NATURE|Vol 459|14 May 2009|doi:10.1038/nature08059

The microbial ocean from genomes to biomes

Edward F. DeLong[1]

Numerically, microbial species dominate the oceans, yet their population dynamics, metabolic complexity and synergistic interactions remain largely uncharted. A full understanding of life in the ocean requires more than knowledge of marine microbial taxa and their genome sequences. The latest experimental techniques and analytical approaches can provide a fresh perspective on the biological interactions within marine ecosystems, aiding in the construction of predictive models that can interrelate microbial dynamics with the biogeochemical matter and energy fluxes that make up the ocean ecosystem.

Just 40 years ago, the number of microorganisms in each millilitre of sea water was underestimated by a staggering three orders of magnitude. Astronauts may have been exploring the Moon, but most of the microbial life on Earth remained largely undiscovered. The situation changed dramatically in the late 1970s and early 1980s, when accurate estimates of total cell numbers in the sea became available. Over the next 25 years or so, local, regional and global estimates of microbial numbers, along with their bulk production and consumption rates in ocean surface waters, were quantified and mapped. These data provided increasingly accurate estimates of the total biomass of planktonic microorganisms and their turnover, enlarging their perceived role and significance in ocean food webs. Although this information was extremely useful, more specific data on the biology of planktonic Bacteria and Archaea have only recently become available, allowing us to address a new range of questions. Which taxa of marine Bacteria and Archaea are most dominant or biogeochemically important in particular ocean provinces or depth strata? What are the most common microbial metabolic pathways, and how do they vary within and between communities and environments? How do dynamic population shifts and species interactions shape the ecology and biogeochemistry of the seas?

Unlike eukaryotic plankton, which can often be taxonomically and metabolically categorized according to directly observable phenotypes, it has been more difficult to ascertain the core identities and physiological properties of planktonic Bacteria and Archaea. Recent advances in cultivation-independent metagenomics, in which DNA from the microbial community is collected, sequenced and analysed en masse, as well as new cultivation technologies, have had a dramatic influence on our knowledge of non-eukaryotic microorganisms. The integrated perspective provided by a combination of cultivation-independent phylogenetic surveys, microbial metagenomics and culture-based studies has delivered a more detailed understanding of microbial life in the sea. Here I discuss some of the contributions and synergy of metagenomics and the new cultivation approaches, focusing on recent advances achieved using these new techniques.

Phylogenetic surveys and model systems

One of the drivers for developing cultivation-independent approaches for the phylogenetic identification of microorganisms[1] was the recognition that only a small proportion of the microbial cells sampled from the environment can be readily cultivated using conventional techniques[2]. The development of ribosomal-RNA-based phylogenetic surveys in the 1980s led to less biased assessments of the distribution of uncultivated bacterial, archaeal and protistan phylotypes in natural populations[1]. The number of newly recognized bacterial and archaeal phylogenetic divisions has increased markedly. Indeed, in many habitats, some of the most abundant microbial phylotypes have no close relatives that have been cultured[3]. These and other results from cultivation-independent surveys have fundamentally changed our perspective on microbial phylogeny, evolution and ecology. These discoveries subsequently inspired more directed cultivation strategies, aimed at isolating some of the more environmentally abundant microbial phylotypes that had previously escaped cultivation[4–6].

Directed cultivation still has an important role in describing the nature and properties of marine Bacteria and Archaea. For example, the ocean's most abundant cyanobacterium, *Prochlorococcus*, which was first discovered by ship-board flow cytometry[7], was successfully cultivated soon after its discovery[8]. Isolates of *Prochlorococcus* now provide an environmentally relevant system for modelling the biology and ecology of planktonic cyanobacteria. Physiological characterization of *Prochlorococcus* genotypic variants led to the idea of 'ecotypes', which are highly related yet physiologically and genetically distinct populations that are adapted to different environmental conditions. An oceanographic survey of six *Prochlorococcus* ecotype variants in the Atlantic Ocean confirmed their distinct environmental distributions across broad environmental isoclines. *Prochlorococcus* isolates have also been used in detailed studies of phage diversity, host range, genome content, host–phage genetic exchange[9] and gene-expression dynamics[10]. The integration of *Prochlorococcus* lab-based physiological modelling and field-based surveys has also helped constrain and validate some computational ecosystem models that can successfully recapitulate known *Prochlorococcus* ecotype distributions in the environment[11], suggesting promising future directions in microbial oceanography.

The development of 'dilution to extinction' cultivation techniques[4] is another important advance aimed at culturing the new phylotypes discovered in rRNA-based environmental surveys. The basic approach involves preparing sterilized sea water, which is distributed into tissue-culture wells and subsequently inoculated with serially diluted bacterioplankton[6]. Growth in these low-density cultures is monitored by cell counting. These approaches have been hugely successful with respect to the recovery in pure culture of many dominant surface-water bacterioplankton[4–6,12].

[1]Departments of Biological Engineering and Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

As with any approach, however, there are practical limitations, such as uncertainties when using undefined and variable seawater media and the low probability of isolating rare organisms in the dilution-to-extinction approach. Indeed, the reasons why some predominant groups are readily cultivated, whereas others continue to resist cultivation, are still not well understood[4]. Nevertheless, the isolation and partial characterization of more representative bacterioplankton strains is having a major impact on our understanding of their genomic, phenotypic and physiological properties. The effects of these new approaches to cultivation are especially evident in the isolation of *Pelagibacter ubique*[12], a member of perhaps the most abundant bacterial group in the oceans. Isolates of *P. ubique* are now yielding fresh data on the phenotype[12,13], genome content[14], genetic variability[5,14] and physiology[15,16] of this major bacterioplankton taxon.

The cultivation of resident microorganisms is a valuable part of the drive to describe microbial processes in the environment, but it is not enough on its own. Although pure cultures provide readily manipulated models, there are fundamental limitations to their utility when it comes to inferring ecological processes. Some physicochemical variables can be well controlled in cultures, but patterns of temperature, pressure, pH, nutrient concentrations and redox balance, and their naturally occurring gradients, may sometimes be difficult to reproduce in the laboratory. Additionally, many microorganisms have evolved to interact closely with other organisms and are often engaged in obligatory symbiotic relationships. For these and other reasons, it is unreasonable to assume that pure-culture microbial models will be available for all the ecologically important microorganisms. Cultivation-independent phylogenetic and genomic surveys will continue to have an important role in describing uncultured microorganisms and their population genetics and biogeochemical and ecological interactions, which cannot be well studied or modelled in laboratory systems.

Microbial metagenomics and cultivation

For the purpose of this Review, 'metagenomics' is defined as the cultivation-independent genomic analysis of microbial assemblages or populations. Although still in its infancy, metagenomics has already contributed to our knowledge of genome structure, population diversity, gene content and the composition of naturally occurring microbial assemblages. In low-complexity populations, metagenomic studies have led to the assembly of almost complete genomes from the abundant genotypes[17] and have provided composite genomic representations of dominant populations[18,19]. Advances and improvements in sequencing technologies are propelling the field forward rapidly (Box 1). Despite the large data sets now available, high allelic variation in microbial populations, high species richness and a relatively even representation among species still render whole-genome assemblies of individual genotypes mostly impractical, given current sequencing and assembly technologies[20–22] (Box 2).

The coupling of metagenomics and culture-based approaches is particularly useful. Every methodology has its own shortcomings (see Box 2), but metagenomic surveys have already contributed significantly to our understanding of the microorganisms in the environment. For example, metagenomic data sets have allowed the directed enrichment and isolation of new isolates with specific and predicted functional and genetic properties[23]. In metagenomic surveys along environmental gradients, direct observations of gene distributions in the water column have revealed patterns of vertical stratification of functional genes, bacteriophage and other genetic properties, providing clues about the differential distribution of metabolic processes, phage–host interactions and evolutionary dynamics along the depth continuum[24]. A more recent survey using the latest pyrosequencing technologies compared more than 70 marine metagenomic data sets and revealed statistically significant differences in gene content among the nine major biomes compared[25]. In a recent dramatic example of cell-specific metagenomics, the genome content of an uncultivated nitrogen-fixing cyanobacterium population (UCYN-A) recovered by flow cytometry has been reported[26]. The genome sequences of the UCYN-A cell population revealed that these cyanobacteria, as expected, contained all the genes required for nitrogen fixation and all the components of photosystem I. The big surprise was that UCYN-A lacked the genes required for carbon dioxide fixation and oxygenic photosynthesis that are found in all other known free-living cyanobacteria[26]. The metagenomic data suggest that these cyanobacteria are not oxygen-generating photoautotrophs. This study provides an excellent example of how metagenomics can be used to identify the metabolic capabilities of uncultivated microbial phylotypes, a crucial goal in microbial ecology.

Metagenomic analyses of bacterial and archaeal populations have often presaged the later findings of culture-based studies. More specifically, metagenomic data have revealed unexpected phylogenetic and environmental distributions of genes and metabolisms. Early metagenomic studies, for example, revealed the unexpected presence of a bacteriorhodopsin-like photoprotein gene in an abundant marine bacterioplankton group (SAR86)[27]. Biophysical and functional characterization of the proteorhodopsin gene product confirmed its ability to function as a light-driven proton pump[27]. Later metagenomic surveys revealed the high abundance and global distribution of these rhodopsins in marine planktonic Bacteria and Archaea[20,21,28–35]. Subsequent genome sequencing of cultivated marine isolates then confirmed the widespread distribution of rhodopsin genes in many taxa of marine Bacteria[13,36,37]. Similarly, metagenomics revealed new types of aerobic, anoxygenic photosynthetic

Box 1 | Evolving genomic technologies

The range of genomic and metagenomic data now available for marine microorganisms is expanding rapidly for a variety of reasons. First, the acquisition of whole genome sequences from cultivated strains of microorganisms is becoming much faster and cheaper, so genome sequences are accumulating rapidly, with thousands now in the pipeline. With respect to marine microorganisms, hundreds of whole or draft bacterial and archaeal genome sequences are already available in public databases. In addition, nucleic-acid sequences recovered directly from total microbial assemblages are fast outstripping microbial whole-genome sequence data. The drivers for this include an increasing awareness of the usefulness of such data, a few major expeditions that have contributed large volumes of shotgun sequence data, and advancing technologies[60] that are making large amounts of sequence data readily available.

In addition to the size of metagenomic data sets, the heterogeneity of data types and environments sampled is also expanding dramatically. Original data sets mainly included Sanger-based shotgun sequence data of cloned DNA captured in small insert clone libraries (about 3 kilobase pairs, kbp) or longer genome fragments (40–100 kbp) in bacterial artificial chromosomes (BACs). More recently, pyrosequencing techniques[60] that do not require DNA clone libraries (eliminating the associated labour and cost overheads) have rapidly evolved from initial read lengths of 100 bp to 450 bp. Other next-generation technologies that involve sequencing by synthesis but generate very short reads (around 25 bp) may also prove useful in metagenomics, if sufficient long-read reference databases are available. On the horizon are technologies that will allow even higher-throughput, longer-read, single-molecule sequencing[61,62]. These advances will make a huge difference with respect to the amount of data that can be collected, as well as the bioinformatic infrastructure that will be required for analysis and synthesis to occur.

Single-cell genome sequencing using multiple displacement amplification (MDA) techniques coupled with new sequencing technologies also promises better genomic access to uncultivated or rare microorganisms[63–65], although significant challenges remain[64,65]. Chief among these are contamination problems associated with the 'extreme amplification' of large amounts of DNA from a single cell. Additionally, inherent mechanisms of the MDA reaction itself result in uneven amplification and coverage of even single, pure genotypes[65]. Partial draft genomes can be produced from single cells but currently not without extraordinary efforts to reduce contamination and to normalize for uneven coverage[63,66]. Nevertheless, incremental improvements in single-genome sequencing in the future are likely to allow the recovery of more partial draft genomes from as-yet-uncultivated Bacteria and Archaea. These are expected to both provide benefits to and derive benefit from the more traditional metagenomic approaches currently in common use.

Box 2 | Problems with metagenomic methods

The technical constraints of microbial sampling, changes in sequencing technologies and the sheer complexity and size of the data sets all present significant challenges for interpreting and comparing genomic data from microbial communities. Some of the larger challenges are discussed below.

There are numerous technical challenges associated with even the seemingly simple task of obtaining representative and reproducible samples. Sampling strategies are always context dependent and are influenced by the type of microbial community, its environment, the spatial scale sampled, the population density and the presence of contaminating substances. There are many relevant questions. Do the cells need to be purified away from a soil, sediment or rock matrix? To reduce sample complexity, will the cells be separated by size from larger eukaryotic species? Do the cells need to be concentrated before the DNA is extracted? These and other concerns about sampling are central to the interpretation of the resultant data sets.

The methods used to recover and sequence DNA from microbial communities are also critical. Past approaches using Sanger sequencing have predominantly relied on the cloning of individual DNA molecules. Cloning biases are well known, and in some cases specific genes[68] (as well as specific phylogenetic groups[69,70]) may be under-represented in genomic and metagenomic clone libraries. However, problems with such biases have been largely overcome by pyrosequencing[61] and other next-generation sequencing technologies that sidestep the need to clone individual DNA molecules.

Another problem relates to functional gene predictions and annotation. Even preliminary tasks of gene characterization, including calling open reading frames, identifying taxonomic origins and inferring functional properties, are non-trivial enterprises in analyses of metagenomic data sets. Complicating factors include short sequence read lengths, poor sequence quality, the absence of gene-linkage context, and having extremely large data sets and uneven coverage. Several strategies for metagenomic open-reading-frame prediction[22,71,72], phylogenetic assignment[73,74] and functional predictions[22,75,76] have recently been developed, and improvements and new approaches to these fundamental tasks continue to evolve. For example, a study combining homology searches and gene neighbourhood analyses succeeded in specific functional gene predictions for 76% of the 1.4 Mbp examined[77]. Such advances, alongside customized metagenomic databases[51–53], promise to improve current capabilities for gene identification and the annotation of metagenomic data sets.

Statistical approaches for the comparison of metagenomic data sets have only recently been applied, so their development is at an early stage. The size of the data sets, their heterogeneity and a lack of standardization for both metadata and gene descriptive data continue to present significant challenges for comparative analyses. Statistical approaches to examine gene distributions in the environment have so far included gene-enrichment probability estimates in three-way comparisons[75], bootstrap resampling methods that evaluate gene-abundance confidence intervals deviating from the median in pairwise sample comparisons[78], canonical discriminant analyses that identify the genes that most influence distributional variance[25], and canonical correlation analyses that interrelate metabolic-pathway occurrence with multiple environmental variables[79]. However, only highly disparate sample types have been the subject of much statistical scrutiny. It will be interesting to learn the sensitivity limits of such approaches, along more fine-scale taxonomic, spatial and temporal microbial community gradients, for example in the differences between the microbiomes of human individuals[44]. As the availability of data sets and comparable metadata fields continues to improve, quantitative statistical metagenomic comparisons are likely to increase in their utility and resolving power.

bacteria in marine plankton[38], an observation that was later confirmed by strain-isolation studies[39,40].

The predictive power of metagenomics was also demonstrated in the finding of genes associated with ammonia oxidation in Archaea, a character previously found in just a few bacterial groups. Two concurrent metagenomic studies[20,41] reported that a specific clade of Crenarchaeota seemed to have the genes diagnostic for chemolithotrophic ammonia oxidation. At about the same time, enrichment cultures using ammonia as the sole energy source and CO_2 as the sole carbon source yielded an ammonia-oxidizing crenarchaeal isolate[42]. Parallel metagenomic analyses of the genome sequence from an uncultured crenarchaeon extended previous studies beyond a single gene in the pathway and suggested specific functional differences between the archaeal and bacterial ammonia-oxidizing metabolic pathways[18,43]. In a very short time period, Archaea came to be recognized as potentially important contributors to a part of the nitrogen cycle previously thought to be regulated solely by Bacteria.

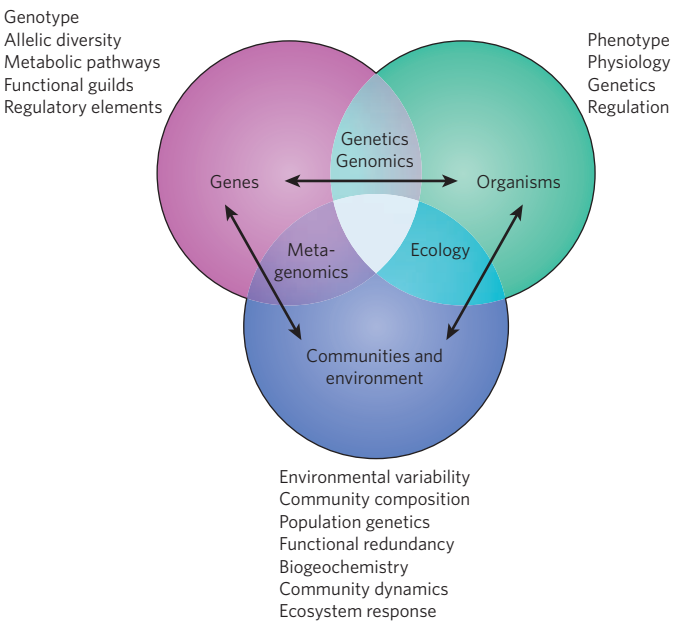

Figure 1 | The intersection of traditional disciplines and metagenomics. The pink, green and blue regions represent the fundamental elements of study: genes, organisms and the environment. Areas of investigation associated with each are indicated in the text. The intersections between the elements show the disciplinary overlaps: genetics/genomics, metagenomics and ecology. The pale blue area in the middle identifies the 'sweet spot' in which information from cultured-based studies, environmental studies and metagenomics can be integrated and modelled.

These and other examples have clearly indicated the value of integrating and comparing metagenomic and culture-based studies. Indeed, the deficiencies of each approach are largely compensated for by the strengths of the other. Phenotype, metabolism and physiology are mainly inferred from laboratory culture-based experiments, whereas detailed information on environmental distributions and ranges, population genetics, and community interactions and dynamics are best viewed through the lens of cultivation-independent strategies, including metagenomics. It is also clear that reference genome sequences from cultivated microorganisms greatly aid metagenomic studies. The integration of metagenomics, cultivation-based studies and environmental surveys leads to insights not previously open to microbiologists (Fig. 1), at the intersection of genes, organisms and the environment. More specifically, the integration of cultivation-dependent and cultivation-independent approaches partly bridges the gap between genomics, population genetics, biochemistry, physiology, biogeochemistry and ecology. Approaches that combine cultivation and metagenomic perspectives will undoubtedly be more common in future collaborative microbiological studies. Plans for human microbiome studies are a good case in point[44].

Nucleic-acid sequences as analytes in ecosystem studies

The development of metagenomic methods has helped to expand the repertoire of known microbial genes, their environmental distributions

and their allelic diversity. The associated bioinformatic analyses are useful for generating new hypotheses, but other methods are required to test and verify *in silico* hypotheses and conclusions in the real world. It is a long way from simply describing the naturally occurring microbial 'parts list' to understanding the functional properties, multi-scalar responses and interdependencies that connect microbial and abiotic ecosystem processes. New methods will be required to expand our understanding of how the microbial parts list ties in with microbial ecosystem dynamics. Experimental technologies that can leverage massively parallel sequencing technologies, or that can link information from pre-existing sequence data sets with experimental observations in natural assemblages, seem particularly promising.

Several approaches are available that have the potential to link DNA sequences found in the microbial community with specific microorganisms and their activities in the environment. One method uses the thymidine analogue 5-bromodeoxyuridine (BrdU) to tag actively growing substrate-responsive cells. The BrdU-labelled DNA is immuno-captured and subsequently sequenced to identify taxa and genes specific to a given experimental treatment[45]. Stable-isotope analyses also have significant potential for tracking specific microbial groups that incorporate labelled organic or inorganic compounds into living tissues. Stable-isotope tracers have been used to identify methanotrophic Archaea, to localize nitrogen-fixing symbionts in host tissues, and to verify autotrophic metabolism in planktonic Crenarchaeota. A novel approach that has the potential to link DNA sequence information directly to substrate-specific incorporation is stable-isotope probing, where nucleic acids labelled with a 'heavy' isotope are physically isolated by buoyant density centrifugation and subsequently sequenced[46].

The application of gene-expression technologies to track microbial sensing and responses in the environment is another exciting development. In this approach, bacterial and archaeal total RNA is extracted from microbial assemblages, converted to complementary DNA and sequenced (Fig. 2). Early studies began with the analysis of randomly primed cDNA clone libraries by Sanger-based capillary sequencing to survey abundant transcripts from a coastal seawater sample[47]. Advances such as pyrosequencing, which sidesteps the need for clone libraries, have allowed the analysis of larger data sets obtained from more rapidly collected, smaller-volume samples of marine bacterioplankton[48]. Pyrosequencing of both genomic DNA and cDNA from the same sample allows the normalization of transcript abundance to the corresponding gene copy number of the community's collective gene pool[48] (Figs 2, 3).

Early high-throughput, pyrosequence-based studies[48] of the transcriptome of planktonic microbial communities have led to several new insights. Not surprisingly, genes associated with the key metabolic pathways of open-ocean microbial species (including photosynthesis, carbon fixation and nitrogen acquisition) were found to be highly expressed in the photic zone at a depth of 75 m in the North Pacific Subtropical Gyre. Both genomic and transcriptomic data sets showed high coverage of some dominant community members, such as *Prochlorococcus,* with hypervariable genomic regions showing some of the highest transcript abundances. Many of the microbial community transcripts were similar to previously predicted genes found in ocean metagenomic surveys, but about half seemed to be unrelated to predicted protein sequences in available databases[48]. The transcriptomic data sets in such studies contain several categories of RNA, including rRNAs, messenger RNAs and small RNAs[49], some of which have an important role in regulating gene expression. Each of the molecular species recovered — rRNA, mRNA and small RNA — has the potential to shed light on the dynamics and variability of the phylogenetic composition, functional properties and regulation of natural microbial communities.

The application of transcriptomic methods to microbial communities is creating a new research agenda in which sequence data are the analytes in experimental field studies. This approach allows the measurement of gene expression in microbial assemblages, in microcosms, mesocosms or natural samples, as a function of environmental variability over time (Fig. 3). The environmental variation examined can be natural (for example, tracking changes in gene expression as a function of the daily cycle) or applied (for example, monitoring changes in gene expression following changes to nutrient levels). By tracking which genes are responsive to specific environmental perturbations, it should soon be possible to track environmental variations that are first observed as changes in gene expression but later may lead to shifts in community composition (Fig. 3). Quantifying the variability and kinetics of gene expression in natural assemblages has the

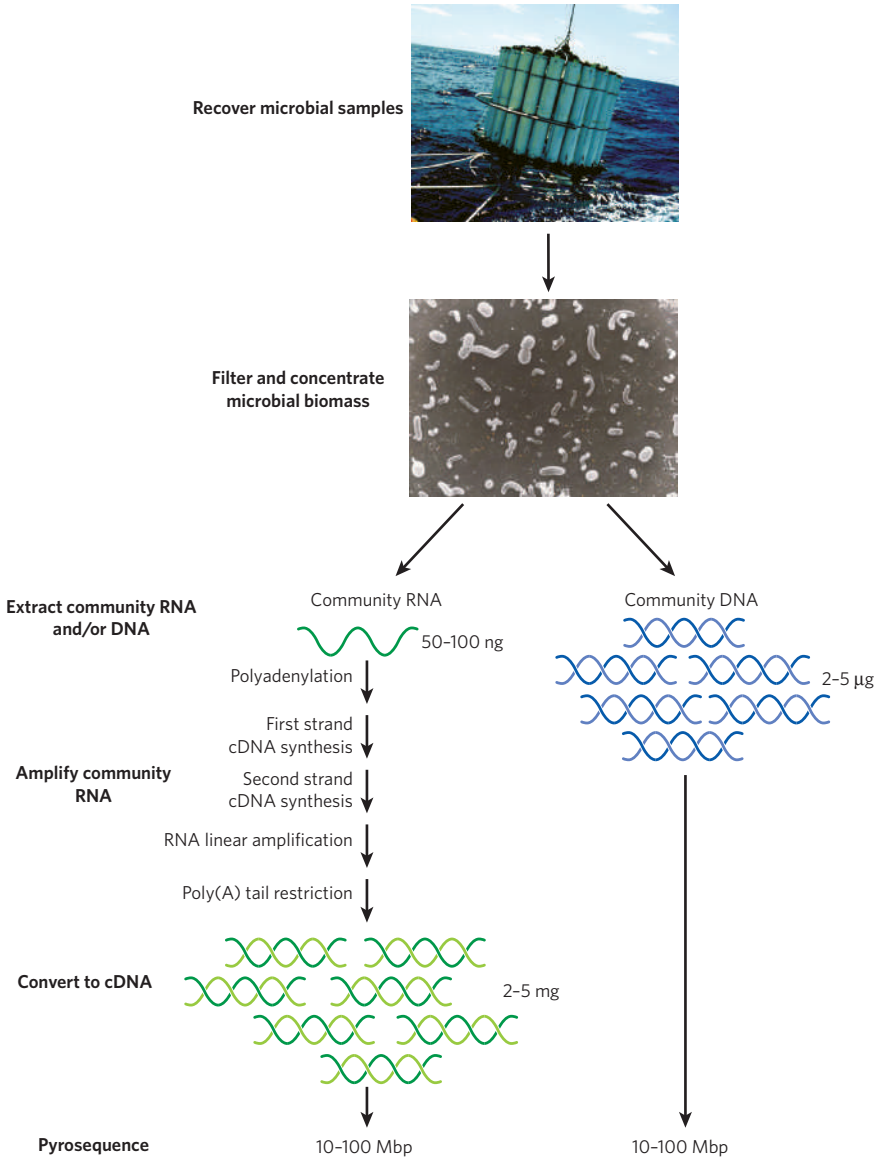

Figure 2 | Transcriptome sequencing protocol for marine microbial assemblages. Cells are collected and processed to produce genomic DNA, or cDNA from total RNA[48]; samples for RNA extraction are collected in smaller volumes (less than 1 litre) and filtered as rapidly as possible (about 10 min). After RNA amplification and conversion to cDNA, cDNA and genomic DNA from the same assemblage are sequenced and compared.

potential to provide a fresh perspective on microbial community dynamics. Can expression patterns provide clues to the functional properties of putative genes? What are the key community responses to environmental perturbation? What fundamental community-wide regulatory responses are common to different taxa? Are certain taxa or metabolic pathways more or less responsive to particular environmental changes? Are specific changes in gene expression indicative of changes in community composition? These and other questions can now be addressed more directly by applying these new experimental approaches.

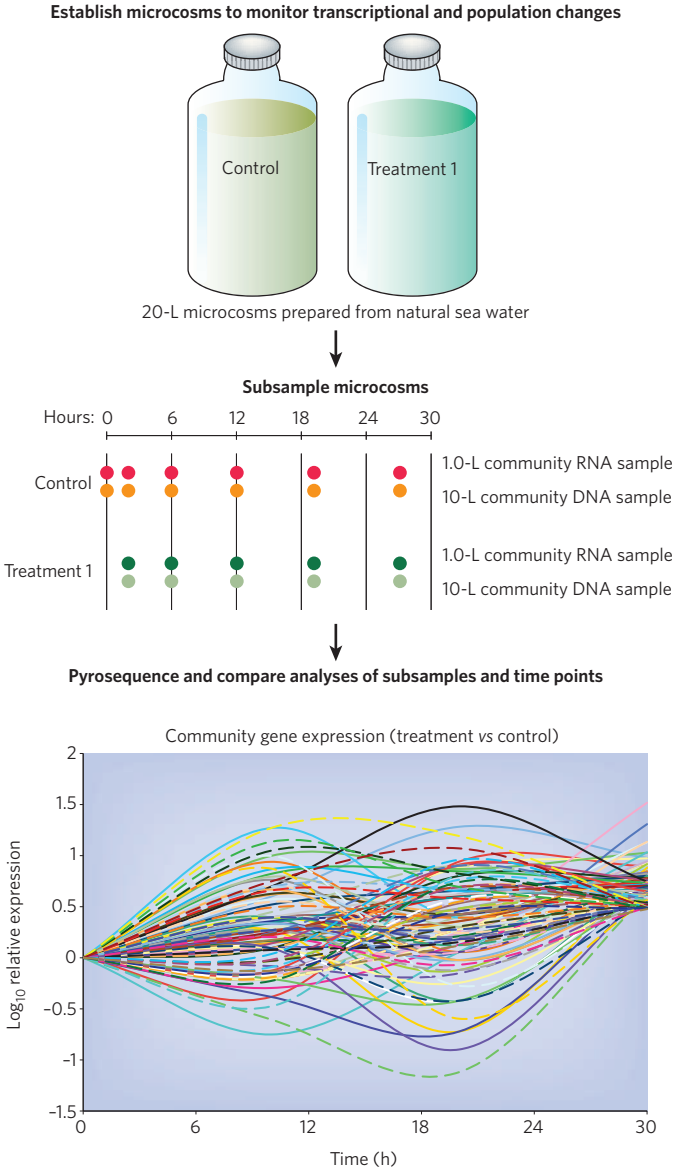

Figure 3 | Quantifying microbial responses to environmental variability using environmental transcriptomics. The experiments shown have been made possible by tandem metagenomic and 'metatranscriptomic' pyrosequencing (Fig. 2). Initially, microcosms containing aquatic microbial communities are established. The untreated sample is a control for intrinsic incubation effects, as well as natural daily variation in gene expression. Different experimental treatments could measure a variety of physical or environmental perturbations, including the effects of light, nutrients, temperature or anthropogenic compounds. Microbial-assemblage DNA and RNA subsamples are taken at various time points, subjected to pyrosequencing (see Fig. 2) and analysed and compared. Differential gene expression between control and treatment communities (bottom panel) is used to identify microbial responses to environmental perturbation. Coloured lines represent individual gene categories that are overexpressed or underexpressed relative to the control.

Information management from genes to ecosystems

One of the major challenges facing the emerging metagenomic and 'metatranscriptomic' studies is the sheer size of the data sets, and the methods and tools that are therefore needed to deal with them. Large data sets create challenges with respect to data management, computational resources, sampling and analytical strategies, and database architectures. It is encouraging that the research community has recognized the need to establish clear standards for the submission and reporting of data so that primary sequence data can be related across relevant environmental parameters. The Genomic Standards Consortium (http://gensc.org) is promoting schemes reminiscent of the MIAME standards for microarray data (http://www.mged.org/Workgroups/MIAME/miame.html). These would capture metadata associated with genomes (minimum information about a genome sequence) and metagenomic data (minimum information about a metagenome sequence)[50]. For comparative analyses of archived data sets, such metadata field standardization and reporting will be critical.

We are entering a new era in microbial ecology and biology in which experimental high-throughput sequencing data will increasingly be analysed (Fig. 3). The coordination of experimental reports from such studies will be important, and MIAME-like standards for such reporting (minimum information about a high-throughput sequencing experiment) have recently been proposed (http://www.mged.org/minseqe). Even simple annotation, archiving and accessing of sequence-data types and experiments, along with associated and relevant metadata, pose serious challenges for the biological community. These challenges are being addressed by the development of new metagenomic databases[51–53], analytical strategies and statistical approaches (Box 2).

Efficient bioinformatics management and analytical practices will not be a panacea for the larger challenge of describing microbial biology at an ecosystem level. There is still a mismatch with respect to the integration of 'bottom up', reductionist molecular, approaches with 'top down', integrative ecosystems, analyses. Molecular data sets are often gathered in massively parallel ways, but acquiring equivalently dense physiological and biogeochemical process data[54] is not currently as feasible. This 'impedance mismatch' (the inability of one system to accommodate input from another system's output) is one of the larger hurdles that must be overcome in the quest for more realistic integrative analyses that interrelate data sets spanning from genomes to biomes.

The road ahead

The microbial parts list of the genes and genomes in metagenomic data sets is growing rapidly, but work to understand their functional and ecological relevance is proceeding more slowly. DNA sequence data and bioinformatic analyses fall short of describing which gene suites are being expressed, and which metabolic pathways are being used, in any given environmental context. A large number of hypothetical proteins that have been identified may be ecologically important but have functions that remain unknown. How do community composition, gene content and variability influence biogeochemical function, turnover rates and ecosystem processes? How important are functional redundancy and allelic diversity to community function and stability? How does the process of succession play out, from the initial environmental change to shifts in microbial community composition? Can we predict the probability of lateral gene transfer and gene fixation for particular functional properties or gene categories? Can suites of genes and their variability be correlated with larger-scale biogeochemical and ecological patterns and processes? Can we determine the functional properties and roles of as-yet-uncharacterized proteins that share little or no homology with functionally annotated proteins? How representative are the activities and responses of microbial isolates in the laboratory, with respect to their physiological and metabolic behaviour in the environment? Fresh approaches will be required to address these and other questions that are currently being raised.

We need to develop and explore new strategies to bridge the gaps between microbial genomics, metagenomics, biochemistry, physiology, population genetics, biogeochemistry, oceanography and ecosystem

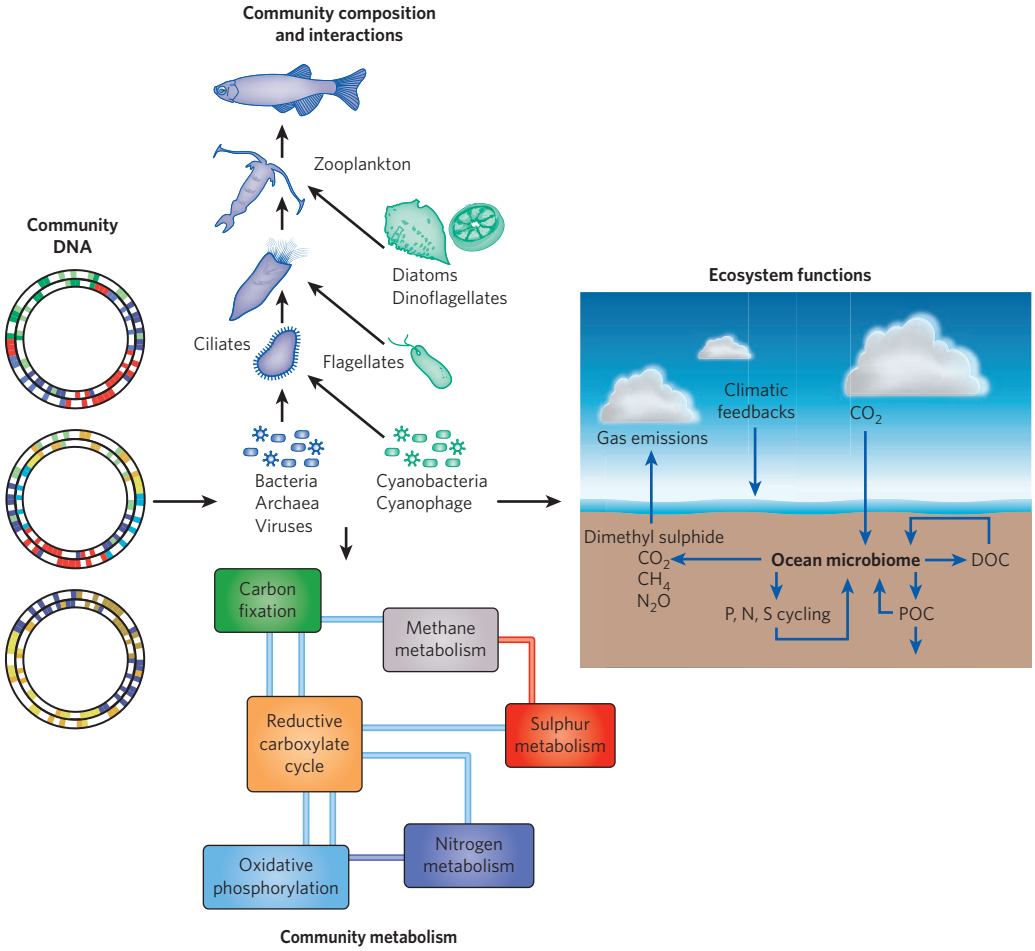

Figure 4 | The network instructions encoded in microbial genomes drive ecosystem processes. This schema shows hypothetical linkages between the genomic information of the microbial assemblage and the collective ecological interactions and community metabolism that in part regulate and sustain biogeochemical and ecosystem processes. Each DNA circle in the left panel represents a genome derived from a marine bacterioplankton species. Co-occurring microorganisms that inhabit the same environment collectively form the pool of genes sampled in metagenomic studies. These instructions modulate community interactions, metabolism and ecosystem function. DOC, dissolved organic carbon; POC, particulate organic carbon.

biology. Integrative and interdisciplinary interactions will be key to future studies because microbial diversity, metabolism and biogeochemistry are all intertwined over multiple temporal and spatial scales. One central hypothesis that drives metagenomics is that the network instructions for metabolic processes, biogeochemical function and ecological interactions are encoded in the collective microbial genomes and expressed in response to environmental variability. These network instructions are eventually expressed as the biological drivers of ecosystem processes (Fig. 4).

Microbial metabolic diversity and environmental variation together lead to changes in biological matter and energy flux. Time series[55] and mesocosm studies[56] are being used to investigate how microorganisms and their activities co-vary with environmental change. Efforts to integrate microbial diversity and process data with quantitative models that incorporate physical oceanography and biogeochemistry are still in their infancy[11,24,54,56–59]. Momentum is building, however, and direct observations of microbial diversity, variability and processes will soon inform models that will in turn inform and direct further field-oriented surveys, experiments and measurements. Observation, experiment and theory can together provide, verify and integrate information from genomics, metagenomics, microbial physiology, biogeochemistry and ecology, creating a clearer picture of emergent properties in the microbial systems that drive energy and matter flux in ocean ecosystems. The challenges to integrating work across disciplinary and conceptual boundaries are formidable, but the need for a more interdisciplinary understanding of the microbial ocean is clear. The reward will be a greatly improved qualitative and quantitative perspective on the living ocean system, from genomes to biomes. ■

1. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734–740 (1997).
2. Staley, J. T. & Konopka, A. Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* **39,** 321–346 (1985).
3. Rappe, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu. Rev. Microbiol.* **57,** 369–394 (2003).
4. Giovannoni, S. & Stingl, U. The importance of culturing bacterioplankton in the 'omics' age. *Nature Rev. Microbiol.* **5,** 820–826 (2007).
5. Stingl, U., Tripp, H. J. & Giovannoni, S. J. Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. *ISME J.* **1,** 361–371 (2007).
6. Connon, S. A. & Giovannoni, S. J. High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. *Appl. Environ. Microbiol.* **68,** 3878–3885 (2002).
   This is the first report of a dilution-to-extinction cultivation approach that was successful in isolating a wide variety of the predominant marine bacterioplankton types.
7. Chisholm, S. W. *et al.* A novel free-living prochlorophyte occurs at high cell concentrations in the oceanic euphotic zone. *Nature* **334,** 340–343 (1988).
8. Chisholm, S. W. *et al. Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll *a* and *b*. *Arch. Microbiol.* **157,** 297–300 (1992).
9. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424,** 1047–1051 (2003).
10. Lindell, D. *et al.* Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449,** 83–86 (2007).
11. Follows, M. J., Dutkiewicz, S., Grant, S. & Chisholm, S. W. Emergent biogeography of microbial communities in a model ocean. *Science* **315,** 1843–1846 (2007).
12. Rappe, M. S., Connon, S. A., Vergin, K. L. & Giovannoni, S. J. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418,** 630–633 (2002).
    This paper reports the first isolation in pure culture of a strain from the SAR11 clade, a representative of one of the most abundant bacterial groups in marine plankton.
13. Giovannoni, S. J. *et al.* Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438,** 82–85 (2005).

14. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309,** 1242–1245 (2005).
15. Tripp, H. J. *et al.* SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452,** 741–744 (2008).
16. Tripp, H. J. *et al.* Unique glycine-activated riboswitch linked to glycine–serine auxotrophy in SAR11. *Environ. Microbiol.* **11,** 230–238 (2009).
17. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37–43 (2004).
18. Hallam, S. J. *et al.* Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc. Natl Acad. Sci. USA* **103,** 18296–18301 (2006).
19. Allen, E. E. *et al.* Genome dynamics in a natural archaeal population. *Proc. Natl Acad. Sci. USA* **104,** 1883-1888 (2007).
20. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
21. Rusch, D. B. *et al.* The *Sorcerer II* Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5,** e77 (2007).
22. Yooseph, S. *et al.* The *Sorcerer II* Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5,** e16 (2007).
This was the first attempt to identify protein-family clusters globally, based on public data combined with some 6 million newly predicted peptides from a metagenomic sampling of surface-water marine bacterioplankton.
23. Tyson, G. W. *et al.* Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. *Appl. Environ. Microbiol.* **71,** 6319–6324 (2005).
24. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311,** 496–503 (2006).
25. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452,** 629–632 (2008).
26. Zehr, J. P. *et al.* Globally distributed uncultivated oceanic N_2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322,** 1110–1112 (2008).
27. Béjà, O. *et al.* Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289,** 1902–1906 (2000).
This paper reports the first observation and characterization of ion-pumping rhodopsins in the domain Bacteria, a discovery enabled by metagenomic sampling and analyses.
28. Béjà, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. Proteorhodopsin phototrophy in the ocean. *Nature* **411,** 786–789 (2001).
29. Sabehi, G. *et al.* Novel proteorhodopsin variants from the Mediterranean and Red Seas. *Environ. Microbiol.* **5,** 842–849 (2003).
30. de la Torre, J. R. *et al.* Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc. Natl Acad. Sci. USA* **100,** 12830–12835 (2003).
31. Sabehi, G., Beja, O., Suzuki, M. T., Preston, C. M. & DeLong, E. F. Different SAR86 subgroups harbour divergent proteorhodopsins. *Environ. Microbiol.* **6,** 903–910 (2004).
32. Sabehi, G. *et al.* New insights into metabolic properties of marine bacteria encoding proteorhodopsins. *PLoS Biol.* **3,** e273 (2005).
33. Frigaard, N. U., Martinez, A., Mincer, T. J. & DeLong, E. F. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439,** 847–850 (2006).
34. McCarren, J. & DeLong, E. F. Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. *Environ. Microbiol.* **9,** 846–858 (2007).
35. Martinez, A., Bradley, A. S., Waldbauer, J. R., Summons, R. E. & DeLong, E. F. Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. *Proc. Natl Acad. Sci. USA* **104,** 5590–5595 (2007).
36. Gomez-Consarnau, L. *et al.* Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* **445,** 210–213 (2007).
37. Stingl, U., Desiderio, R. A., Cho, J. C., Vergin, K. L. & Giovannoni, S. J. The SAR92 clade: an abundant coastal clade of culturable marine bacteria possessing proteorhodopsin. *Appl. Environ. Microbiol.* **73,** 2290–2296 (2007).
38. Béjà, O. *et al.* Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415,** 630–633 (2002).
39. Cho, J. C. *et al.* Polyphyletic photosynthetic reaction centre genes in oligotrophic marine Gammaproteobacteria. *Environ. Microbiol.* **9,** 1456–1463 (2007).
40. Fuchs, B. M. *et al.* Characterization of a marine gammaproteobacterium capable of aerobic anoxygenic photosynthesis. *Proc. Natl Acad. Sci. USA* **104,** 2891–2896 (2007).
41. Treusch, A. H. *et al.* Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic Crenarchaeota in nitrogen cycling. *Environ. Microbiol.* **7,** 1985–1995 (2005).
This is the first report presenting definitive metagenomic evidence for the occurrence of nitrification-associated genes (encoding ammonia monooxygenase subunits A and B) in non-thermophilic Crenarchaeota.
42. Konneke, M. *et al.* Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437,** 543–546 (2005).
This is the first report to prove the existence of chemolithoautotrophic, ammonia-oxidizing, non-thermophilic Crenarchaeota by their isolation in pure culture.
43. Hallam, S. J. *et al.* Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol.* **4,** e95 (2006).
44. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449,** 804–810 (2007).
45. Mou, X., Hodson, R. E. & Moran, M. A. Bacterioplankton assemblages transforming dissolved organic compounds in coastal seawater. *Environ. Microbiol.* **9,** 2025–2037 (2007).
46. Neufeld, J. D., Wagner, M. & Murrell, J. C. Who eats what, where and when? Isotope-labelling experiments are coming of age. *ISME J.* **1,** 103–110 (2007).
47. Poretsky, R. S. *et al.* Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* **71,** 4121–4126 (2005).
48. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl Acad. Sci. USA* **105,** 3805–3810 (2008).
This is the first report of a directed, massively parallel approach that sequenced a sample of the transcriptome in planktonic microbial assemblages by using pyrosequencing.
49. Shi, Y., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459,** 266–269 (2009).
50. Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnol.* **26,** 541–547 (2008).
51. Markowitz, V. M. *et al.* IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* **36,** D534–D538 (2008).
52. Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P. & Frazier, M. CAMERA: a community resource for metagenomics. *PLoS Biol.* **5,** e75 (2007).
53. Meyer, F. *et al.* The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9,** 386 (2008).
54. Anderson, R. *et al.* A new vision of ocean biogeochemistry after a decade of the Joint Global Ocean Flux Study (JGOFS). *Ambio* **10,** 4–30 (2001).
55. Karl, D. M., Bidigare, R. R. & Letelier, R. M. Long-term changes in plankton community structure and productivity in the North Pacific Subtropical Gyre: the domain shift hypothesis. *Deep-Sea Res. II* **48,** 1449–1470 (2001).
56. Karl, D. M. Microbial oceanography: paradigms, processes and promise. *Nature Rev. Microbiol.* **5,** 759–769 (2007).
57. DeLong, E. F. Towards microbial systems science: integrating microbial perspective, from genomes to biomes. *Environ. Microbiol.* **4,** 9–10 (2002).
58. Doney, S., Abbott, M., Cullen, J., Kar, I. D. & Rothstein, L. From genes to ecosystems: the ocean's new frontier. *Front. Ecol. Environ.* **2,** 457–466 (2004).
59. Fuhrman, J. A. *et al.* Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc. Natl Acad. Sci. USA* **103,** 13104–13109 (2006).
60. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
61. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–138 (2009).
62. Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl Acad. Sci. USA* **105,** 1176–1181 (2008).
63. Binga, E. K., Lasken, R. S. & Neufeld, J. D. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J.* **2,** 233–241 (2008).
64. Ishoey, T., Woyke, T., Stepanauskas, R., Novotny, M. & Lasken, R. S. Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.* **11,** 198–204 (2008).
65. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.* **7,** 19 (2007).
66. Marcy, Y. *et al.* Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl Acad. Sci. USA* **104,** 11889–11894 (2007).
67. Sorek, R. *et al.* Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318,** 1449–1452 (2007).
68. Béjà , O. *et al.* Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2,** 516–529 (2000).
69. Pham, V. D., Konstantinidis, K. T., Palden, T. & DeLong, E. F. Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical profile in the North Pacific Subtropical Gyre. *Environ. Microbiol.* **10,** 2313–2330 (2008).
70. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **34,** 5623–5630 (2006).
71. Krause, L. *et al.* Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics* **22,** e281–e289 (2006).
72. Krause, L. *et al.* Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.* **36,** 2230–2239 (2008).
73. von Mering, C. *et al.* Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315,** 1126–1130 (2007).
74. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308,** 554–557 (2005).
75. Dalevi, D. *et al.* Annotation of metagenome short reads using proxygenes. *Bioinformatics* **24,** i7–i13 (2008).
76. Harrington, E. D. *et al.* Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc. Natl Acad. Sci. USA* **104,** 13913–13918 (2007).
77. Rodriguez-Brito, B., Rohwer, F. & Edwards, R. A. An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7,** 162 (2006).
78. Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl Acad. Sci. USA* **106,** 1374–1379 (2009).
Canonical correlation analyses identified metabolic pathways in metagenomic data sets that maximally co-varied with multiple environmental variables, revealing co-variation of amino-acid transport and cofactor synthesis across communities and environments.

Acknowledgements I thank my current and former students, colleagues and co-workers for sharing their ideas, insights, enthusiasm and inspiration. Work in my laboratory is supported by grants from the US National Science Foundation, the US Department of Energy, the Gordon and Betty Moore Foundation and the Agouron Institute. This article is a contribution from the NSF Science and Technology Center, and the Center for Microbial Oceanography: Research and Education.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (delong@mit.edu).