



# Integration of genomics and proteomics into marine microbial ecology

Torsten Thomas, Suhelen Egan, Dominic Burg, Charmaine Ng, Lily Ting, Ricardo Cavicchioli\*

School of Biotechnology and Biomolecular Sciences, The University of New South Wales, Sydney, New South Wales 2052, Australia

**ABSTRACT:** Genomics and proteomics of microorganisms are revolutionizing our understanding of marine microbial ecology. In this essay we address this by discussing (1) what microbial genome resources are available for marine ecologists, (2) how single-organism genomics and proteomics have revealed new microbial functions in the marine ecosystem, and (3) how the integration of metagenomics, metaproteomics and biogeochemical studies will further advance the field of marine microbial ecology. Comprehensive knowledge of the genetic blueprints, the functions and the interactions of microbial communities will provide insight into the evolution of marine ecosystems and enable rational predictions of how microbial processes will affect, and be affected by, environmental changes.

**KEY WORDS:** Environmental genomics · Metagenomics · Metaproteomics · Proteomics · Marine ecology · Marine microbiology

*Resale or republication not permitted without written consent of the publisher*

## INTRODUCTION

Marine microbial ecology has advanced over the last decade through a progression of approaches that has included taxonomic and physiological studies of culturable isolates, molecular community analysis (e.g. rRNA), analyses of complete genome sequences of individual isolates and, most recently, metagenomic analyses of entire microbial communities. Concomitant with technological advances of the genomic era has been an exponential increase in the extent of data pertaining to marine microorganisms. This has provided on the one hand an enormous capacity to learn, while on the other a daunting overload of information. It is clear that irrespective of the quantity, information lacks real value unless the intelligent means are available to process it effectively. This essay reflects on how genomics and proteomics may empower marine ecological studies. It considers strengths and difficulties of marine genomics and proteomics, and discusses the need to integrate these data with the full gamut of all available data (e.g. physical, geochemical) that de-

scribe the marine system. It is apparent that getting the most out of the genome stockpile will require healthy and informed interplay among scientists in many disciplines.

## MICROBIAL GENOME RESOURCES FOR MARINE ECOLOGISTS

In the last 10 years, microbial genomics has experienced one of the most dramatic developments and advances of any scientific field. Microbiologists are now facing a genomic 'data flood' with more than 300 finished, bacterial or archaeal genome sequences available and over 900 more in progress (Liolios et al. 2006). Improved sequencing technologies and strategies (Margulies et al. 2005, Goldberg et al. 2006, Zhang et al. 2006) will continue to support this trend. In the near future, genome sequencing of new microorganisms will become a standard tool for microbial characterisation, analogous to the use of Gram staining in the past. Large-scale sequencing programs, such as

\*Corresponding author. Email: r.cavicchioli@unsw.edu.au

the Microbial Genome Sequencing Project of the Betty and Gordon Moore Foundation, are already involved in the sequencing of numerous (>100) marine microbial species ([www.moore.org/microgenome](http://www.moore.org/microgenome)). Several of these organisms have been isolated in global diversity studies, and genome sequencing will now fast-track the understanding of their biology.

There would be few scientists who would argue against the immense value of the growing number of marine microbial genome sequences, and most microbiologists engaged in laboratory-based physiological or evolutionary studies will have effectively integrated available genome-based knowledge into their research. However, the integration of microbial genomics into marine ecology has not been as rapidly or as widely adopted. This may be due in part to the traditional background training of marine ecologists, which has focused less on the properties of individual organisms and more on the broader properties of the ecosystem, and to the lack of ecological data linked to genome sequence data of marine microorganisms (see last paragraph of this section). Some of this can be remedied immediately by simply tapping into available web-based resources (e.g. becoming familiar with what type of information is available and beginning to find out fundamental information about target organisms). Comprehensive databases and user-friendly, web-based interfaces have made genomic information increasingly accessible, without the need for specialized bioinformatics training or knowledge. The Integrated Microbial Genome (IMG) database of the Joint Genome Institute (JGI) and the Comprehensive Micro-

bial Resource (CMR) of The Institute of Genomic Research (TIGR) are just 2 examples of the excellent tools available that allow the user to view, browse, analyse and compare microbial genome information. Specialized interest groups also provide databases dedicated to particular microbial groups, such as the Roseobase (<http://roseobase.org/>), which deals with genomic information of the abundant, marine *Roseobacter* clade. Table 1 lists some databases and web-based tools relevant to marine microbial genomics.

There are a number of issues in genomics that broadly affect the genomics community and that haven't been resolved, and there are additional aspects that need to be addressed in order to effectively facilitate ecological studies. Maintaining data quality is a broadly important issue with genome sequence data. The sheer volume of DNA sequence data in combination with limited human resources has made it extremely difficult to carefully and manually revise and curate the data. This has resulted in genome sequences being wrongly assembled from raw data (Salzberg & Yorke 2005) and automated gene prediction or annotation processes being inaccurate (Nielsen & Krogh 2005). Database users should therefore be cautious with predicted genome properties (particularly from auto-annotation pipelines) and be aware of the need to critically review the evidence for assigned gene function. For example, if a gene has been annotated based on experimental evidence or high similarity to a gene that has been experimentally characterised, the functional prediction is likely to be sound. However, there are numerous examples of annotations

Table 1. Web resources for marine, microbial genomics

Name	Weblink	Description
Integrated Microbial Genomes	<a href="http://img.jgi.doe.gov">http://img.jgi.doe.gov</a>	Comparative database, all publicly available genomes
Comprehensive Microbial Resource	<a href="http://cmr.tigr.org">http://cmr.tigr.org</a>	Comparative database, all publicly available microbial genomes
Microbial Genome Database for Comparative Analysis	<a href="http://mbgd.genome.ad.jp">http://mbgd.genome.ad.jp</a>	Comparative database, all publicly available microbial genomes
ERGO	<a href="http://ergo.integratedgenomics.com">http://ergo.integratedgenomics.com</a>	Private, comprehensive database
Center for Biological Sequence Analysis	<a href="http://www.cbs.dtu.dk/index.shtml">www.cbs.dtu.dk/index.shtml</a>	Comprehensive database and several web-based tools
Megx.net	<a href="http://www.megx.net">www.megx.net</a>	Database resource for marine ecological genomics; in development
Camera	<a href="http://camera.calit2.net">http://camera.calit2.net</a>	Cyberinfrastructure for marine microbial ecology research and analysis
Roseobase	<a href="http://roseobase.org">http://roseobase.org</a>	Specialised database for marine <i>Roseobacter</i> strains
Cyanobase	<a href="http://www.kazusa.or.jp/cyano">www.kazusa.or.jp/cyano</a>	Specialised database for cyanobacterial genomes
Moore Foundation Microbial Genome Sequencing Project	<a href="http://www.moore.org/microgenome">www.moore.org/microgenome</a>	links to maps and genome database

(classification of gene function) linked with low or high similarity to a gene with poorly defined properties. For example, a number of genes in Archaea, likely to be aminopeptidase genes (Ando et al. 1999), have been annotated as cellulase genes. Owing to an initial mis-annotation in 1 genome, subsequent archaeal genes with high levels of similarity were consequently mis-annotated. As there is presently no easy solution to this widespread problem, it argues strongly for individuals to carefully examine important gene targets of interest and to enhance the level of functional analysis of genes in order to experimentally determine their functions.

Genome databases have mainly been designed to enhance understanding of the biology and evolution of individual organisms, and there is clearly a need to include information that is relevant to ecological research (Lombardot et al. 2006). Database fields that are lacking include information describing the physical habitat from where an organism has been isolated or is typically present (e.g. temperate or tropical waters, planktonic or surface-associated), additional biological information about the environment (e.g. competitors, viruses, predators), information about seasonal and spatial abundance of organisms in the community, and a summary of relevant oceanography and physico-chemical properties (e.g. O<sub>2</sub>, minerals, salinity, dissolved/particulate organic carbon). This would allow macrobiotic or abiotic parameters to be linked with molecular or genomic properties, and hence provide a straightforward bridge between ecology and genomics. Engaging ecologists with database designers/curators would help to create more ecologically useful databases.

## MARINE ECOLOGY IS ALREADY BENEFITING FROM MICROBIAL GENOMICS AND PROTEOMICS

### Genome sequences

In an analogous manner to the analysis of the human genome to predict specific drug targets and candidates for gene therapy, genome sequencing of ecologically relevant microorganisms and microbial communities (metagenomics) can provide new insight or generate testable hypotheses about ecosystem function (see 'The way forward with marine microbial ecology is through an integrated 'meta' approach'). A striking example is the discovery of the light-dependent proton pump, bacterial proteorhodopsin, which was first discovered from the sequences of cloned environmental DNA (eDNA), and is thought to play a major role in the generation of energy for microbial metabolism in the oceans (Beja et al. 2000, 2001). Prior to this discovery,

light-driven processes were mainly linked to processes such as primary production by photosynthetic cyanobacteria. Another good example is the prediction of archaeal-driven nitrification processes derived from the analysis of metagenomic data (Schleper et al. 2005, Hallam et al. 2006), and the verification of this ability through the isolation of a chemolithoautotrophic ammonia-oxidizing member of the Crenarchaeota (Konneke et al. 2005).

The sequencing of genomes of single microbial species is also clearly of value for deriving inferences about ecology of marine bacteria. For example, genomic studies of ubiquitous planktonic bacteria (the SAR11 isolate *Pelagibacter ubique*, and a member of the Roseobacter clade *Silicibacter pomeroyi*) have greatly enhanced our understanding of how some microorganisms have adapted and evolved to become numerically abundant within the marine environment (Moran et al. 2004, Giovannoni et al. 2005). *P. ubique* has the smallest known genome (1.3 Mb) of any free-living microorganism, and points to an evolutionary adaptive strategy involving genome streamlining; i.e. optimizing growth efficiency by minimizing the genomic and cellular complement that needs to be reproduced in order for the species to survive and remain evolutionarily competitive (Giovannoni et al. 2005). Despite the relatively small size of the genome, *P. ubique* still possesses the capacity to synthesise all 20 amino acids and all core functions required for a free-living bacterium. The small genome size appears to have been selected through a process leading to the minimization of non-functional DNA, extra-chromosomally derived genetic elements (e.g. phage, integrons or transposons), and duplicated genes. Comparative studies with genome sequences of species from the same ecosystem indicate that adaptation to oligotrophy in this organism involves a low level of gene regulation and an investment in genes devoted to energy metabolism and high-affinity nutrient uptake.

*Silicibacter pomeroyi* is a dominant member of the coastal bacterioplankton (Moran et al. 2004). Based on genome sequence analysis, it appears that *S. pomeroyi* takes an opportunistic strategy towards nutrient acquisition. Genes for cell-density-dependent regulation, rapid growth and uptake systems for algal-derived compounds are present, suggesting that the organism is capable of associating with nutrient-rich hot-spots such as algal plankton and other suspended particles. Furthermore, the presence of gene clusters encoding enzymes for the oxidation of reduced inorganic compounds (e.g. carbon monoxide and sulphide) suggests that *S. pomeroyi* is a lithoheterotroph that gains energy from inorganic compounds and uses organic carbon compounds that are at low abundance for generating bacterial biomass. Arising from this genome

sequence analysis, a range of experimental studies can be designed and performed to assess the proposed metabolic capabilities and ecological function of *S. pomeroyi*. In view of the apparent wide-spread capacity for lithoheterotrophy that has been deduced from the analysis of metagenome data (e.g. the Sargasso Sea library; Venter et al. 2004) and the potential impact of this on global nutrient cycling in the oceans (Moran et al. 2004), performing functional studies to better understand this process is of considerable importance (see 'Functional genomics' below).

As great advances in marine ecology can come from the analysis of genome sequence data of individuals and communities of marine microorganisms, a strong argument can be made for a greater emphasis to be placed on the training of scientists with the necessary expertise in genomics, in order to more fully exploit the potential of genomic data for marine ecology research. In particular, there is an important need to develop synergies between bioinformaticians and ecologists/biologists, in order to translate the raw stock piles of genome sequence data into valuable science.

### Functional genomics

Global functional studies, such as proteomics and transcriptomics, have the potential to most rapidly advance our understanding of functional cellular processes, and hence likely ecological processes. Studies relating to how an organism (or community) responds to environmental change provide insight into core physiological properties and adaptive strategies. Technological advances in the functional 'omics' have developed in concert with genome sequencing technology, particularly owing to the need for high-throughput procedures to keep pace with the growth in genome sequence data. Metagenomic data is relatively new (see 'The way forward with marine microbial ecology is through an integrated 'meta' approach') and, as a result, functional 'omic' studies of marine microorganisms have almost exclusively been linked to genome sequences of individual organisms.

Mass spectrometry (MS)-based proteomics provides a powerful means of determining proteins expressed under 1 growth condition (proteome snap-shot), or by comparing at least 2 growth conditions (differential expression). Proteomic coverage can be maximized by reducing the complexity of samples through the use of fractionation regimes (e.g. to identify less-abundant proteins) and the sampling of sub-proteomes (e.g. intracellular, membrane, secreted). This type of approach was used to analyse the proteome of the marine bacterium *Alcanivorax borkumensis* in order to determine the metabolic functions involved in petroleum

degradation (Sabirova et al. 2006). Whole-cell, soluble fractions are typically used for proteomic analysis. However, it is important to pay attention to other fractions. Secreted proteins may play important roles in antimicrobial activity and cell-cell signalling (Milton 2006). Membrane sub-fractions are technically challenging to analyse but are likely to provide important insight into the mechanisms of how cells sense and respond to their environment. Although 2-dimensional gel electrophoresis (2DE)-based methods are useful (e.g. for fractionation, visualizing post-translational modifications), recent developments in more rapid 'shotgun' approaches using liquid chromatography mass spectrometry (LC-MS) have proven particularly valuable for analysing less-soluble sub-proteomes (e.g. hydrophobic proteins) by providing rapid, high-throughput and broad coverage of these proteins (Wu & Yates 2003, Martosella et al. 2006).

It is not only important to successfully identify proteins, but to accurately quantify protein levels in order to determine the abundance of individual proteins relative to other proteins in the cell (i.e. differential expression). There are 3 major MS-based approaches for quantifying protein levels: 2DE-MS, intensity-based quantification, and stable isotope labeling. Stable isotope labeling is the most comprehensive approach for globally measuring protein abundances and can be performed by *in vitro* (e.g. isotope coded affinity tag: ICAT) and *in vivo* (e.g. metabolic labeling) approaches (Gygi et al. 1999, Krijgsveld et al. 2003, Zhong et al. 2004).

The method developments and refinements of approaches in the field of proteomics (Wilkins et al. 2006) should help to make this technology accessible and immensely useful to the microbial marine biology/ecology field. Reflective of the way in which proteomic methodology has developed, studies of marine microorganisms have primarily involved 2DE-MS approaches (Goodchild et al. 2004a, Gade et al. 2005, Kan et al. 2005, Kim et al. 2005, Sabirova et al. 2006). However, 2DE-MS, LC-MS and ICAT have been applied to *Methanococcoides burtonii* (Goodchild et al. 2004a,b, 2005), and metabolic labeling combined with DNA microarrays with *Methanococcus maripaludis* (Xia et al. 2006). The ability to successfully apply these methods to fastidious, strict anaerobes highlights the potential ease of application of these types of methods to many microorganisms.

Studies of the marine, surface associated bacterium *Pseudoalteromonas tunicata* provides a good example of how genomics and functional genomics can be used to generate new hypotheses about marine ecology. The genome sequence not only provided detailed knowledge of the ability of *P. tunicata* to associate with eucaryal hosts and synthesise novel bioactive metabo-

lites, but enabled comparative proteomic and transcriptomic studies between a wildtype and a mutant (Stelzer et al. 2006). These studies showed that the mutant, which no longer produced bioactives, exhibited an unexpected overexpression of genes involved in iron scavenging and sensing (Stelzer et al. 2006). As a result of these findings, awareness was created about the likely ecological relevance of iron, and this has prompted a series of experimental programs that target the role of iron in bacterial-host interactions in the marine environment.

### THE WAY FORWARD WITH MARINE MICROBIAL ECOLOGY IS THROUGH AN INTEGRATED 'META' APPROACH

Tools for characterising the diversity of marine microorganisms have progressed through 3 phases. Initially, marine microbial populations were characterised by isolation and culturing of strains. However, it is now well-accepted that cultivation introduces large qualitative and quantitative biases into ecological studies (Suzuki et al. 1997, Eilers et al. 2000) and only canvases a very small proportion of the total marine diversity (Rappe & Giovannoni 2003). This is primarily a result of the fact that most microorganisms are unable to be cultured using current methods, highlighting the need for access to culture-independent technologies. A second phase arose through molecular ecology studies, which allowed non-culturable bacteria and Archaea to be characterised via molecular fingerprinting, specific molecular probes and sequencing of selected genes (e.g. 16S rDNA and selected functional genes). These PCR-based methods suffer from biased amplification of target sequences and often fail to correctly reflect community composition (Suzuki et al. 1997, Marchesi et al. 1998, von Wintzingerode et al. 1999, Schmalenberger et al. 2001). A third approach, 'metagenomics' (also termed 'environmental genomics', 'ecogenomics' or 'community genomics') has recently emerged, which involves the extraction of DNA from all microorganisms (or size-fractionated components of all microorganisms) from the environment (Handelsmann 2004, Riesenfeld et al. 2004). The eDNA is either cloned as small or large fragments into *Escherichia coli* plasmids and sequenced by the method of Sanger (1977) (e.g. using an ABI 3730 sequencer) (Tyson et al. 2004, Venter et al. 2004), or sequenced directly by high-throughput pyrosequencing (e.g. using a GS20/454 sequencer) (Margulies et al. 2005, Edwards et al. 2006). The eDNA plasmid libraries represent the genomes of the environmental population of microorganisms irrespective of whether the microorganisms are culturable, and can also be

used for functional or phylogenetic screening (Handelsmann 2004, Riesenfeld et al. 2004).

The extent of DNA sequencing that is required for the successful reconstruction of genome sequences of microbial communities is directly proportional to the complexity of the environment. Preliminary molecular ecology studies can provide a useful indication of species richness and therefore an estimation of the number of sequencing reactions required. Metagenomic studies of less complex communities are not only less expensive (per sample site) and more easily analysed (e.g. reconstruction of genome sequences of individual species), but are more amenable to metafunctional studies. The metagenome (Tyson et al. 2004) and metaproteome (Ram et al. 2005) study of a biofilm from the acid mine drainage of Iron Mountain is a powerful illustration of what can be achieved in microbial ecology when using a meta-approach. From less than 100 Mb of sequence data, genome sequences for the dominant bacterium (*Leptospirillum* Group II) and archaeon (*Ferroplasma* Type II) and partial genome sequences of several others were obtained. Reconstruction of metabolic pathways led to inferences about nitrogen fixation, which subsequently enabled a successful cultivation strategy to be derived for *Leptospirillum ferrodiazotrophum*, a previously uncultured organism (Tyson et al. 2005).

Metaproteome analysis of the biofilms from the Iron Mountain site led to up to 48% of the predicted proteins being identified from an individual member of the biofilm, a percentage that exceeds the number of proteins typically detected from proteomic studies of microbial isolates. For example, in proteomic studies of Archaea, proteome coverage has been reported as approximately 50% for *Methanocaldococcus jannaschii*, 25% for *Methanococcoides burtonii*, 10% for *Methanosarcina acetivorans* and 34% for *Halobacterium salinarum* (Cavicchioli et al. 2006). In proteomic studies of isolates, monocultures are typically grown in nutrient excess under controlled conditions of growth phase and abiotic influence (e.g. temperature, pH). It is likely that the acid mine biofilm contained cells exhibiting a broad range of phenotypes in response to varying levels of nutrient limitation, interactions with other microorganisms, growth state (e.g. actively growing, dead, planktonic, sessile) and other natural, undefined environmental effectors.

Technological advancement in genome sequencing and proteomics shows no sign of plateauing. Therefore, these meta-approaches will become increasingly feasible for application to more complex environmental samples. Moreover, genomic/proteomic and metagenomic/metaproteomic programs can be run in parallel to more effectively annotate genome sequences and obtain a direct measure of functional gene expression in terms of the presence, relative

abundance and modification states of proteins. The potential of, and challenges for, meta-based studies of microbial communities have been evaluated (Banfield et al. 2005, Foerstner et al. 2006, Ward 2006, Wilmes & Bond 2006), and methods for the preparation of DNA and proteins from soil/sediment or water from environmental samples have been reported (Tsai & Olsen 1991, Purdy et al. 1996, Miller et al. 1999, Schulze 2004, Daniel 2005, Kan et al. 2005, Schulze et al. 2005, Tringe & Rubin 2005).

Isolating representative DNA from communities in soil/sediment is probably one of the most challenging of all natural environmental samples, owing to the complexity and the diversity of microbial populations ( $\geq 2 \times 10^4$  bacterial or archaeal species per gram of soil; Daniel 2005), and the fact that microbial cells and free DNA from dead cells adhere to the soil/sediment matrix. DNA can be extracted directly (cells are lysed within sample material) or indirectly (cells are first separated from the environmental sample) (Miller et al. 1999, Daniel 2005), and similar approaches have been adopted for extracting proteins (e.g. Schulze et al. 2005). Marine water samples are easier to manipulate than sediment and tend to have lower complexity. Microbial biomass can be collected by size-fractionated filtration on membrane filters and tangential flow centrifugation, and methods have been developed for subsequent protein extraction and analysis (Schulze 2004, Venter et al. 2004, Kan et al. 2005).

In addition to the use of a meta-approach for obtaining information about microorganisms that are difficult to study (e.g. non-culturable) (Rodriguez-Valera 2004, Tringe & Rubin 2005), proteogenomic studies of individual isolates can also form the basis for rationalizing the need for meta-studies. *Sphingopyxis alaskensis* was isolated as a numerically abundant microorganism from Resurrection Bay in Alaska and oligotrophic waters near Japan, and has served as a model ultramicrobacterium (Cavicchioli et al. 2003). A broad range of laboratory studies including proteomics (e.g. Ostrowski et al. 2004) have defined physiological characteristics that distinguish it from typical copiotrophic bacteria, such as *Photobacterium angustum* S14 (Cavicchioli et al. 2003). Despite being isolated by extinction dilution methods and representing a numerically abundant organism at the time of sampling, *S. alaskensis* has not been reported to be as widely distributed as SAR11, which is apparently one of the most cosmopolitan microorganisms in oligotrophic oceanic waters. Metagenomics of distinct oceanic sites along the path of the Sorcerer II expedition (Venter et al. 2004) have revealed an astounding level of total microbial genetic diversity. To date, metagenome studies have not included North Pacific waters where *S. alaskensis* was

isolated. It has been proposed that *S. alaskensis* may circulate between locations that are 10 000 km apart by ocean currents in the North Pacific (Eguchi et al. 2001), and it will be valuable to assess the genomic variation that exists between populations of *S. alaskensis* from the geographically distinct regions of the North Pacific from where it was previously isolated. Moreover, when the analysis of *S. alaskensis* genome sequence is complete, similarities and differences with SAR11 will be able to be documented. It is already clear that *S. alaskensis* has a significantly larger genome (~3.2 Mb) than SAR11 (1.3 Mb). It has previously been argued that multiple strategies may have evolved to enable microorganisms to compete effectively in oligotrophic waters (Cavicchioli et al. 2003). It will be valuable to assess the metaproteome of both *S. alaskensis* and SAR11 in their native environments to determine which component of their genetic complement is expressed, and to infer how this may affect their individual adaptation strategies.

The value of individual microorganisms guiding metagenome studies is also well illustrated by studies of psychrophilic Archaea. Cold-adapted Archaea perform diverse functional roles in a wide range of cold environments, and the extent to which they transform the cold biosphere can be appreciated from their phylogenetic and functional diversity, abundance and range of cold biotopes they inhabit (Cavicchioli 2006). They represent an important fraction of cold marine environments and have been detected in deep ocean waters and sediment, sea ice and marine-derived Antarctic lakes. *Methanococcoides burtonii* was isolated from a marine-derived lake (Ace Lake) in Antarctica, and through studies of cold adaptation that addressed protein structure, intracellular solutes, membrane lipids, tRNA modification, gene regulation, comparative genomics and proteomics, it has developed into the model psychrophilic archaeon (Cavicchioli 2006). In addition to *M. burtonii*, *Methanogenium frigidum* (Ace Lake) and *Halorubrum lacusprofundi* (Deep Lake) were isolated from Antarctica. The studies of these individual isolates from Antarctic lakes have generated a broad range of questions that can be addressed most successfully by performing metagenomic and associated functional studies. The types of questions include: are genes that are preferentially expressed at 4°C under laboratory growth conditions (and therefore thought to be involved in cold adaptation) expressed in the environment (Goodchild et al. 2004a, 2005)? Are genes that have been linked to genome plasticity (e.g. transposons) (Goodchild et al. 2004b) expressed in the environment, and how does this affect the overall microheterogeneity of species such as *M. burtonii*, *M. frigidum* and *H. lacusprofundi*? Which hypothetical proteins are synthesized and

therefore important for growth of the organism in the environment (Saunders et al. 2005)?

By coupling metagenomic analysis with a range of other experiments performed at the time of sampling, including isolations, physical and chemical measurements and labeling experiments, and integrating this with established physical and geochemical data of the lakes, a comprehensive understanding of the microbial system can be defined. The metagenomics will particularly facilitate the ability to define community structure, individuals within the communities and their associated biological processes. Key biological properties that underpin the biogeochemical process will also be able to be derived from the genomic properties of key microbial groups (e.g. methanogens and methylotrophs), as will an understanding of how key chemical cycles (e.g. methane) are microbially driven, and of what biological properties define life at abiotic limits (e.g. cold, hypersalinity, oligotrophy). While these examples pertain to specific Antarctic lakes and the archaeal isolates, the principles of the approaches are applicable to other environments and individual isolates from those environments. This clearly illustrates how studies of cultivated organisms can be greatly facilitated by subsequent metagenomic studies, in addition to the obvious advantages metagenomics offers to studies of uncultivated species in their respective environments.

### PERSPECTIVE

Metagenomic studies have identified the high level of microheterogeneity that can exist within populations. One of the first studies of this type documented the existence of 2 major variants of *Cenarchaeum symbiosum* that coexist in a marine sponge (Schleper et al. 1998). However, it is not clear what biological and abiotic factors control the extent and tempo of genomic heterogeneity. Metagenomics of samples from the Sargasso Sea highlighted the microheterogeneity within marine *Prochlorococcus marinus* populations (Venter et al. 2004), even though it is not clear over what time period this has occurred. Biofilms can exhibit large changes in genetic composition over their life-time (e.g. within a few days) (Webb et al. 2004, Mai-Prochnow et al. 2006), and may be major ecological drivers of genetic diversity in 'real-time', and hence model systems for studying genome evolution. In fact, natural biofilm populations have been shown to not contain discrete genome sequences for a particular species, but rather to possess a highly diverse, mosaic genomic complement (Tyson et al. 2004, Allen & Banfield 2005). In contrast to this dynamic system, microheterogeneous communities that double at very slow

rates (e.g. deep subsurface) may be considered repositories of genetic codes, rather than as genomic solutions that have evolved subject to the influences of modern-day perturbations.

Phylogenetic analysis of 16S rRNA genes was largely responsible for the revolution in evolutionary biology that led to the definition of the 3 domains of life (Woese et al. 1990), and to the great expansion of the diversity of known species. Genome sequencing of individual microorganisms corroborated the concept of the 3 domains of life, and uncovered the extent to which genetic diversity exists within apparently coherent species (e.g. *E. coli*). Metagenomics is revealing the extent to which genetic diversity exists within natural communities, and is challenging the concept of a microbial species. By understanding the extent, tempo and mode of genome evolution it will be possible to gain a practical understanding of microbial community evolution and infer the effects of human impact on the microbial gene pool, and greatly enrich our understanding of how life has evolved along its 3.8 billion yr old trajectory.

### LITERATURE CITED

- Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* 3:489–498
- Ando S, Ishikawa K, Ishida H, Kawarabayasi Y, Kikuchi H, Kosugi Y (1999) Thermostable aminopeptidase from *Pyrococcus horikoshii*. *FEBS Lett* 447:25–28
- Banfield JF, Verberkmoes NC, Hettich RL, Thelen MP (2005) Proteogenomic approaches for the molecular characterization of natural microbial communities. *OMICS* 9: 301–333
- Beja O, Aravind L, Koonin EV, Suzuki MT and 8 others (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289:1902–1906
- Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF (2001) Proteorhodopsin phototrophy in the ocean. *Nature* 411:786–789
- Cavicchioli R (2006) Cold adapted Archaea. *Nature Rev Microbiol* 4:331–343
- Cavicchioli R, Ostrowski M, Fegatella F, Goodchild A, Guixa-Boixereu N (2003) Life under nutrient limitation in oligotrophic marine environments: an eco/physiological perspective of *Sphingopyxis alaskensis* (formerly *Sphingomonas alaskensis*). *Microb Ecol* 45:203–217
- Cavicchioli R, Goodchild A, Raftery M (2006) Proteomics of Archaea (Chapter 5). In: Humphery-Smith I, Hecker M (eds) *Microbial proteomics—functional biology of whole organisms*. Wiley, New York
- Daniel R (2005) The metagenomics of soil. *Nature Rev Microbiol* 3:470–478
- Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M and 6 others (2006) Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. *BMC Genomics* 7:57
- Eguchi M, Ostrowski M, Fegatella F, Bowman J, Nichols D, Nishino T, Cavicchioli R (2001) *Sphingomonas alaskensis*, strain AF01: an abundant oligotrophic ultramicrobacterium from the North Pacific. *Appl Environ Microbiol* 67: 4945–4954

- Eilers H, Pernthaler J, Gloeckner FO, Amann R (2000) Culturability and *in situ* abundance of pleagic bacteria from the North Sea. *Appl Environ Microbiol* 66:3044–3051
- Foerster KU, von Mering C, Bork P (2006) Comparative analysis of environmental sequences: potential and challenges. *Phil Trans R Soc Lond B* 361:519–523
- Gade D, Theiss D, Lange D, Mirgorodskaya E and 8 others (2005) Towards the proteome of the marine bacterium *Rhodospirellula baltica*: Mapping the soluble proteins. *Proteomics* 5:3654–3671
- Giovannoni SJ, Tripp HJ, Givan S, Podar M and 10 others (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309:1242–1245
- Goldberg SM, Johnson J, Busam D, Feldblyum T and 15 others (2006) A sanger/pyrosequencing hybrid approach for generation of high quality draft assemblies of marine microbial genomes. *Proc Natl Acad Sci USA* 103:11240–11245
- Goodchild A, Saunders NFW, Ertan H, Raftery M, Guilhaus M, Curmi PMG, Cavicchioli R (2004a) A proteomic determination of cold adaptation in the Antarctic archaeon, *Methanococoides burtonii*. *Mol Microbiol* 53:309–321
- Goodchild A, Raftery M, Saunders NFW, Guilhaus M, Cavicchioli R (2004b) The biology of the cold adapted archaeon, *Methanococoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry. *J Proteome Res* 3:1164–1176
- Goodchild A, Raftery M, Saunders NFW, Guilhaus M, Cavicchioli R (2005) Cold adaptation of the Antarctic archaeon, *Methanococoides burtonii* assessed by proteomics using ICAT. *J Proteome Res* 4:473–480
- Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol* 17:994–999
- Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM, DeLong EF (2006) Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol* 4:e95
- Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 68:669–685
- Kan J, Hanson TE, Ginter JM, Wang K, Chen F (2005) Meta-proteomic analysis of Chesapeake Bay microbial communities. *Saline Syst* 19:7–19
- Kim YK, Yoo WI, Lee SH, Lee MY (2005) Proteomic analysis of cadmium-induced protein profile alterations from marine alga *Nannochloropsis oculata*. *Ecotoxicol* 14:589–596
- Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437:543–546
- Krijgsveld J, Ketting RF, Mahmoudi T, Johansen J, Artal-Sanz M, Verrijzer CP, Plasterk RHA, Heck AJR (2003) Metabolic labeling of *C. elegans* and *D. melanogaster* for quantitative proteomics. *Nature Biotechnol* 21:927–931
- Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* 34:D332–334
- Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glockner FO (2006) Megx.net—database resources for marine ecological genomics. *Nucleic Acids Res* 34:D390–393
- Mai-Prochnow A, Webb JS, Ferrari BC, Kjelleberg S (2006) Ecological advantages of autolysis during biofilm development and dispersal of *Pseudoalteromonas tunicata*. *Appl Environ Microbiol* 72:5414–5420
- Marchesi JR, Sato T, Weightman AJ, Martin TA, Fry JC, Hiom SJ, Dymock D, Wade WG (1998) Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Appl Environ Microbiol* 64:795–799
- Margulies M, Egholm M, Altman WE, Attiya S and 52 others (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
- Martosella J, Zolotarjova N, Liu H, Moyer SC, Perkins PD, Boyes BE (2006) High recovery HPLC separation of lipid rafts for membrane proteome analysis. *J Proteome Res* 5:1305–1312
- Miller DN, Bryant JE, Madsen EL, Ghiorse WC (1999) Evaluation and optimization of DNA extraction and purification procedures of soil and sediment samples. *Appl Environ Microbiol* 65:4715–4724
- Milton DL (2006) Quorum sensing in vibrios: complexity for diversification. *Int J Med Microbiol* 296:61–71
- Moran MA, Buchan A, Gonzalez JM, Heidelberg JF and 31 others (2004) Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* 432:910–913
- Nielsen P, Krogh A (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* 21:4322–4329
- Ostrowski M, Fegatella F, Wasinger V, Corthals G, Guilhaus M, Cavicchioli R (2004) Cross species identification of proteins from proteome profiles of the marine oligotrophic ultramicrobacterium, *Sphingopyxis alaskensis*. *Proteomics* 4:1779–1788
- Purdy KJ, Embley TM, Takii S, Nedwell DB (1996) Rapid extraction of DNA and rRNA from sediments by a novel hydroxyapatite spin-column method. *Appl Environ Microbiol* 62:3905–3907
- Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW and 5 others (2005) Community proteomics of a natural microbial biofilm. *Science* 308:1915–1920
- Rappe M, Giovannoni SJ (2003) The uncultured microbial majority. *Annu Rev Microbiol* 57:369–394
- Riesenfeld C, Schloss P, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38:525–552
- Rodríguez-Valera F (2004) Environmental genomics, the big picture? *Microbiol Lett* 231:153–158
- Sabirova JS, Ferrer M, Regenhart D, Timmis KN, Golyshin PN (2006) Proteomic insights into metabolic adaptations in *Alcanivorax borkumensis* induced by alkane utilization. *J Bacteriol* 188:3763–3773
- Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. *Bioinformatics* 21:4320–4321
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain termination. *Proc Natl Acad Sci USA* 74:5463–5467
- Saunders NFW, Goodchild A, Raftery M, Guilhaus M, Curmi PMG, Cavicchioli R (2005) Predicted roles for hypothetical proteins in the low-temperature expressed proteome of the Antarctic archaeon *Methanococoides burtonii*. *J Proteome Res* 4:464–472
- Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV (1998) Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. *J Bacteriol* 180:5003–5009
- Schleper C, Jurgens G, Jonuscheit M (2005) Genomic studies of uncultivated archaea. *Nat Rev Microbiol* 3:479–488

- Schmalenberger A, Schwieger F, Tebbe CC (2001) Effect of primers hybridizing to different evolutionarily conserved regions of the small-subunit rRNA gene in PCR-based microbial community analyses and genetic profiling. *Appl Environ Microbiol* 67:3557–3563
- Schulze W (2004) Environmental proteomics—what proteins from soil and surface water can tell us: a perspective. *Biogeosciences* 1:195–218
- Schulze WX, Gleixner G, Kaiser K, Guggenberger G, Mann M, Schulze ED (2005) A proteomic fingerprint of dissolved organic carbon and of soil particles. *Oecologia* 142: 335–343
- Stelzer S, Egan S, Larsen MR, Bartlett DH, Kjelleberg S (2006) Unravelling the role of the ToxR-like transcriptional regulator WmpR in the marine antifouling bacterium *Pseudoalteromonas tunicata*. *Microbiology* 152:1385–1394
- Suzuki MT, Rappe MS, Haimberger ZW, Winfield H, Adair N, Strobel J, Giovannoni SJ (1997) Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample. *Appl Environ Microbiol* 63: 983–989
- Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nature Rev Microbiol* 6: 805–814
- Tsai YL, Olson BH (1991) Rapid method for direct extraction of DNA from soil and sediments. *Appl Environ Microbiol* 57:1070–1074
- Tyson GW, Chapman J, Hugenholtz P, Allen EE and 6 others (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43
- Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, Banfield JF (2005) Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. *Appl Environ Microbiol* 71:6319–6324
- Venter J, Remington K, Heidelberg JF, Halpern AL and 19 others (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74
- von Wintzingerode F, Gobel UB, Stackebrandt E (1999) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* 21:213–219
- Ward N (2006) New directions and interactions in metagenomics research. *FEMS Microbiol Ecol* 55:331–338
- Webb JS, Lau M, Kjelleberg S (2004) Bacteriophage and phenotypic variation in *Pseudomonas aeruginosa* biofilm development. *J Bacteriol* 186:8066–8073
- Wilkins MR, Appel RD, Van Eyk JE, Chung MCM and 12 others (2006) Guidelines for the next 10 years of proteomics. *Proteomics* 6:4–8
- Wilmes P, Bond PL (2006) Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol* 14:92–97
- Woese CR, Kandler O, Wheelis ML (1990) Toward a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87: 4576–4579
- Wu CC, Yates JR (2003) The application of mass spectrometry to membrane proteomics. *Nature Biotechnol* 21:262–267
- Xia Q, Hendrickson EL, Zhang Y, Wang T and 6 others (2006) Quantitative proteomics of the archaeon *Methanococcus maripaludis* validated by microarray analysis and real time PCR. *Mol Cell Proteomics* 5:868–881
- Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM (2006) Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnol* 24:680–686
- Zhong H, Marcus SL, Li L (2004) Two-dimensional mass spectra generated from the analysis of <sup>15</sup>N-labeled and unlabeled peptides for efficient protein identification and de novo peptide sequencing. *J Prot Res* 3:1155–1163

*Editorial responsibility: Howard Browman (Associate Editor-in-Chief), Storebø, Norway*

*Submitted: June 29, 2006; Accepted: July 28, 2006  
Proofs received from author(s): February 7, 2007*