The microbial ocean from genomes to biomes

Edward F. DeLong

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, 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 was the recognition that only a small proportion of the microbial cells sampled from the environment can be readily cultivated using conventional techniques. 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. 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. These and other results from cultivation-independent surveys have fundamentally changed our perspective on microbial phylogenesis, 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.

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, was successfully cultivated shortly after its discovery. 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 isolines. Prochlorococcus isolates have also been used in detailed studies of phage diversity, host range, genome content, host–phage genetic exchange and gene-expression dynamics. 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, suggesting promising future directions in microbial oceanography.

The development of ‘dilution to extinction’ cultivation techniques 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. 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.

1Departments 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. 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*\(^1\), a member of perhaps the most abundant bacterial group in the oceans. Isolates of *P. ubique* are now yielding fresh data on the phenotype\(^{6,11}\), genome content\(^1\), genetic variability\(^{11}\) and physiology\(^{11,12}\) 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\(^5\) and have provided comprehensive genomic representations of dominant populations\(^6,13,14\). 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\(^1\). 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 oxygen photosynthesis that are found in all other known free-living cyanobacteria\(^{26}\). The metagenomic data suggest that these cyanobacteria are not oxygen-generating phototrophs. 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 (SR86)\(^{27}\). Biophysical and functional characterization of the bacteriorhodopsin 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\(^{28-31}\). Subsequent genome sequencing of cultivated marine isolates then confirmed the widespread distribution of rhodopsin genes in many taxa of marine Bacteria\(^{28-31}\). Similarly, metagenomics revealed new types of aerobic, anaerobic photosynthetic
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 genes68 (as well as specific phylogenetic groups69,70) may be under-represented in genomic and metagenomic clone libraries. However, problems with such biases have been largely overcome by pyrosequencing 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 prediction71,72, phylogenetic assignment73,74 and functional predictions75,76,77 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 examined78.

Such advances, alongside customized metagenomic databases79,80, 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 comparisons75, bootstrap resampling methods that evaluate gene-abundance confidence intervals deviating from the median in pairwise sample comparisons75, canonical discriminant analyses that identify the genes that most influence distributional variance79, and canonical correlation analyses that interrelate metabolic-pathway occurrence with multiple environmental variables80. 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 individuals44. 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 plankton81, an observation that was later confirmed by strain-isolation studies82,83.

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 studies84,85 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 CO₂ as the sole carbon source yielded an ammonia-oxidizing crenarchaeal isolate86. 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 pathways87,88. 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. 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 point44.

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. 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.

The application of gene-expression technologies to track microbial sensing and responses in the environment is another exciting development. In this approach, bacterial and archaean 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. 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. 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 (Figs 2, 3).

Early high-throughput, pyrosequence-based studies 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. The transcriptional data sets in such studies contain several categories of RNA, including rRNAs, messenger RNAs and small RNAs, 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
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.

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)²⁰. 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/miniseq). 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 databases²¹⁻⁵, 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 data²⁵ 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.


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. 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.
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.


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).