field Tests Across an Upwelling Plume in the Monterey Bay. Aquat. Microbial. Ecol. 24:117-127 [doi:10.3354/ame024117].

Tetu, S. G., B. Brahamsha, D. A. Johnson, V. Tai, K. Phillipy, B. Palenik, and I. T. Paulsen. 2009. Microarray analysis of phosphate regulation in the marine cyanobacterium Synechococcus sp. WH8102. ISME J. 3:835-849 [doi:10.1038/ismej.2009.31].

Todd, J. D., and others. 2007. Structural and regulatory genes required to make the gas dimethyl sulfide in bacteria. Science 315:666-669 [doi:10.1126/science.1135370].

———, A. R. Curson, C. L. Dupont, P. Nicholson, and A. W. Johnston. 2009. The dddP gene, encoding a novel enzyme that converts dimethylsufoniopropionate into dimethyl sulfide, is widespread in ocean metagenomes and marine bacteria and also occurs in some Ascomycete fungi. Environ. Microbiol. 11:1376-1385 [doi:10.1111/j.1462-2920.2009.01864.x].

Tripp, H. J., J. B. Kitner, M. S. Schwalbach, J. W. H. Dacey, L. J. Wilhelm, and S. J. Giovannoni. 2008. SAR11 marine bacteria require exogenous sulphur for growth. Nature 452:741-744 [doi:10.1038/nature06776].

———, and others. 2010. Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. Nature 464:90-94 [doi:10.1038/nature08786].

Urbach, E., D. L. Robertson, and S. W. Chisholm. 1992. Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. Nature 355:267-270 [doi:10.1038/355267a0].

Venter, J. C., and others. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74 [doi:10.1126/science.1093857].

Vila-Costa, M., J. M. Rinta-Kanto, S. Sun, S. Sharma, R. Poretsky, and M. A. Moran. 2010. Transcriptomic analysis of a marine bacterial community enriched with dimethylsulfoniopropionate. ISME J. 4:1410-1420 [doi:10.1038/ismej.2010.62].

Whittaker, R. H. 1969. New concepts of kingdoms or organisms. Science 163:150-160 [doi:10.1126/science.163.3863.150].

Wilmes, P., and P. L. Bond. 2009. Microbial community proteomics: elucidating the catalysts and metabolic mechanisms that drive the Earth's biogeochemical cycles. Curr. Opin. Microbiol. 12:310-317 [doi:10.1016/j.mib.2009.03.004].

Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Nat. Acad. Sci. 74:5088-5090 [doi:10.1073/pnas.74.11.5088].

Yooseph, S., and others. In press. Genomic and functional adaptation in surface ocean planktonic prokaryotes. Nature.