Microbial Metagenomics: Beyond the Genome

Jack A. Gilbert[1,2,3] and Christopher L. Dupont[4]

[1]Plymouth Marine Laboratory, Plymouth PL1 3DH, United Kingdom

[2]Argonne National Laboratory, Argonne, Illinois 60439

[3]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637; email: gilbertjack@gmail.com

[4]Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, California 92121; email: cdupont@jcvi.org

Annu. Rev. Mar. Sci. 2011. 3:347–71

First published online as a Review in Advance on November 12, 2010

The *Annual Review of Marine Science* is online at marine.annualreviews.org

This article's doi: 10.1146/annurev-marine-120709-142811

1941-1405/11/0115-0347$20.00

Keywords

bacteria, communities, bioinformatics, annotation, next generation sequencing, function

Abstract

Metagenomics literally means "beyond the genome." Marine microbial metagenomic databases presently comprise ~400 billion base pairs of DNA, only ~3% of that found in 1 ml of seawater. Very soon a trillion-base-pair sequence run will be feasible, so it is time to reflect on what we have learned from metagenomics. We review the impact of metagenomics on our understanding of marine microbial communities. We consider the studies facilitated by data generated through the Global Ocean Sampling expedition, as well as the revolution wrought at the individual laboratory level through next generation sequencing technologies. We review recent studies and discoveries since 2008, provide a discussion of bioinformatic analyses, including conceptual pipelines and sequence annotation and predict the future of metagenomics, with suggestions of collaborative community studies tailored toward answering some of the fundamental questions in marine microbial ecology.

INTRODUCTION

Marine microbial metagenomics (for a definition, see the sidebar Metagenomics, below) is one of the most data-rich areas of marine ecology and oceanography. Unlike previous data-intensive phenomena in oceanography, such as the World Ocean Circulation Experiment, metagenomics also produces a large amount of uninterpretable data. It is a relatively young research area very much driven by technology, and as a result it is also one of the most comprehensively reviewed fields (e.g., Schloss & Handelsman 2003, 2005; Béjà, 2004; Cowan et al. 2004; Falkowski & de Vargas 2004; Handelsman 2004; Riesenfeld et al. 2004; Rodriguez-Valera 2004; Delong 2005; Steele & Streit 2005; Green & Keller 2006; Pedrós-Alió 2006; Schwartz 2006; Ward 2006; Xu 2006; Edwards & Dinsdale 2007; Karl 2007; Kennedy et al. 2007; Schmeisser et al. 2007; Warnecke & Hugenholtz 2007; Marco 2008; Kennedy et al. 2008; Sleator et al. 2008; Langridge 2009; Singh et al. 2009; Steele et al. 2009; Tyson & Hugenholtz 2009; Wooley et al. 2010). Typing the search terms marine and metagenomic into the National Center for Biotechnology Information's database reveals 157 related articles, 23 of which are reviews; by searching for metagenomics, we can increase this to 97 reviews! As might be expected, when there are this many reviews on a relatively young subject, there also tends to be a lot of repetition. For example, we are constantly reminded of two key discoveries provided by metagenomic studies, namely, the ubiquity of proteorhodopsin (Béjà et al. 2000, Venter et al. 2004; reviewed by Fuhrman et al. 2008) and the discovery of the importance of archaeal ammonia oxidizers (Venter et al. 2004, Schleper et al. 2005, Treusch et al. 2005; reviewed by Prosser & Nicol 2008). Although these are indeed interesting highlights, a wealth of additional information has been uncovered by metagenomics, such as an incredible diversity, vast swathes of uncharacterized metabolism, increased complexity of biogeochemical pathways, and even some paradigm shifts in our understanding of marine microbial ecology.

WHAT CONSTITUTES A METAGENOMIC STUDY?

Metagenomics is presently most appropriately divided into two research areas driven by technological application, environmental single-gene surveys and random shotgun studies of all environmental genes (**Figure 1**). The first can be seen as a directed, focused metagenomic study. Single targets are amplified using the polymerase chain reaction (PCR), and then the products are sequenced, providing an analysis of the range of different orthologs (or paralogs, but this is not always discernable) for that gene within a given community. Random shotgun metagenomics is a study in which total DNA has been isolated from a sample and then sequenced—resulting

METAGENOMICS

The basic definition of metagenomics is the analysis of genomic DNA from a whole community; this separates it from genomics, which is the analysis of genomic DNA from an individual organism or cell. In fact, the most appropriate translation of meta in Greek is "beyond," and hence the term literally means "beyond the single genome study." The term was first published in 1998 in a study of soil microbes using random cloning of environmental DNA (Handelsman et al. 1998). Subsequently, definitions have varied to include any study whereby a whole community is analyzed, e.g., directed studies of 16S rDNA diversity from an environment to isolation and analysis of total DNA from environmental samples without prior cultivation (Chen & Pachter 2005). It could be argued that prior cultivation of communities, in the case of enrichment studies or community cell-encapsulation cultures, can also be analyzed using metagenomics, and hence such definitions must be kept broad.

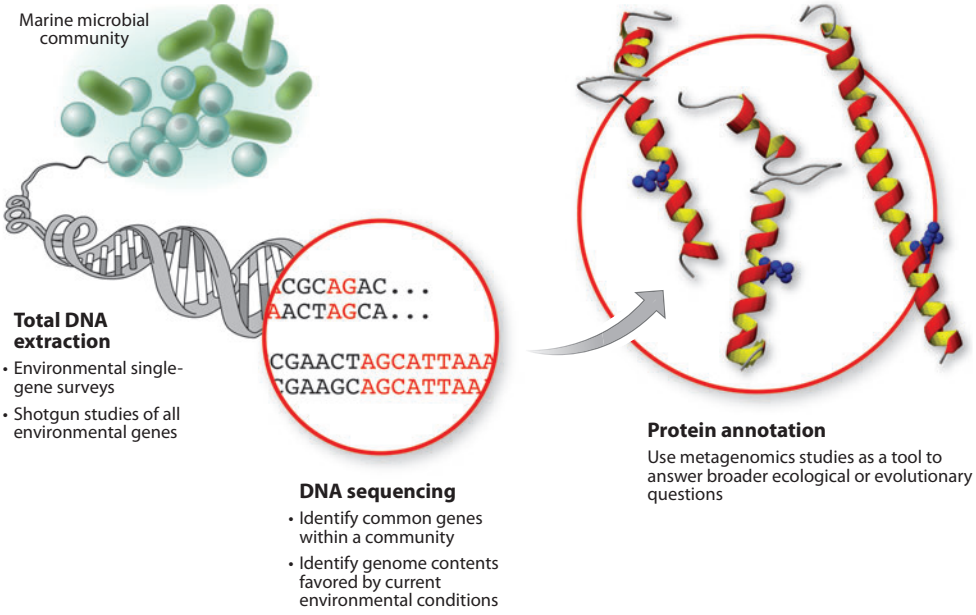

Figure 1

Metagenomics can be divided into two key research areas, environmental gene surveys and random shotgun sequencing of all genes at once. Following the latter path, we annotate the genes and determine their relationship to the environment and then use this to identify proteins that can be synthesized by the metagenome, so as to infer the proteome of a community and ultimately derive the metabolic state of a community by examining the potential of proteins to consume or produce metabolites.

in a profile of all genes within the community. The community coverage of both approaches is entirely dependent on the depth of sequencing, that is, how many gene fragments are obtained during sequencing. In this review, we will focus on the random shotgun metagenomic studies and what it has taught us about marine ecology. To prevent significant overlap with previous reviews, we will also focus only on studies released since January 2008. In this, we will capture what the most recent studies and technological advances have shown us and how this is influencing future metagenomic analyses. Additionally, we will highlight some key questions that need answering in this field.

HISTORY

How many marine microbial shotgun metagenomic studies have there been since 1998? By our calculation there have been approximately 45 major studies since the mid-nineties, with the vast majority occurring after 1998, when Jo Handelsman and colleagues coined the term "shotgun" metagenomic studies (Handelsman et al. 1998). We understand that this may not be as comprehensive as one would like, and we apologize to our colleagues in this field if we have missed any studies of relevance. These studies can be divided into the following categories.

(1) Fosmid, cosmid, and bacterial artificial chromosome (BAC)-derived metagenomic studies (e.g., Stein et al. 1996; Vergin et al. 1998; Béjà et al. 2000; Suzuki et al. 2001; Nesbø et al. 2005; Schirmer et al. 2005; Tring et al. 2005; Treusch et al. 2005; Delong et al. 2006; Kim & Fuerst 2006; Lee et al. 2006; Stokes et al. 2006; Woyke et al. 2006; Hardeman & Sjoling

2007; Martín-Cuadrado et al. 2007, 2009; Woebken et al. 2007; Gilbert et al. 2008b, 2009; Neufeld et al. 2008; Brazelton & Baross 2009; Huang et al. 2009; Kim et al. 2009; Martinez et al. 2010).

(2) Sanger sequencing–derived shotgun metagenomic studies (Venter et al. 2004, Yutin & Béjà, 2005, Martiny et al. 2006, Rusch et al. 2007, Yooseph et al. 2007, Harrington et al. 2007, Wilhelm et al. 2007, Yutin et al. 2007, Gianoulis et al. 2009, Sebastian & Ammerman 2009).

(3) Next generation sequencing–derived shotgun metagenomic studies (Dinsdale et al. 2008, 2008b; Frias-Lopez et al. 2008; Gilbert et al. 2008a,b, 2009; Mou et al. 2008; Hewson et al. 2009; Thurber et al. 2009; Willner et al. 2009; Palenik et al. 2009; Temperton et al. 2009; Tripp et al. 2010).

It was Rondon et al. (1999) who first used BAC inserts in *Escherichia coli* to study *Bacillus cereus*, a Gram-positive bacterium, opening the way to environmental DNA studies [e.g., Rondon et al. (2000), who (with soil DNA) produced a library with $>10^9$ base pairs of DNA from all soil microbes]. Béjà et al. (2000) first used BAC library cloning to isolate marine microbial metagenomic DNA, providing important information about marine *Archaea*. Béjà et al. (2002) then created fosmid libraries (a fosmid vector maintains an average insert size of 40,000 base pairs) to characterize the marine archaeal phylum *Crenarchaeota* from the Antarctic Ocean and from deep waters of the temperate Pacific Ocean. As fosmid libraries were, for a number of technical reasons (see Gilbert 2010), considerably easier to produce than BAC libraries, they became by far the most commonly used technique in marine studies. For example, Grzymski et al. (2006) produced a fosmid library from the coastal waters of the Antarctic and found proteins with specific adaptations to living in this extremely cold ecosystem. Other studies have highlighted the use of phylogenetic classification of annotated genes within a fosmid insert to infer the taxonomy of the original genome and, additionally, to identify gene clusters that occurred from horizontal gene transfer (Nesbø et al. 2005). One of the major fosmid studies to date is that of DeLong et al. (2006) at the Hawaii Ocean Time-series (HOT) station in the Pacific. This study produced 64 Mbp of DNA sequence, one of the largest studies at the time. Analysis of this data demonstrated that proteins and contiguous protein clusters occurred at specific depths—potentially equivalent to classical species zonation in terrestrial habitats. One of the most important aspects of this data set was the accessibility of the associated (so-called meta-) information with a well-characterized 20 years of environmental, biogeochemical, and biological data (Karl 2007). The lasting impact of this study demonstrates the exceptional importance of publishing these environmental characteristics with a metagenome, so as to provide context to the gene profile and aid comparison with other studies.

In 2004, the power of sequencing small-insert clones with the well-established Sanger sequencing technology was demonstrated. Venter et al. (2004) used this technique to characterize oceanic microbial assemblages from the Bermuda Atlantic Time-series Study site in the Sargasso Sea. This study produced in excess of 1 billion nonredundant base pairs and used novel bioinformatics approaches to examine the data. They reported 1,800 unique genomes, 48 unknown bacterial phylotypes, and 1.2 million previously unknown genes. It should be mentioned that all the 48 previously unknown bacterial taxa could be grouped into known 16S clades (Giovannoni & Stingl 2005), which raises some very interesting questions regarding how we define taxa derived from metagenomic data. This groundbreaking study was the first to apply techniques commonly used to sequence individual genomes to go beyond the genome. Unfortunately, even studies of this magnitude were still only scraping the surface. With approximately 1 million bacteria per milliliter of seawater and an estimated average genome size of 2 million bp, the Sargasso Sea project sequenced only 0.05% of the genomic information in a single milliliter—a proverbial drop in the ocean. Subsequent studies by the same group, for instance, the Global Ocean Sampling (GOS) expedition (**Figure 2**), used the same technology on a subset of samples from the northwest Atlantic and

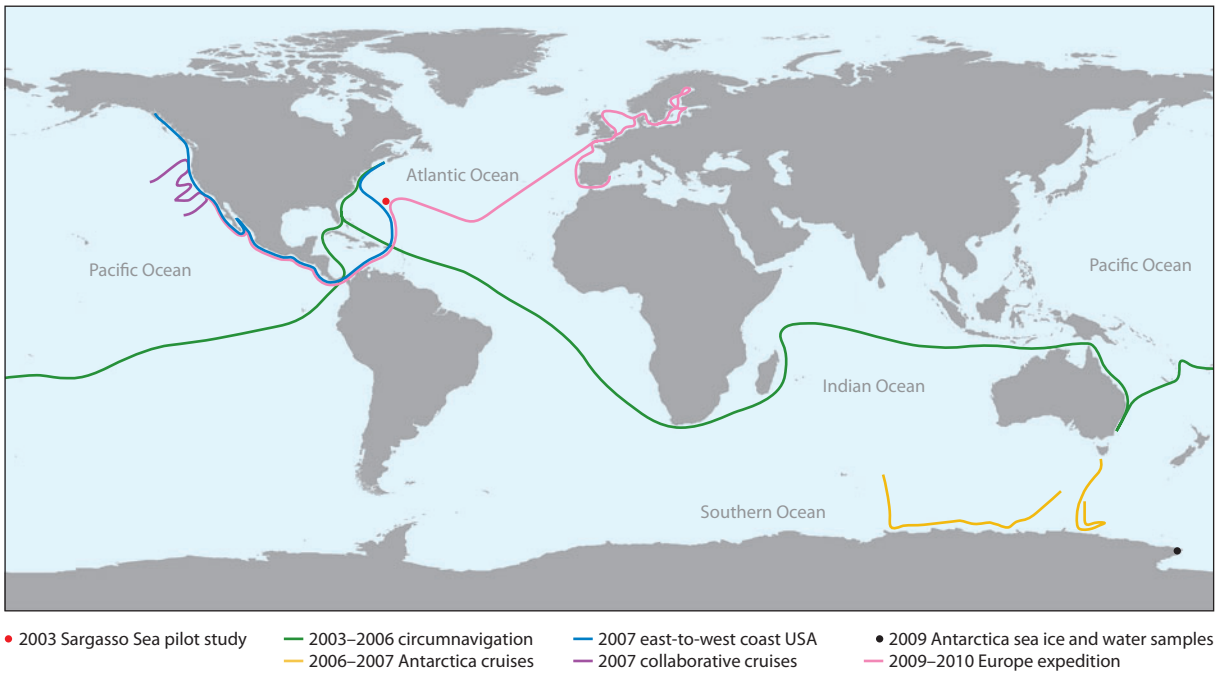

Figure 2

Map of the Global Ocean Sampling expedition.

eastern tropical Pacific, producing 6.3 billion base pairs from 7.7 million sequence-reads (Rusch et al. 2007, Yooseph et al. 2007). The public availability of this sequence data has facilitated an impressive number of independent bioinformatic studies by other researchers but has also provided a means for extending laboratory results into the natural environment. What follows is a selection of publications from several disparate specialties that have made use of the data acquired during these studies.

GLOBAL OCEAN SAMPLING: BACTERIAL DIVERSITY

Other than the primary analyses from Venter et al. (2004) and Rusch et al. (2007), a number of additional studies have used the data to investigate the biogeography and diversity of specific bacterial groups. For example, Yutin et al. (2007) analyzed the GOS data set to assess the abundance and spatial distribution of aerobic, anoxygenic, photosynthetic bacteria. They showed that this group is an important component of the bacterioplankton, with a near constant contribution between the GOS sampling locations of between 1–5% of the whole community. However, in one coastal sample south of Nova Scotia, they contributed >10% of the community, suggesting that environmental conditions may influence their relative abundance. Wilhelm et al. (2007) demonstrated the natural variation in genome content and organization within the SAR11 (*Pelagibacter ubique*) clade at different ocean sites. Importantly, this study highlighted the potential of natural selection in streamlining core features of the genome of this generalist; however, there was considerable diversity within the hypervariable regions of the genome, potentially resulting from biogeographical isolation. This suggests that microbes do exist in distinct populations between

different marine provinces. Alternatively, each potential SAR11 genome type is in each location, but the present environmental conditions favor a specific set of genome contents.

Rusch et al. (2010) used fragment recruitment and a phylogenetic analysis of marker genes to identify two previously unobserved genotypes of marine *Prochlorococcus*. After identifying enriched GOS samples, nearly complete consensus genomes were assembled using Celera software. These consensus genomes represent an ensemble view of the genome organization of populations and are not equivalent to a complete, or even draft, genome assembly from a clonal isolate. However, the consensus assemblies allow for an analysis of metabolism, revealing an economization of iron (Fe) requirements through gene loss. This metabolic adaptation was consistent with a biogeographical analysis showing the domination of these strains in the Fe-limited, high-nutrient, low-chlorophyll regions of the eastern tropical Pacific. These results suggest that the availability of trace elements like Fe might drive speciation and genetic differentiation in marine microbes.

GOS: PHOSPHORUS CYCLING

Using microarray technology, Martiny et al. (2006) experimentally determined the gene clusters from *Prochlorococcus* that are regulated by phosphate availability. They then extended these laboratory results to the environment by demonstrating that the metagenome from the oligotrophic, phosphate-limited Sargasso Sea environment is replete with the genes from these clusters. In a subsequent comparison of the GOS phase I sites, Martiny et al. (2009) examined the genomic incorporation of proteins involved in phosphate sensing, uptake, and organophosphate utilization in *Prochlorococcus*. Strikingly, they found that above an environmental phosphate concentration of 100 nM, *Prochlorococcus* genomes did not contain these proteins, which directly links genome content and environmental conditions. Sebastian & Ammerman (2009) used the Sargasso Sea data to show that the traditional alkaline phosphatase protein (phoA) used by bacteria to access inorganic and organic phosphate was considerably less abundant in the GOS databases than a more novel phoX phosphatase protein. This was an excellent example of how our understanding of nutrient cycling, which has been derived from cultured bacteria, may not always apply to the natural ecosystems. Studies by Quinn et al. (2007) and Martinez et al. (2010) highlighted the widespread distribution of enzymes specific to the degradation of the phosphonates (a class of organic phosphates characterized by a stable carbon-phosphorus bond). This group of organic phosphorus was previously thought to be recalcitrant to biological use, yet studies like GOS have demonstrated that this phosphonate assimilation is widespread in microbial communities and hence that phosphonate represents a significant phosphorus resource for marine microbes. Finally, Luo et al. (2009) found that many marine bacterial genomes contain both secreted and cytoplasmic alkaline phosphatases, whereas nearly all genomes contain a suite of genes involved in glycerol phosphate uptake. Essentially, it appears that the uptake of monoesters and diesters of glycerol phosphate with subsequent phosphate liberation in the cytoplasm may be a widespread strategy for phosphorus acquisition in marine bacteria.

GOS: SULFUR

Several recent studies have highlighted the use of heterologous expression in identifying the bacterial genes involved in the production of the climate gas dimethyl sulfide (DMS) from dimethylsulfonopropionate (DMSP; Curson et al. 2008, Todd et al. 2009). Of the three discovered, *dddP* was found to be particularly abundant in the GOS survey databases. The most abundant DMSP-processing gene is *dmdA*, which codes for a methylase that catalyzes DMSP assimilation rather than DMS production (Todd et al. 2009).

GOS: NITROGEN

As with sulfur, Johnston et al. (2005) investigated the presence of genes involved in nitrogen fixation from the Sargasso Sea database and found that although many of the genes in the *nif* operon could be identified in this database, there were still a number that could not. The relatively low abundance of identified *nif* genes in this ecosystem was indicative of the relatively low abundance of diazotrophic bacteria within the whole population. This highlighted the problem of coverage associated with metagenomics, wherein we see only what is abundant because, presently, metagenomic projects, even those as expansive as GOS, still produce far less than 1% coverage of the DNA in an ecosystem. Given the advances in our understanding of bacterial processing of organic phosphorus and sulfur, one remarkable failure of metagenomics has been the lack of insight into the processing of the dissolved organic nitrogen found in seawater.

GOS: EUKARYOTES

Most of the sequencing effort to date in the GOS expedition has focused on the smallest fraction collected (0.1–0.8 μm), which would exclude most eukaryotes. Despite this, a few groups have recently delved into this data set to analyze eukaryotes. Piganeau et al. (2008) identified GOS contigs (small assemblies of multiple, overlapping reads) of eukaryotic origin by examining the phylogenetic affinity of 10 phylogenetic marker genes where they occurred. Even though only 41 eukaryotic contigs were discovered, they represented five of the six supergroups of eukaryotic phylogeny, reinforcing the incredible diversity at the most fundamental phylogenetic scale. Not et al. (2009) compared the diversity of 18S sequences within the GOS data set with clone libraries constructed via PCR (rDNA) or RT-PCR (rRNA) and found significant differences between each data set. Although an understandable approach, these studies focused only on reads and scaffolds containing a phylogenetic marker, and as each GOS metagenome is dramatically undersampled, this means that most of the sequence information available for analysis is discarded at the first pass.

AN ASIDE ON EUKARYOTIC MICROBIAL METAGENOMICS

Fossils of single-celled eukaryotes, or protists, have been found in rocks over 1.6 billion years old (Javaux et al. 2001, Knoll et al. 2006). Not surprisingly, given this evolutionary time frame and the potential for endosymbiotic events to result in a wholesale acquisition of genetic material, modern protists comprise the bulk of eukaryotic phylogenetic diversity and an astounding array of morphologies, physiologies, and ecological activities (Baldauf 2003, Falkowski et al. 2004, Caron et al. 2009). Photosynthetic protists account for roughly 75% of marine primary production, whereas heterotrophic or mixotrophic protists are the major contributors to bacterial, archaeal, and eukaryotic mortality (Sanders et al. 1992, Sherr & Sherr 2002). Remarkably, small (<5-μm-diameter) organisms account for the bulk of bacterivory (Zubkov & Tarran 2008). Yet, the ecological role and functional traits of specific lineages of aquatic microbial eukaryotes remain poorly understood. Marine eukaryotic communities are now emerging as an important topic in marine and environmental sciences and biogeochemistry (Caron et al. 2009), and sequencing-based approaches are certain to be of particular significance.

The application of metagenomic techniques to marine microbial eukaryotes has lagged behind the efforts applied to prokaryotic communities. The foremost reason for this has been the carefully planned choice of the researchers involved; most marine metagenomic sampling protocols include a prefiltration step to remove larger organisms. The GOS prefilters originally used a 20-μm

(now 200-μm) plankton net prior to serial filtration onto 3.0-, 0.8-, and 0.1-μm pore size filters. However, the vast majority of sequencing efforts has focused on the 0.1-μm filters, with sequencing from only five 0.8-μm filters and one 3.0-μm filter presently deposited in the CAMERA database.

One of the underlying reasons why scientists using metagenomic tools have forsaken the eukaryotes is the cost required compared with that for prokaryote-focused projects. Even the smallest free-living eukaryote, the photosynthetic prasinophyte *Ostreococcus*, has a genome five times larger than that of an average marine bacterium (Derelle et al. 2006, Palenik et al. 2007). At the more extreme end of the spectrum, some dinoflagellate genomes appear to be much larger than the human genome (Hackett et al. 2005). Confounding this issue, eukaryotic genomes are far less gene-dense than those of bacteria and archaea, meaning that equivalent sequencing efforts will yield much more information for prokaryotes.

Another hurdle is the combination of the inherent diversity of the eukaryotic superkingdom, the lack of reference genomes, and the phylogenetic complexity of eukaryotic genomes. There are six hypothesized phylogenetic supergroups of Eukarya (Lane & Archibald 2008), five of which have been found to be present in seawater based upon 18S surveys (Massana & Pedrós-Alió 2008, Piganeau et al. 2008). Among these supergroups, there is a massive bias in the availability of completed reference genomes toward the Opisthokonta (fungi, metazoans) and Archaeplastida (plants, green and red algae). This bias means that most of the sequences acquired from the Chromaveolata or Rhizaria within a marine microbial community will either have no similarity to any other sequence or may even be misassigned. A further complication of this lack of reference genomes is the inherent genomic complexity that has been introduced by endosymbiotic events, whereby entire genomic complements of organisms from different supergroups are recombined into a new nuclear genome. For example, the dinoflagellate *Karenia brevis* derived its plastid from a tertiary endosymbiotic event involving a haptophyte (Van Dolah et al. 2009), and metagenomic sequencing of a bloom would yield sequences that would appear to be haptophyte in origin (provided the lack of a *K. brevis* reference sequence).

Most molecular work investigating environmental protistan ecology to date has been limited to taxonomic diversity surveys through use of amplified 18S ribosomal RNA libraries, which has revealed unexpected diversity and clarified some of the phylogenetic relationships of protists in the environment (Massana & Pedrós-Alió 2008). Even metagenomic studies targeting phylogenetic marker genes are potentially impeded; eukaryotic ribosomal DNA sequences are more often found in tandem repeats in genomes, and polymorphisms within these would result in a overassessment of organismal diversity within a clone library. Some eukaryotic groups also have long and/or GC-rich rDNA, both of which are biased against by traditional general-eukaryotic SSU rDNA primers (Liu et al. 2009). Even for the studies that have occurred, little morphological or physiological data can be inferred for the novel and abundant organisms, as most of them are not in culture.

STUDIES SINCE 2008

Using Metagenomics as a Tool

With falling sequencing costs and the aforementioned large-scale studies like Venter et al. (2004) and DeLong et al. (2006) providing the proof of concept and a template for analysis, metagenomic techniques have become accessible to individual laboratories or small collaborations. There have been many marine metagenomic studies since January 2008 using a variety of techniques to acquire the gene profiles from a variety of different environments. One of the wonderful aspects of these studies has been the diverse use of selective techniques to target specific organisms or hypotheses. These sorts of techniques highlight how "omics" techniques have become a tool for studying

ecological or evolutionary questions. What follows is a comprehensive review of the majority of these studies and what they have told us about marine microbial ecology. If we have missed some in error then we apologize to the colleagues responsible for those studies. In the interest of compartmentalizing this review, we have broken the studies into broad ecosystem definitions, within which we will deal with the contribution of large-insert-library studies and next generation sequencing studies.

Coastal Pelagic Ecosystems

A coastal pelagic ecosystem is one that is commonly defined as a coastal shelf sea rather than by the geopolitical definition, which gives a specific distance from shore (~5 km). Only one major study using large-insert libraries has been performed since January 2008: Neufeld et al. (2008) utilized stable isotope–probing techniques with ^{13}C-labeled methanol to examine the organisms metabolizing this substrate in the surface waters of the western English Channel L4 sampling site. They used multiple displacement amplification (phi29-mediated rolling circle amplification) to amplify the picogram quantities of ^{13}C-DNA acquired to the microgram quantities required for the construction of a 10,000-clone fosmid library. This library was screened for taxonomic markers that demonstrated that the dominant group involved in methanol metabolism in this ecosystem were most closely related to the *Methylophaga* genus. However, these were not dominant taxa; in fact, the gene (*mxaF*) encoding methanol dehydrogenase isolated from an individual fosmid clone showed relatively low similarity to known marine homolog in the nr database in GenBank, which the authors conclude is evidence of the low relative abundance of the *Methylophaga* genus in marine ecosystems.

In a groundbreaking study, Mou et al. (2008) coupled immunocapture with sequencing to examine microbes actively responding to the presence of dissolved organic carbon substrates, namely, DMSP or vanillate. Microbial communities were isolated from the surface waters off the coast of Sapelo Island, Georgia, in the United States. Two separate 20-l mesocosms were then supplemented with either DMSP or vanillate (100 nM) and bromooxyuridine (BrdU, 10 μM). Following a 12-h incubation, any new replicated DNA would contain BrdU instead of thymidine, facilitating immunocapture. This resulted in an enrichment of the organisms that responded to the pulse of dissolved organic carbon (DOC). Immunocaptured DNA was pyrosequenced, producing ~300,000 reads, which demonstrated significant overlaps in the types of genes and bacterial phyla enriched by this process. Specifically, a dramatic increase in the relative abundance of the Alteromonadales and Oceanospirillales was observed, suggesting that these groups are important in DMSP and vanillate metabolism. In the DMSP mesocosm, both demethylation and cleavage enzymes were found in abundance (e.g., dmdA, dddD, dddL, etc.). Mou and colleagues suggest that lack of evidence for specialization in the metagenomic gene profile between DMSP and vanillate enrichments provides evidence of generalism in the microbial population in this ecosystem in response to heterogeneity in the normal supply of DOC. They go on to suggest that predation or physical disturbance may be more responsible than DOC quality in the dynamic fluctuations observed in pelagic communities, although with only one time point, such a conclusion must be treated cautiously.

Palenik et al. (2009) used another enrichment strategy to increase the relative abundance of cells from the genus *Synechococcus* from surface water near the Scripps Institution of Oceanography pier in La Jolla, California. Flow cytometry sorting was used for cells with the size and fluorescent characteristics of this cyanobacterium, from which DNA was extracted and pyrosequenced, producing 370,000 DNA sequences. Much of the enriched sequence database could be aligned to the sequenced genomes of *Synechococcus*, including two genomes isolated from the same environment.

Whereas many genes shared significant homology, select areas of the sequenced genomes had little homology to the enriched sequence libraries, implying that these genomic locations were genetic hot spots with diverse contents within natural populations. This is similar to the result observed for SAR11 and *Prochlorococcus* ecotypes in the GOS data set (Rusch et al. 2007, Coleman et al. 2006). The authors suggest that horizontal gene transfer and mobile genetic elements, specifically plasmids assembled from the metagenomes, could explain some of the vast diversity of this genus.

Woyke et al. (2009) also used flow cytometry to isolate a population but focused instead on Flavobacteria. After sorting, multiple displacement amplification was used to amplify the genomes of two individual cells, with subsequent shotgun sequencing and assembly. The two Flavobacteria genomes were dramatically different from those of cultured Flavobacteria, being remarkably streamlined but far more representative of the Flavobacteria found in the GOS data set. Metabolic reconstruction suggests that these organisms are specialized at the incorporation of organic rather than inorganic carbon, nitrogen, and sulfur, potentially using proteorhodpsin to fuel uptake.

Gilbert et al. (2008a) examined the genetic diversity and gene expression in a large-scale (11,000-l) coastal mesocosm on a flotilla off the coast of Bergen, Norway. Whereas the overarching aim of the mesocosm experiment was to determine the impact of ocean acidification on marine microbes, the initial report detailed the production of 1 million metagenomic sequences and half a million metatranscriptomic reads from four samples, producing more than 300 million base pairs. This intersection of genomic and transcriptomic data sets revealed a wealth of unassignable transcripts, highlighting the immense gap in knowledge that exists for protein-encoding gene annotation in marine ecosystems. The data sets also demonstrated that there was an increase in the abundance of genes associated with carbohydrate and amino acid metabolism following the collapse of an induced phytoplankton bloom, which corresponded to an increase in relative α-proteobacterial abundance, suggesting that in this semi-artificial environment, the Alphaproteobacteria respond significantly to the release of nutrients following the collapse of a bloom. In a related study, Gilbert et al. (2009) used a combination of fosmid libraries from the western English Channel sampling site, L4, and the metagenomic data from the above-mentioned study to demonstrate the ubiquity and importance of phosphonate-degrading bacteria in coastal marine ecosystems. Their study suggested that despite an abundance of inorganic phosphate, numerous microorganisms were actively using the organic phosphonate fraction, potentially as a niche diversification strategy to avoid competition or as a means to make use of this vast resource.

Oxygen minimum zones, or so-called dead zones, are oceanic layers where total dissolved oxygen is drawn down to <20 μM. The lack of the most energetic electron acceptor results in the reduction of compounds like nitrate, which has a major influence on the global nitrogen cycle. Walsh et al. (2009) constructed a fosmid library from an oxygen minimum zone in Saanich Inlet, British Columbia. Sequencing of fosmids from the uncultured SUP05 γ-proteobacterial lineage revealed an extensive metabolic repertoire, including the ability to oxidize multiple sulfur compounds and reduce nitrate to fuel autotrophic carbon assimilation via the Calvin–Benson–Bassham cycle.

Marine Hydrothermal Vents

Since 2008, there has been only one significant metagenomic study on material isolated from hydrothermal vents, that of Brazelton & Baross (2009). In this study, DNA was isolated from a biofilm growing on the mineral surfaces of the highly porous carbonate chimneys from the Lost City Hydrothermal Field on the mid-Atlantic ridge. These chimneys vent hydrogen and methane-rich fluid at <90°C and a pH of 9–10. Hence, this is a true extreme environment. The biofilms are

dominated by a single phylotype belonging to the *Methanosarcinales*, which constitutes >80% of the community. The DNA was cloned into pUC18 vectors and then randomly end sequenced, similar to the GOS strategy, producing 35 Mbp from 46,316 sequences. The primary finding from this metagenome was that the community gene profile had a considerable abundance of transposases, >8%, which was tenfold higher than any other compared metagenome (the next highest was from bioreactor sludge; Garcia Martin et al. 2006). This not only highlights the importance of marine metagenomic studies from the many diverse environments that constitute marine ecosystems but also the importance of mobile genetic elements in this extreme ecosystem. This study suggests that the majority of the transposases were found on small but abundant extragenomic elements, which could be responsible, as the authors suggest, for rampant horizontal gene transfer in this ecosystem, a potential adaptation mechanism for the dominant population.

Marine Sediments

Two recent studies have focused on marine sediments using cloning strategies. First, Huang et al. (2009) produced a 40,000-clone fosmid library from sediment isolated from 1.2-, 1.3-, and 2.9-km depths in the South China Sea. Clones were screened for their ability to alter the phenotype of the host organism, *E. coli*. One particular clone was identified that produced melanin, and this clone was fully sequenced. The genomic fragment was identified as being most closely related to the γ-proteobacterium *Idiomarina loihiensis*, and further comparative analysis with known genomes of this group suggested that the organisms in question may likely derive their carbon and energy from the metabolism of tyrosine. Second, Kim et al. (2009) produced a fosmid library from intertidal flat sediments of the coastal regions of Saemankum, located in the west of South Korea. Whereas the aim of this study was to identify lipase encoding genes, the gene identified encoded a particularly unique enzyme, with a broad range of pH and temperature tolerance. The authors suggest that the highly variable tidal flat environment supports organisms whose enzymes are capable of functioning in an extreme range of conditions. Such directed studies are more amenable to the identification of biotechnologically relevant genes, rather than providing further information of ecological relevance.

Open Ocean

Two recent studies have used metagenomics to determine microbial functional and taxonomic diversity in open ocean ecosystems. First, in a study similar to that of the coastal study from Gilbert et al. (2008a), and combining metagenomics with metatranscriptomics, Frias-Lopez et al. (2008) demonstrated the value of producing a metagenomic sequence database complementary to a metatranscriptomic database, highlighting the power of multi-omic data sets. DNA was extracted from a water sample taken from the HOT sampling site in the North Pacific. Ecologically, they demonstrated that Cyanobacteria and unknown bacterial taxa contributed the largest fraction of gene transcripts. A further highlight was the large number of transcripts encoding for genes found in the hypervariable regions of the cyanobacterial genomes, reinforcing the idea that these fine-scale genomic variations are critical to niche differentiation.

Hewson et al. (2009) presented an exciting biogeographic study of the surface waters of the open ocean at seven locations between the North Pacific and South Pacific subtropical gyres. They produced 1.1 million pyrosequence reads, which when annotated demonstrated highly significant differences in the gene profiles found between samples in gyre ecosystems, countercurrent habitats, and equatorial environments. Again, the Proteobacteria and Cyanobacteria dominated in all ecosystems, but the relative abundance of *Synechococcus* and *Prochlorococcus* varied significantly

between different samples. The patterns within cyanobacterial species composition matched those expected from the physiological potential suggested by the individual genomes. Despite the changes in community composition, the metabolic characteristics of each ecosystem were very similar, suggesting that the same niches exist but that specific parameters drive the selection of specific taxa in different habitats. As expected in locations with extremely low inorganic phosphate concentrations, genes involved in phosphate scavenging and accumulation had a greater relative abundance within the profile. Interestingly, phosphonate utilization–encoding genes were at a relatively equal abundance throughout the transect, suggesting that this was a core function for some members of the community that existed in all ecosystems and was not necessarily affected by a decrease in the availability of inorganic phosphate.

In another example of the utility of selective enrichment prior to metagenomic sequencing, Tripp et al. (2010) used paired-end shotgun sequencing to assemble a complete genome of a unicellular nitrogen-fixing Cyanobacterium from the UCYN-A clade. Lacking a complete tricarboxylic acid cycle and photosystem II, UCYN-A appears to be a photoheterotroph dependent upon other organisms for vital compounds like several amino acids.

Dead Sea

Although technically not marine, a recent study looking at the metagenomic diversity of samples collected during a microbial bloom and standard present-day brine conditions in the Dead Sea demonstrates the capability of the techniques to expand our understanding of the ecology of an ecosystem. Bodaker et al. (2010) produced two fosmid libraries from samples collected during 1992 (significant microbial bloom condition) and 2007 (normal brine conditions). As might be expected, the bloom condition was less diverse, with several dominant archaeal taxa. Specifically, the metabolic potential of the 2007 sample was very similar to that of a previous metagenomic sample from a Spanish saltern in terms of genes associated with divalent cation antiporters and transposable elements (the latter was also seen in the hydrothermal vent biofilm; Brazelton & Baross 2009). These two functions are potential adaptive mechanisms for this environment. Studying such extreme ecosystems can help us to unravel the genomic potential of more mesophilic systems such as pelagic marine samples.

Host-Associated Communities

Many marine microbes are associated with particles and/or other organisms. A number of recent studies have focused on these associated communities in organisms such as corals and sponges. For example, Dinsdale et al. (2008) produced a pyrosequencing-derived metagenomic study of coral-associated microbial communities from four different coral atolls across a gradient of anthropogenic influence. In those systems with a greater degree of human influence, the corals were less healthy and the microbial metagenomic functional profiles were dominated by heterotrophic processes.

Another coral study (Thurber et al. 2009) took a single species and its holobiont (associated microbial community), exposed it to four stressors (temperature, nutrient loading, DOC loading, and reduced pH), and produced pyrosequenced metagenomes from the resulting holobiont commmunities. The microbial community of the stressed corals exhibited an increase in genes associated with virulence, stress resistance, sulfur and nitrogen metabolism, motility and chemotaxis, fatty acid and lipid utilization, and secondary metabolism. Additionally, taxonomic analysis of the communities demonstrated a shift from a healthy holobiont to one dominated by Bacteroidetes, Fusobacteria, and Fungi, all indicative of a diseased state in corals. Additionally, the metagenomic

profile was significantly altered by low-abundance *Vibrio* spp., suggesting that this group of known coral pathogens is potentially opportunistic as a result of the stressed coral state.

Comparative Metagenomic Pyrosequencing Studies

Recently, two key studies have involved the comparison of multiple pyrosequenced metagenomes to determine the relationship between environmental parameters and taxonomic and functional profiles. In truth, these are comparative biogeographic studies that take advantage of the extensive buildup of metagenomic data sets to date. The first study by Dinsdale et al. (2008) compared 15 million sequences from 45 microbiomes and 42 viromes and demonstrated that there were strong metabolic differences between different environments. Importantly, the changes associated with each environment were the product of relative changes in the abundance of specific functions, and the prevalence of different functional groups could be used to predict the environmental conditions of each ecosystem. A follow-up study to this used pyrosequenced data from 86 microbial and viral metagenomes and analyzed the di-, tri-, and tetranucleotide coding frequency across different habitats (Willner et al. 2009). This again showed distinct profiles driven by broad ecosystem descriptors such as marine, freshwater, and so forth; however, 80% of the variance could be described by the dinucleotide coding frequencies alone. Two hypotheses are given for this bias: The ecosystems select proteins with specific characteristics encoded by specific dinucleotide frequencies, and the ecosystems select specific taxonomic groups that can be shown through comparative genomic analysis to have taxa-specific dinucleotide coding frequencies.

One major concern for comparative metagenomics involves the possible experimental bias introduced when different methods are used. Morgan et al. (2010) examined the community structure of in vitro simulated communities of 10 well-known Bacteria using different DNA extraction and sequencing techniques. Remarkably, different extraction protocols resulted in very different metagenome-based estimates of the original community structure, whereas Sanger and 454 sequencing approaches generated comparable results. A sobering conclusion is that metagenomes generated using different DNA extraction protocols are unsuitable for comparative analyses.

Metagenomic Studies of Eukaryotes

Liu et al. (2009) used specialized PCR primers designed to amplify the haptophyte 18S RNA prior to clone library construction. They found that the vast majority of marine haptophytes have no cultured representative and that they are likely noncalcifying. The abundance of these "nude" haptophytes explains an historical discrepancy between cell counts of haptophytes and the abundance of the diagnostic pigment, 19′-hexanoylfucoxanthin. Shi et al. (2009) collected natural populations of chlorophyll a–containing picoeukaryotes using flow cytometry and found that most of the 18S sequences were derived from uncultured organisms, a remarkable finding considering the availability of several prasinophyte genomes. In some ways, this scenario eerily mirrors that facing microbiologists studying prokaryotes in the late 1990s.

Massana et al. (2008) identified several fosmids of eukaryotic origin through an analysis of SSU rDNA sequences and completely sequenced a 35.5-Kbp fosmid from an uncultured marine aveolate. The fosmid contained a tandem repeat array of three rRNA genes, the exact genomic entity that has been evoked as a confounding influence on estimates of diversity based on clone libraries of rDNA. Crucially, the lack of a single nucleotide polymorphism within matching elements of the repeat array suggests that the remarkably high diversity found in the rDNA clone libraries represents actual organismal rather than intragenomic diversity in natural populations.

Cuvelier et al. (2010) used flow cytometry to isolate eukaryotic picoplankton from the Florida Straits in the North Atlantic. Whole-genome amplification, followed by both Sanger and 454 sequencing, was used to study the genomic characteristics of uncultured and environmentally abundant picoprymnesiophytes. The sorted picoprymnesiophyte genome is gene-dense and codes for a modest protein repertoire (11,000–13,000 total genes) of mosaic evolutionary origin. The picoprymneisiophyte genome has far more proteins of likely Archaeplastidia origin than was expected for a Chromaveolate, suggesting a prominent role for Archaeplastidia in the endosymbiosis-driven history of the Chromaveolates. This mosaic evolutionary history highlights a major challenge in studying eukaryotes with shotgun metagenomics.

Bioinformatic Resources and Requirements

Large-scale sequencing projects such as GOS (Venter et al. 2004, Rusch et al. 2007) and the HOT-ALOHA (A Long-term Oligotrophic Habitat Assessment) fosmid database (Delong et al. 2006) really brought to our attention the need for high-throughput computational techniques to cope with the analysis of millions of sequencing reads. Whereas a traditional study may have produced 2–3 thousand sequences that would often be aligned and annotated using programs with considerable user input, the analysis of the 7 million sequences from the GOS expedition (Rusch et al. 2007) meant that we had to find a way to trust the programs to do the job, and to analyze the output of these programs rather than scrutinizing the whole process every time.

The development and rapid improvement of second generation sequencing technologies, for example, 454-pyrosequencing and the Illumina platform, meant that for the first time labs with comparatively small budgets could produce immense-scale metagenomic sequencing projects. There is no doubt that access to these platforms is revolutionizing metagenomics through the production of more sequencing data than was ever possible with Sanger analysis. Yet, the continued development of more capacity and faster technologies with longer read lengths is producing an analysis bottleneck, whereby it is the bioinformatic software and computational capacity that now limit our analytical capability. For a comprehensive description of the bioinformatic capabilities and programs presently available, please consult Wooley et al. (2010).

Presently, a standard sequence analysis pipeline (**Figure 3**) will provide a number of different analytical routes. For example, you can take the raw metagenomic sequences, which will be of different sizes depending on the sequencing technology used (e.g., 454, 500, 800 bp; Illumina: 200–300 bp; SOLiD: 50–75 bp; HeliScope, Complete Genomics, and SMRT: variable), and annotate them directly against a database of annotated genes (see the sidebar A Word of Caution on Annotation, below) such as the National Center for Biotechnology Information's nucleotide database nt. Alternatively, you can first cluster the DNA sequences using cd-hit (Li & Godzik, 2006), allowing you to group together sequences with similar sequence homology (e.g., 95% nucleotide identity); a representative from each group can then be annotated to a known function or compared between multiple metagenomic data sets to determine distribution. The most appropriate way of screening out the potentially uninformative sequences is to use gene-prediction software to determine the likelihood of a DNA sequence being a gene or not [e.g., orf-finder (code written by Weizhong Li, weizhong@ucsd.edu). Orf-finder uses a rule of thumb that identifies an ORF as being between the beginning of a sequence and the end of a stop codon, or between a start codon and the end of the sequence or a stop codon, and being of a certain prespecified length (e.g., 40 amino acids)]. This is very useful for next generation sequencing reads (e.g., Illumina) because it is far less likely to have sequence motifs such as ribosome binding sites on the fragment that can be used to reliably predict a gene, which is more akin to MetaGene (unfortunately now unsupported software but still available at **http://metagene.cb.k.u-tokyo.ac.jp/metagene**). There is

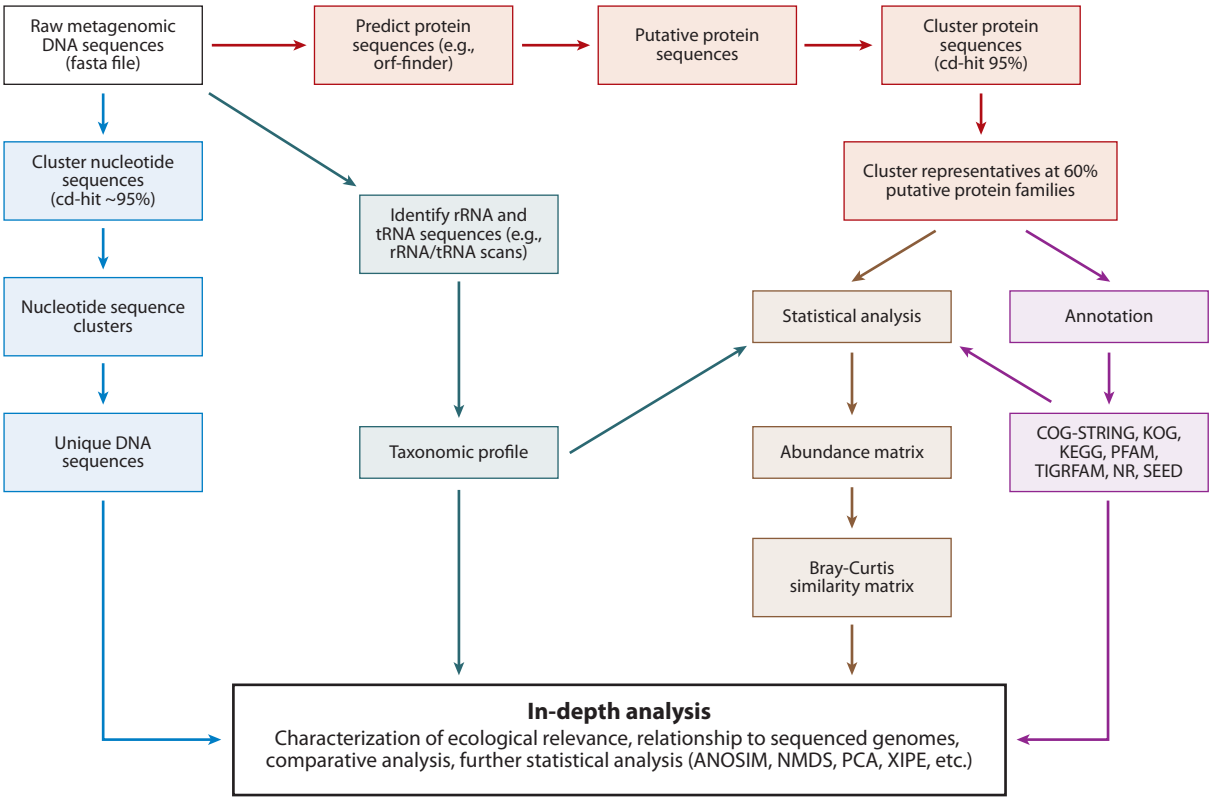

Figure 3

Standard route to market, from raw metagenomic reads to ecosystem analytical capability. Presently used computational programs are also shown, as well as protein databases PFAM, TIGRFAM, and COG. Reproduced with the kind permission of Jeffrey S. Grethe. Abbreviations: ORF, open reading frame, a potential protein coding sequence; rRNA, ribosomal RNA; tRNA, transfer RNA.

a vast array of other software and more in development that will improve on this method. Once we have predicted proteins, we can again cluster and annotate these, or annotate the raw protein sequences but using different databases, such as TIGRFAM, PFAM, or COG (**Figure 3**).

Once a metagenome has been annotated or clustered, it is necessary to use this data to test the hypothesis, which was of course the reason you produced the metagenome in the first place! To do this effectively, you will often need to have (*a*) an appropriate way of visualizing this data, such as sequence hits to a KEGG metabolic map or a pie chart of functional hits, and/or (*b*) the contextual data from the sample from which the metagenome was derived, for example, MIMS standard environmental information (Field et al. 2008). Several online tools have become mainstays of metagenomic analysis to help reach these goals; the most comprehensive at this time are the MG-RAST annotation platform (Meyer et al. 2008), which enables you to annotate and compare metagenomes from a list of private or publicly accessible precomputed analyses of various metagenomes, and the CAMERA 2.0 portal (**http://www.camera.calit2.net**), which provides a framework to design and implement your own data analysis pipeline. Other more recent additions include the Galaxy pipeline, which has over 100 tools for interrogating your sample, including statistical packages for making biological inferences about sequence patterns between samples (Kosakovsky Pond et al. 2009).

A WORD OF CAUTION ON ANNOTATION

Annotation is a very human concept; we feel comfortable labeling sequences to contextualize them within a framework of our understanding. The ideal is to provide an unknown sequence with an established nomenclature based on previous experimental analysis of a sequence with similar sequence homology. However, there are two potential problems with this. First, since the inception of PCR and sequencing of clone libraries, there have been far more sequences entering public databases with no experimentally determined function than those that have been proved experimentally to code for a protein with that function. Hence, the first known sequence was used to annotate the first unknown sequence, but the first unknown sequence will have been used to annotate the second unknown sequence; this continues and continues until we are, more often than not, ascribing function or taxonomy to a gene based entirely on homology with tenuous evidence. An excellent example of the power of experimentally proven function improving our interpretation of metagenomic data comes from the identification and functional characterization of the DMSP enzyme–encoding gene *dmdA* (Reisch et al. 2008).

The second problem occurs as a result of a lack of linearity between function and sequence homology. A well-established principle of structural bioinformatics is that proteins with proven, identical functions and extremely similar three-dimensional structures may share less than 20% sequence identity (Bourne et al. 2010); thus, sequence clusters that have identical function may appear to be disparate. The only way to alleviate these issues is through faster and more comprehensive association of a sequence to an actual protein function, or by the application of faster and computationally less expensive methods for in silico protein folding and, hence, structural comparability analyses to overcome issues of ascribing function based entirely on sequence homology.

In all cases, bioinformatics annotation of any type is far better for producing a profile of function from millions of sequences than it is at describing actual function for a specific sequence. The limitations of annotation are threefold: First, as highlighted above, a sequence may have no sequence or structural homology to known proteins and as such remains unknown; second, homology to a known protein can be correctly assigned, but unknown to the interpreter, a new function has evolved in the gene family (neofunctionalization); and third, when at last it is possible to identify homology and assign the correct function to a metagenomic sequence, that function may not coexist in any metabolic context, making accurate interpretation of its ecological significance extremely unlikely.

HOW WHAT WE HAVE LEARNED CAN SHAPE METAGENOMIC STUDIES

The time for proof of principle is over. In the last two years, there has been an increasing number of hypothesis-driven studies that have used metagenomic analysis as a tool. Importantly, modern high-throughput metagenomic techniques have the capability to characterize the diversity and ecological function of microbial communities in a way never before imagined. This is driven by the depth of perspective provided by the volume of sequencing. It is vital that we use the scientific method in an incremental way and apply metagenomic tools to answer specific questions. For example, a latitudinal gradient in plant and animal diversity has been a central tenet of ecology since the early nineteenth century, but the lack of morphological distinction in microbes prevented an analogous study. In 2008, Fuhrman et al. (2008b) produced a revolutionary biogeographic analysis of microbial diversity through a transect of the globe from north to south, and using statistical analysis of the community composition, they were able to demonstrate that latitude was the major factor influencing the composition of the microbial communities, probably driven by temperature and nutrient availability. Additionally, the elegant study by Schattenhoffer et al. (2009) demonstrated similar properties in distribution for the prokaryotic picoplankton.

To improve this analysis, we need to perform a directed global transect, whereby over a period of two years a similar marine community—that is, open ocean throughout the Pacific—is subject to monthly metagenomic analysis to determine the seasonal and interannual variability in the communities at each location, in essence, a combined temporal and biogeographical analysis of microbial communities. Such a study should be accompanied by a full suite of environmental data at each site. International efforts such as the TARA-OCEANS (**http://oceans.taraexpeditions.org**) study, which aims to travel the globe producing ecologically relevant metagenomic analyses, are one step; however, as each sample is taken at a different location and at a different time, it will be difficult to disentangle biogeographic effects from seasonality in the communities. A more reductionist study would use high-throughput metagenomics to determine the true spatial heterogeneity in a specific sediment basin, for example, an enclosed shallow coastal bay. A systematic coordinated collection of thousands of samples in a statistically designed, temporally restricted study would produce a comprehensive map of microbial function and diversity across a local-scale sediment area. If taken with appropriate metadata, this study could have far-reaching implications for our understanding of wide-scale sediment diversity analysis. A third and more reductionist study would identify protein structures from intensively sequenced metagenomic projects. Ubiquitous protein clusters with unknown function are excellent targets for high-throughput protein folding computation (e.g., Rosetta; **http://boinc.bakerlab.org/rosetta/**) and for directed high-throughput automated protein crystallization and X-ray characterization of protein folding structures (e.g., Babnigg & Joachimiak 2010). This is an essential area of study if we wish to contextualize future metagenomic databases but one that demands constant and considerable resources, and hence, it is unlikely to move forward significantly unless a concerted global effort is applied.

Technical advances and creative sampling will continue to shape the field of metagenomics. A promising approach involves the coupling of selective collection techniques such as flow cytometry sorting and either shotgun sequencing or multiple displacement amplification single-genome sequencing. As shown by several studies, this approach has the potential to enrich a sample for sequences from one to several organisms, allowing for a better bioinformatic analysis or even assembly after sequencing (Woyke et al. 2009, Palenik et al. 2007). Assembly refers to the reconstruction of genomes or parts of genomes from shotgun sequencing. With robust assemblies, genome contents and physiological traits can tentatively be associated with specific organisms. Although a variety of methods exist (Pop 2009), each depends to some degree on sequences overlapping to an extent that alignment can be made. Previous attempts to produce adequate reconstructions of microbial genomes from metagenomics data (e.g., *Burkholderia* and *Shewanella* genomes from the Sargasso Sea data; Venter et al. 2004) suffered from a misunderstanding of the level of genetic variability found in marine microbial genomes. Within a single species, genomes can exhibit considerable vairability, notwithstanding hypervariable regions as identified in the SAR11 clade (Wilhelm et al. 2007). The concept of the pangenome (Medini et al. 2005) may explain this variability, which acts to provide a single population with enough variability to adapt to environmental conditions. As previously mentioned, even deep-sequencing approaches sample only a small fraction of the genomic variability available in a natural environment. This undersampling, along with the progression toward the shorter read lengths of next generation sequencing techniques, provides a substantial impediment to a robust assembly. Some approaches have shown promise (Rusch et al. 2007, 2010), and the development of bioinformatics techniques for the assembly of next generation sequencing data is a current research focus in bioinformatics (Pop 2009, Kingsford et al. 2010). Advances in metagenomic assembly will have the direct benefit of facilitating the linkage of function and phylogeny through an analysis of the assembly itself, but the assembly also becomes a scaffold for examining genetic variation and environmental distribution

through fragment recruitment of different metagenomes to the assemblies (Rusch et al. 2007, 2010; Woyke et al. 2009).

It is useful to consider the two methods for analyzing metagenomic DNA in more detail. The rival concepts of gene-centered versus organism-centered analysis are at the core of all metagenomic analyses. For example, the vast majority of GOS-derived studies has focused on the detection of either keystone genes or patterns, usually skipping metabolic pathway reconstruction, to infer diversity or distribution of a particular taxonomic marker or functional gene (e.g., Sebastian & Ammerman 2009, Todd et al. 2009). The organism-centered approach is far more akin to systems biology, whereby reconstruction of specific metabolic pathways from metagenomic data leads to interpretations as to the fucntional potential of an ecosystem. One important way to achieve this is to group pathways based on their host genome, either through the assembly of metagenomic DNA as described above or via the binning of metagenomic sequences against known sequenced genomes. As the number of sequenced genomes increases through the efforts of the Genomic Encyclopedia of Bacteria and Archaea (GEBA; Wu et al. 2009), we will see a concomitant rise in our ability to interpret metagenomic information from an organism-centered perspective.

Metatranscriptomics, or the sequencing of cDNA (also known as expressed sequence tag libraries, or RNA-seq) derived from community RNA, has an attractive future for eukaryotes in particular. Because eukaryotic gene coding RNA can be selectively converted to cDNA through the use of poly-dT primers, a study can focus on only the coding portions of a eukaryotic genome but also avoid the overabundant rRNA sequences present in samples. Theoretically, this provides information on the community composition through phylogenetic affinity of the sequence and on the environmental physiology of each component in the community. To date, most expressed sequence tag (EST) studies have focused on single cultures of eukaryotes, but the potential for community studies is immense. Highlighting this, a recent analysis of eukaryotic sequences from a transcriptomic library focused on prokaryotes revealed a large number of transcripts of diatom origin for carbamoyl phosphate synthase III (Poretsky et al. 2009). This enzyme is the keystone of the urea cycle, believed previously to be a metazoan-only biochemical entity surprisingly discovered in diatoms (Allen et al. 2006, Parker et al. 2008). Although the exact function of this cycle remains unknown, transcriptomics have bolstered the suspicion that it is important to the ecophysiology of a phytoplankton responsible for every eighth breath of oxygen.

Although the sequencing of the genomes of aquatic eukaryotic microbes has generally lagged, the future looks much brighter. At the moment, pending genome sequences for marine protists include a calcifying haptophyte (*Emiliania huxleyi*), two noncalcifying haptophytes (*Phaeocystis antarctica* and *P. globosa*), two diatoms (*Fragillariopsis cylindrus* and *Pseudonitzschia multiseries*), four labyrinthulids (*Aurantiochytrium limacinum*, *Aplanochytrium kerguelense*, *Labyrinthula terrestris*, *Schizochytrium aggregatum*), two chrysophytes (*Ochromonas sp.* CCMP1393, *Paraphysomonas imperforata*), and a pelagophyte (*Aureococcus anophagefferens*). With the completion of each genome, we suspect that a portion of the publicly available metagenomic data will gain some phylogenetic and potentially physiological context. Further, the availability of these genomes facilitates functional studies, which in turn allows for a more robust annotation of metagenomic data.

FROM A QUALITATIVE TO A QUANTITATIVE METAGENOMICS

To date, all metagenomic studies have focused on how to interpret data from a descriptive perspective. This is essentially a qualitative analysis: Who is there and how do they change between ecosystems? Whereas most of the comparisons are essentially based on the changes in relative abundance of different genes or taxa, this is not an effective quantitative report, as the comparison is based entirely on relative abundance rather than absolute values. A suite of statistical techniques

both new and old have been employed to cope with the interpretation of these qualitative data sets. Traditional techniques such as principal component analysis, nonmetric multidimensional scaling, and dendrogram clustering have been enployed to demonstrate how similar or different the profiles are between different metagenomes. Additionally, heat maps have become very popular for visualizing how genes or taxa change (Willner et al. 2009), whereas new visualization techniques are constantly being developed (e.g., Mitra et al. 2010). More often than not, metagenomic studies have very few samples with no replication, which is because of the cost of sequencing. This means that traditional approaches cannot be used to infer statistically significant differences. Therefore, techniques such as XIPE (Rodriguez-Brito et al. 2006) and STAMP (Parks & Beiko 2010) have been developed to provide a descriptive statistic to infer biologically relevant differences between samples.

As the price of sequencing falls to a level where statistical replication is technically and financially feasible, proper statistical analysis of metagenomic data sets will become more common. This will continue to improve the design and implementation of metagenomic studies, leading to further developments such as increased sequencing depth. Within the next few years, we will start to see the production of terabase metagenomic data sets (i.e., >1 trillion base pairs per sample), which will provdide unparalleled access to appropriate quantification of genes and taxa. Every marine metagenomic study to date has focused on water sample volumes of 1–1,000 l of seawater, which contain an approximate average of 5 quadrillion (1×10^{15}) to 5 quintillion (1×10^{18}) base pairs of DNA (based on an average of 1 million microbial cells per milliliter and an average genome size of 2 million base pairs). Based on sequencing effort in each study, we generally have provided only <0.000001% coverage of the DNA in a sample (the latest studies have produced between 500 million and 5 billion base pairs, often over a number of samples). By increasing the sequencing depth to 4–5 trillion base pairs, we will be able to increase coverage to 0.001% per litre of water, which is still a massive undersampling. Techniques are presently being developed, including improved sensitivity for high-throughput sequencing platforms, that will enable smaller quantities of DNA to be sequenced directly without prior amplification. Once we can extract and sequence the metagenome from 1 ml of seawater or 10 mg of sediment, it will be possible to sequence the entire metagenomic DNA, providing a coverage of 100% at a depth of 1—basically, sequencing every base pair once. Initiatives are under way to implement this approach over hundreds of thousands of sampling locations across the globe. The so-called Earth Microbiome Project aims to generate 10 quadrillion base pairs of metagenomic and metatranscriptomic data before 2014 from more than 160,000 sampling sites, including marine envrionments (R. Stevens, F. Meyer, and J. Gilbert, personal communication). It is hoped that these studies will be able to answer fundamental questions regarding the diversity and functionality of the global microbiome.

The ultimate goal of metagenomics is to provide a descriptive and eventually predictive metabolic and taxonomic model of an ecosystem. Terabase metagenomics tied to sample replication and appropriate statistics will provide the framework onto which we can start to build models of metabolite flux over time and space (Larsen et al., manuscript in review; Henry et al. 2010) with the ultimate goal of being able to describe the full interactome of a system, so-called systems biology. Once we have a near-complete understanding of how one element of a system impacts every other element in the system, either directly or indirectly, we can start to extrapolate out these concepts to derive predictive models of ecosystem changes. This is essential if we are to determine the role of climate change, ocean acidification, sea surface temperature rise, etc., on marine ecology. This is so important because it is our only way of predicting food availability, human health impacts, and eventually human migatory patterns. The latter is essentially an economic status model that is driven by the metabolic potential of the microbial ecosystem and the services it provides. These predicted changes in ecosystem services will help us to mitigate changes

in the dynamics of human populations. Essentially, we must look toward the smallest and most ignored among us to understand our future on this planet.

CONCLUSION

We have a long, long way to go. Science is incremental, and each one of these studies is providing another piece of the jigsaw that can hopefully help us to see the bigger picture. Metagenomics is allowing us to increase the resolution of that picture, but conversely, this also means that there are more jigsaw pieces to find! The last two years have witnessed a sequencing revolution, and there is no doubt that in the next two years, we will see a computational revolution to cope with the sequence data. However, just as the arms race drives evolution, it is also driving metagenomics. "Citius, Altius, Fortius," swifter, higher, stronger—the Olympic motto holds considerable resonance for our community; never before have we witnessed such a rate of development of technology in molecular biology, and the hypothesis-driven science may find it hard to keep up.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Allen AE, Vardi A, Bowler C. 2006. An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr. Opin. Plant Biol.* 9:264–73

Babnigg G, Joachimiak A. 2010. Predicting protein crystallization propensity from protein sequence. *J. Struct. Funct. Genomics* 11:71–80

Baldauf S. 2003. The deep roots of eukaryotes. *Science* 300:1703–6

Béjà O. 2004. To BAC or not to BAC: marine ecogenomics. *Curr. Opin. Biotechnol.* 15:187–90

Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, et al. 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289:1902–4

Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, et al. 2002. Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* 415:630–33

Bodaker I, Sharon I, Suzuki MT, Feingersch R, Shmoish M. 2010. Comparative community genomics in the Dead Sea: an increasingly extreme environment. *ISME J.* 4:399–407

Bourne PE, Briedis K, Dupont CL, Valas R, Yang S. 2010. Genome evolution. In *Evolutionary Genomics and Systems Biology*, ed. G Caetano-Anolles, pp. 153–64. Chichester, UK: Wiley-Blackwell

Brazelton WJ, Baross JA. 2009. Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *ISME J.* 3:1420–24

Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. 2009. Protists are microbes too: a perspective. *ISME J.* 3:4–12

Chen K, Pachter L. 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comp. Biol.* 1:24

Coleman ML, Sullivan MB, Steglich C, Delong EF, Chisholm SW. 2006. Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311:1768–70

Cowan DA, Arslanoglu A, Burton SG, Baker GC, Cameron RA, et al. 2004. Metagenomics, gene discovery and the ideal biocatalyst. *Biochem. Soc. Trans.* 32:298–302

Curson ARJ, Rogers R, Todd JD, Brearley CA, Johnston AWB. 2008. Molecular genetic analysis of a dimethylsulfoniopropionate lyase that liberates the climate-changing gas dimethylsulfide in several marine α-proteobacteria and *Rhodobacter sphaeroides*. *Environ. Microbiol.* 10:757–67

Cuvelier ML, Allen AE, Monier A, McCrow JP, Messie M, Tringe SG, et al. 2010. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci. USA.* 107:14679–84

Delong EF. 2005. Microbial community genomics in the ocean. *Nat. Rev. Microbiol.* 3:459–69

Delong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311:496–503

Derelle E, Ferras C, Rombauts S, Rouze P, Worden AZ, et al. 2006. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. USA* 103:11647–52

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. 2008. Functional metagenomic profiling of nine biomes. *Nature* 452:629–33

Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, et al. 2008b. Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE* 3:e1584

Edwards RA, Dinsdale EA. 2007. Marine environmental genomics: unlocking the ocean's secrets. *Oceanography* 20:26–61

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, et al. 2004. The evolution of modern eukaryotic phytoplankton. *Science* 305:354–60

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

Field D, Joint I, Gilbert JA, et al. 2008. Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. *Nat. Biotechnol.* 26:541–47

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, et al. 2008. From the cover: microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. USA* 105:3805–10

Fuhrman JA, Schwalbach MS, Stingl U. 2008. Proteorhodopsins: An array of physiological roles? *Nat. Rev. Microbiol.* 6:488–94

Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, et al. 2008b. A latitudinal diversity gradient in planktonic marine bacteria. *Proc. Natl. Acad. Sci. USA* 105:7774–78

Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, et al. 2006. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.* 24:1263–69

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO et al. 2009. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl. Acad. Sci. USA* 106:1374–79

Gilbert JA. 2010. Aquatic metagenome library (archive; expression) generation and analysis. In *Handbook of Hydrocarbon and Lipid Microbiology*, ed. K Timmis, pp. 4347–52. Berlin: Springer

Gilbert JA, Field D, Huang Y, Edwards R, Li W, et al. 2008a. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* 3:e3042

Gilbert JA, Mühling M, Joint I. 2008b. A rare SAR11 fosmid clone confirming genetic variability in the 'Candidatus Pelagibacter ubique' genome. *ISME J.* 2:790–93

Gilbert J, Thomas S, Cooley NA, Kulakova AN, Field D, et al. 2009. Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters. *Environ. Microbiol.* 11:111–25

Giovannoni SJ, Stingl U. 2005. Molecular diversity and ecology of microbial plankton. *Nature* 437:343–48

Gough J. 2006. Genomic scale subfamily assignment of protein domains. *Nucleic Acids Res.* 34:3625–33

Green BD, Keller M. 2006. Capturing the uncultivated majority. *Curr. Opin. Biotechnol.* 17:236–40

Grzymski JJ, Carter BJ, DeLong EF, Feldman RA, Ghadiri A, et al. 2006. Comparative genomics of DNA fragments from six Antarctic marine planktonic bacteria. *Appl. Environ. Microbiol.* 72:1532–41

Hackett JD, Scheetz TE, Yoon HS, Soares MB, Bonaldo MF, et al. 2005. Insights into a dinoflagellate genome through expressed sequence tag analysis. *BMC Genomics* 6:80–92

Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68:669–85

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5:245–49

Hardeman F, Sjoling S. 2007. Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. *FEMS Microbiol. Ecol.* 59:524–34

Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, et al. 2007. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc. Natl. Acad. Sci. USA* 104:13913–18

Henry CS, DeJongh M, Best AB, Frybarger PM, Linsay B, Stevens RL. 2010. Model SEED: a resource for high-throughput generation, optimization, and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28:977–82

Hewson I, Paerl RW, Tripp HJ, Zehr JP, Karl DM. 2009. Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. *Limnol. Oceanogr.* 54:1981–94

Huang Y, Lai X, He X, Cao L, Zeng Z, et al. 2009. Characterization of a deep-sea sediment metagenomic clone that produces water-soluble melanin in *Escherichia coli*. *Mar. Biotechnol.* 11:124–31

Hugenholtz P, Tyson GW. 2009. Metagenomics. *Nature* 455:481–83

Javaux EJ, Knoll AH, Walter MR. 2001. Morphological and ecological complexity in early eukaryotic ecosystems. *Nature* 412:86–69

Johnston AWB, Li Y, Ogilvie L. 2005. Metagenomic marine nitrogen fixation—Feast or famine? *Trends Microbiol.* 13:416–20

Karl DM. 2007. Microbial oceanography: paradigms, processes and promise. *Nat. Rev. Microbiol.* 5:759–69

Kennedy J, Marchesi JR, Dobson ADW. 2007. Metagenomic approaches to exploit the biotechnological potential of the microbial consortia of marine sponges. *Appl. Microbiol. Biotechnol.* 75:11–20

Kennedy J, Marchesi JR, Dobson ADW. 2008. Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb. Cell Fact.* 7:27

Kim EY, Oh K-H, Lee M-H, Kang C-H, Ph T-K, et al. 2009. Novel cold-adapted alkaline lipase from an intertidal flat metagenome and proposal for a new family of bacterial lipases. *Appl. Env. Microbiol.* 75:257–60

Kim TK, Fuerst JA. 2006. Diversity of polyketide synthase genes from bacteria associated with the marine sponge *Pseudoceratina clavata*: culture-dependent and culture-independent approaches. *Environ. Microbiol.* 8:1460–70

Kingsford C, Schatz MC, Pop M. 2010. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinforma.* 11:21

Knoll AH, Javaux EJ, Hewitt D, Cohen P. 2006. Eukaryotic organisms in Proterozoic oceans. *Philos. Trans. R. Soc. B* 361:1023–38

Kosakovsky Pond S, Wadhawan S, Chiaromonte F, Ananda G, Chung W-Y, et al. 2009. Windshield splatter analysis with the Galaxy metagenomic pipeline. *Genome Res.* 19:2144–53

Lane CE, Archibald JM. 2008. The eukaryotic tree of life: endosymbiosis takes its TOL. *Trends Ecol. Evol.* 23:268–75

Langridge G. 2009. Testing the water: marine metagenomics. *Nature* 7:552

Lee M-H, Lee C-H, Oh T-K, Song JK, Yoon J-H. 2006. Isolation and characterization of a novel lipase from a metagenomic library of tidal flat sediments: evidence for a new family of bacterial lipases. *Appl. Env. Microbiol.* 72:7406–9

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–59

Liu H, Probert I, Ultz J, Claustre H, Aris-Brosou S, et al. 2009. Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proc. Natl. Acad. Sci. USA* 106:12803–8

Luo H, Benner R, Long RA, Hu J. 2009. Subcellular localization of marine bacterial alkaline phophatases. *Proc. Natl. Acad. Sci. USA* 106:21219–23

Marco D. 2008. Metagenomics and the niche concept. *Theory Biosci.* 127:241–47

Martín-Cuadrado A-B, Ghai R, Gonsaga A, Rodriguez-Valera F. 2009. CO dehydrogenase genes found in metagenomic fosmid clones from the deep mediterranean sea. *Appl. Env. Microbiol.* 75:7436–44

Martín-Cuadrado A-B, López-García P, Alba J-C, Moreira D, Monticelli L, et al. 2007. Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* 2:e914

Martinez A, Tyson GW, Delong EF. 2010. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ. Microbiol.* 12:222–38

Martiny AC, Huang Y, Li W. 2009. Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ. Microbiol.* 11:1340–47

Martiny JB, Bohannan BJ, Brown JH, Colwell RK, Fuhrman JA. 2006. Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* 4:102–12

Massana R, Karniol B, Pommier T, Bodaker I, Beja O. 2008. Metagenomic retrieval of a ribosomal DNA repeat array from an uncultured marine aveolate. *Environ. Microbiol.* 10:1335–43
Massana R, Pedrós-Alió C. 2008. Unveiling new microbial eukaryotes in the surface ocean. *Curr. Opin. Microbiol.* 11:213–18
Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15:589–94
Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinforma.* 9:386
Mitra S, Gilbert JA, Field D, Huson DH. 2010. Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *ISME J.* 4:1236–42
Morgan JL, Darling AE, Eisen JA. 2010. Metagenomic sequencing of an in vitro–simulated microbial community. *PLoS ONE* 5:e10209
Mou X, Sun S, Edwards RA, Hodson RE, Moran MA. 2008. Bacterial carbon processing by generalist species in the coastal ocean. *Nature* 451:708–13
Nesbø CL, Boucher Y, Dlutek M, Doolittle WF. 2005. Lateral gene transfer and phylogenetic assignment of environmental fosmid clones. *Environ. Microbiol.* 7:2011–26
Neufeld JD, Chen Y, Dumont MG, Murrell JC. 2008. Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environ. Microbiol.* 10:1526–35
Not F, del Campo J, Balague V, deVargas C, Massana R. 2009. New insights into the diversity of marine picoeukaryotes. *PLoS ONE* 4:e7143
Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, et al. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. USA* 104:7705–10
Palenik B, Ren Q, Tai V, Paulsen IT. 2009. Coastal *Synechococcus* metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environ. Microbiol.* 11:349–59
Parker MS, Mock T, Armbrust EV. 2008. Genomic insights into marine microalgae. *Annu. Rev. Genet.* 42:619–45
Parks DH, Beiko RG. 2010. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26:715–21
Pedrós-Alió C. 2006. Genomics and marine microbial ecology. *Int. Microbiol.* 9:191–97
Piganeau G, Desdevises Y, Derelle E, Moreau H. 2008. Picoeukaryote sequences in the Sargasso Sea metagenome. *Genome Biol.* 9:R5
Pop M. 2009. Genome assembly reborn: recent computational challenges. *Brief. Bioinforma.* 10:354–66
Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP. 2009. Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ. Microbiol.* 11:1358–75
Prosser JL, Nicol GW. 2008. Relative contributions of archaea and bacteria to aerobic ammonia oxidation in the environment. *Environ. Microbiol.* 10:2931–41
Quinn JP, Kulakova AN, Cooley NA, McGrath JW. 2007. New ways to break an old bond: the bacterial carbonphosphorus hydrolases and their role in biogeochemical phosphorus cycling. *Environ. Microbiol.* 9:2392–400
Reisch CR, Moran MA, Whitman WB. 2008. Dimethylsulfoniopropionate-dependent demethylase (DmdA) from *Pelagibacter ubique* and *Silicibacter pomeroy*. *J Bacteriol.* 190:8018–24
Riesenfeld CS, Schloss PD, Handelsman J. 2004. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* 38:525–52
Rodriguez-Brito B, Rohwer F, Edwards RA. 2006. An application of statistics to comparative metagenomics. *BMC Bioinforma.* 7:162
Rodriguez-Valera F. 2004. Environmental genomics, the big picture? *FEMS Microbiol. Lett.* 231:153–58
Rondon MR, August PR, Betterman AD, Brady SF, Grossman TH, et al. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66:2541–47
Rondon MR, Raffel SJ, Goodman RM, Handelsman J. 1999. Toward functional genomics in bacteria: Analysis of gene expression in *Escherichia coli* from a bacterial artificial chromosome library of *Bacillus cereus*. *Proc. Natl. Acad. Sci. USA* 96:6451–55

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. 2007. The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* 5:e77

Rusch DB, Martiny A, Dupont CL, Halpern AL, Venter JC. 2010. Characterization of *Prochlorococcus* clades from iron depleted oceanic regimes. *Proc. Natl. Acad. Sci. USA*. 107:16184–89

Sanders RW, Caron DA, Berninger U-G. 1992. Relationships between bacteria and heterotrophic nanoplankton in marine and fresh water: an interecosystem comparison. *Mar. Ecol. Prog. Ser.* 86:1–14

Schattenhofer M, Fuchs BM, Amann R, Zubkov MV, Tarran GA, Pemthaler J. 2009. Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environ. Microbiol.* 11:2078–93

Schirmer A, Gadkari R, Reeves CD, Ibrahim F, DeLong EF, et al. 2005. Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissolute*. *Appl. Environ. Microbiol.* 71:4840–49

Schleper C, Jurgens G, Jonuscheit M. 2005. Genomic studies of uncultivated archaea. *Nat. Rev. Microbiol.* 3:479–88

Schloss PD, Handelsman J. 2003. Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol.* 14:303–10

Schloss PD, Handelsman J. 2005. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.* 6:229

Schmeisser C, Steele H, Streit WR. 2007. Metagenomics, biotechnology with nonculturable microbes. *Appl. Microbiol. Biotechnol.* 75:955–62

Schwartz K. 2006. Recent advances in marine metagenomics. *MMG 445 Basic Biotechnol. eJournal* 2:165–69

Sebastian M, Ammerman JW. 2009. The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA. *ISME J.* 3:563–72

Sherr EB, Sherr BF. 2002. Significance of predation by protists in aquatic microbial food webs. *Antonie Van Leeuwenhoek* 81:293–308

Shi XL, Marie D, Jardillier L, Scanlan DJ, Vaulot D. 2009. Groups without cultured representatives dominate eukaryotic picophytoplankton in the oligotrophic east Pacific Ocean. *PLoS ONE* 4:e7657

Singh J, Behal A, Singla N, Joshi A, Birbian N, et al. 2009. Metagenomics: concept, methodology, ecological inference and recent advances. *Biotechnol. J.* 4:480–94

Sleator RD, Shortall C, Hill C. 2008. Under the microscope: metagenomics. *Lett. Appl. Microbiol.* 47:361–66

Steele HL, Jaeger KE, Daniel R, Streit WR. 2009. Advances in recovery of novel biocatalysts from metagenomes. *J. Mol. Microbiol. Biotechnol.* 16:25–37

Steele HL, Streit WR. 2005. Metagenomics: advances in ecology and biotechnology. *FEMS Microbiol. Lett.* 247:105–11

Stein JL, Marsh TL, Wu KY, Shizuya H, Delong EF. 1996. Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kb-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* 178:591–99

Stokes HW, Nesbo CL, Holley M, Bahl MI, Gillings MR, et al. 2006. Class 1 integrons potentially predating the association with Tn402-like transposition genes are present in a sediment microbial community. *J. Bacteriol.* 188:5722–30

Suzuki MT, Beja O, Taylor LT, Delong EF. 2001. Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ. Microbiol.* 3:323–31

Temperton B, Field D, Oliver A, Tiwari B, Joint I, et al. 2009. Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. *ISME J.* 3:792–96

Thurber RV, Wilner-Hall D, Rodrigues-Mueller B, Desnues C, Edwards RA, et al. 2009. Metagenomic analysis of stressed coral holobionts. *Environ. Microbiol.* 11:2148–63

Todd JD, Curson ARJ, Dupont CL, Nicholson P, Johnston AWB. 2009. The *dddP* gene, encoding a novel enzyme that converts dimethylsulfoniopropionate into dimethylsulfide, is widespread in ocean metagenomes and marine bacteria and also occurs in some Ascomycete fungi. *Environ. Microbiol.* 11:1376–85

Treusch AH, Leininger S, Kietzin A, Schuster SC, Klenk H-P, et al. 2005. Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environ. Microbiol.* 7:1985–95

Tring SG, von Merring C, Kobayashi A, Salamov AA, Chen K. 2005. Comparative metagenomics of microbial communities. *Science* 308:554–57

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, et al. 2010. Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* 464:90–94

Van Dolah FM, Lidie KB, Monroe EA, Bhattacharya D, Campbell L, et al. 2009. The Florida red tide dinoflagellate *Karenia brevis*: new insights into cellular and molecular processes underlying bloom dynamics. *Harmful Algae* 8:562–72

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74

Vergin KL, Urbach E, Stein JL, Delong EK, Lanoil BD, et al. 1998. Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order *Planctomycetales*. *Appl. Environ. Microbiol.* 64:3075–78

Walsh DA, Zalkova E, Howes CG, Song YC, Wright JJ, et al. 2009. Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* 326:578–82

Ward N. 2006. New directions and interactions in metagenomics research. *FEMS Microbiol. Ecol.* 55:331–38

Warnecke F, Hugenholtz P. 2007. Building on basic metagenomics with complementary technologies. *Genome Biol.* 8:231

Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ. 2007. Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol. Direct* 2:27

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. 2009. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 4:e7370

Willner D, Thurber RV, Rohwer F. 2009. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol.* 11:1752–66

Woebken D, Teeling H, Wecker P, Dumitriu A, Kostadinov I, et al. 2007. Fosmids of novel marine Planctomycetes from the Namibian and Oregon coast upwelling systems and their cross-comparison with Planctomycete genomes. *ISME J.* 1:419–35

Wooley JC, Godzik A, Friedberg I. 2010. A primer on metagenomics. *PLoS Comput. Biol.* 6:e1000667

Woyke T, Teeling H, Ivanova NN, Huntermann M, Richter M, et al. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443:950–55

Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, et al. 2009. Assembling the marine metagenome, one cell at a time. *PLoS ONE* 4:e5299

Wu D, Hudenholtz P, Mavromatis K, Pukall R, Dalin E, et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056–60

Xu J. 2006. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol. Ecol.* 15:1713–31

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. 2007. The *Sorcerer II* Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5:432–66

Yutin N, Béjà O. 2005. Putative novel photosynthetic reaction center organizations in marine aerobic anoxygenic photosynthetic bacteria: insights from metagenomics and environmental genomics. *Environ. Microbiol.* 7:2027–33

Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, et al. 2007. Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific oceans using the global ocean sampling expedition metagenomes. *Environ. Microbiol.* 9:1464–75

Zubkov MV, Tarran GA. 2008. High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature* 455:224–26

Annual Review of Marine Science

Volume 3, 2011

Contents

Errata

An online log of corrections to *Annual Review of Marine Science* articles may be found at http://marine.annualreviews.org/errata.shtml