Evolving perspectives in microbial oceanography from genomes to biomes

The Microbial Ocean from Genomes to Biomes." Nature 459 7244 (2009): 200-06

Evolving perspectives in microbial oceanography from genomes to biomes

Edward F. DeLong, MIT, Cambridge MA 02139

Preface

The ocean's biota is numerically dominated by microbial species. Yet the population dynamics, metabolic complexity and synergistic interactions hidden within marine microbial assemblages remain largely uncharted. A full understanding of the living ocean requires moving beyond the 'parts list' of marine microbial taxa and genome sequences. New experimental techniques and analytical approaches have potential to provide the required model systems, experimental data, and ecosystem observations to facilitate construction of predictive models to interrelate microbial dynamics with the biogeochemical matter and energy fluxes that comprise the living ocean ecosystem.

Evolving perspective on microbial life in the oceans

At the same time astronauts were walking on the moon, the total number of (micro)organisms living in each milliliter of seawater had been underestimated by over three orders of magnitude ! While we were looking toward the skies, the majority of microbial life on Earth was left largely unknown, and certainly under appreciated. A large paradigm shift in microbial oceanography occurred in the late 1970s and early 1980s, when accurate estimates of total cell numbers and bulk rates of microbial production became available. Over the next 25 years or so, local, regional, and global estimates of microbial numbers, and their bulk production and consumption rates in ocean surface waters were quantified and mapped. These data have provided more accurate estimates of total planktonic microbial biomass and its turnover, enlarging the perceived role and significance of bulk microbial biomass and activity in oceanic food webs. While this information has been extremely useful, only recently has more specific information on the biology of planktonic microbial organisms become available. Questions that are now being more specifically addressed include : Which taxa of marine

Bacteria and *Archaea* are most dominant, representative, or biogoechemically relevant in particular ocean provinces or depth strata ? What are the most common microbial energy and carbon sources, and the most relevant catabolic and anabolic pathways ? How do dynamic population shifts (regular or sporadic) and species interactions shape the ecology and biogeochemistry of the seas ?

Unlike eukaryotic plankton, that can often be taxonomically and metabolically categorized based on directly observable phenotypes, planktonic *Bacteria* and *Archaea* have remained more enigmatic with respect to their core identifies and physiological properties. Consequently, advances in cultivation-independent metagenomics, as well as new cultivation technologies, have had specific and dramatic influence on our knowledge of non-eukaryotic microbes. The contributions and synergy of metagenomics and new cultivation techniques is therefore a main focus of this review. The take home lesson is that the integrated perspective provided by cultivation-independent phylogenetic surveys, microbial metagenomics and culture-based studies, all taken together, have critically advanced a more detailed understanding of microbial life in the sea. The utility of the integration of these approaches is fast becoming apparent. Some of recent highlights in new cultivation approaches and metagenomics are described below to emphasize these points.

Phylogenetic surveys, microbial isolates, and model systems

One of the drivers that inspired the development of cultivation-independent approaches for phylogenetically identifying naturally occurring microbes[1] was the general recognition that only a small percentage of total microbial cells can be readily cultivated from the environment using conventional cultivation techniques[2]. The development of rRNA-based phylogenetic surveys in the 1980s led to less biased assessments of the distribution of yet-uncultivated bacterial, archaeal and protistan phylotypes in natural populations[1]. The number of newly recognized major bacterial and archaeal phylogenetic divisions has increased dramatically. Indeed, in many habitats, some of the most abundant microbial phylotypes have no close relatives in culture.[3] These and other

results from cultivation-independent surveys have fundamentally changed our perspective on microbial phylogeny, evolution and ecology. These new discoveries subsequently inspired more directed cultivation strategies, aimed at isolating some of the more environmentally abundant microbial phylotypes that had eluded cultivation previously[4,5,6].

Directed cultivation efforts continue to play an important role for describing the nature and properties of marine *Bacteria* and *Archaea.* For example the ocean's most abundant cyanobacterium *Prochlorococcus,* first discovered via shipboard flow cytometery at sea[7], was successfully cultivated soon after its discovery[8]. *Prochlorococcus* isolates now provide an environmentally relevant system for modeling the biology and ecology planktonic cyanobacteria. Physiological characterization of *Procholococcus* genotypic variants led to the idea of "ecotypes", e.g. highly related yet physiologically and genetically distinct populations presumably adapted to different environmental regimes. An oceanographic survey of six *Prochlorococcus* ecotype variants in the Atlantic confirmed their distinct, environmental distributions on a basin-wide scale. *Procholococcus* isolates have also facilitated detailed studies of their phage diversity, host range, genome content, as well as host-phage genetic exchange[9] and gene expression dynamics[10]. Pointing towards promising future directions in microbial oceanography, the integration of *Prochlorococcus* lab-based physiological modeling and field–based surveys, have helped constrain and validate computational ecosystem models that successfully recapitulate known *Prochlorococcus* ecotype distributions in the environment[11].

The development of dilution-to-extinction cultivation techniques[4] represents a significant advance in directed cultivation efforts, aimed at culturing the newly discovered phylotypes discovered in rRNA-based environmental surveys. The basic approach involves preparing sterilized natural seawater, that is distributed into tissue culture wells and subsequently inoculated with serially diluted bacterioplankton[6]. Successful growth in these low density cultures is monitored by direct cell counting. The results of these approaches have been dramatically successful, with respect to recovery in pure culture of many dominant surface water bacterioplankton representatives[4,5,6,12]. As with any

approach however, there are practical limitations (e.g., uncertainties when using undefined and variable seawater media, low statistical probability of isolating rare organisms, etc.). In fact, the reasons why some predominant groups are readily cultivated, while 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 significant impact of these new cultivation approaches is especially evident in the recent isolation of *Pelagibacter ubique*[12], a member of perhaps the most abundant bacterial group in the oceans . *Pelagibacter ubique* isolates are now yielding new data on the phenotype[12,13], genome content[14], genetic variability[5,14], and physiology[15,16] of this major bacterioplankton taxon.

Cultivation of resident microbes is a necessary but not independently sufficient approach for describing microbial processes in the environment. While pure cultures provide readily manipulated models, there remain fundamental limitations to their utility for inferring ecological process. Although some physicochemical variables can be well controlled in cultures, patterns of temperature, pressure, pH, nutrient concentrations, and redox balance, and their naturally occurring gradients that occur in Nature, may sometimes be difficult to accurately reproduce in the laboratory. Additionally, many microorganisms have evolved to closely interact with other macro- and microorganisms, and often engaged in obligatory symbiotic relationships. For these and other reasons, it is unreasonable to presume that pure culture microbial models will be available for all ecologically important and relevant microbial types. Hence, cultivation-independent phylogenetic and genomic surveys will continue to play an important role in describing uncultured microbes, and their population genetics, biogeochemical and ecological interactions, that cannot be well studied modeled in laboratory systems.

The convergence of microbial metagenomics and cultivation studies

In the context of this review, ‘metagenomics’ is defined as cultivation-independent genomic analysis of microbial assemblages or populations. While still in its infancy,

metagenomics has already contributed significantly to our knowledge of the genomic structure, population diversity, gene content, and composition of naturally occurring microbial assemblages. In low complexity populations metagenomic studies have led to the assembly of near complete genomes from dominant 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 (see Box 1). Despite the large datasets now available however, high allelic variation in microbial populations, high species richness, and relatively even representation among species, still render whole genome assemblies of individual genotypes impractical, given current sequencing and assembly technologies[20,21,22] (Box 2).

Distinct and advantages and synergies are evident from the coupling of metagenomics and culture-based approaches. While no single methodology or approach is without shortcomings (see Box 2), metagenomic surveys have already contributed significantly to our understanding of microbes in the environment. For example, metagenomic datasets have enabled directed enrichment and isolation of new isolates, having specific and predicted functional and genetic properties[23]. In metagenomic surveys along environmental gradients, direct observation of gene distributions in the ocean's water column have revealed discernable patterns of vertical stratification of metabolic genes, bacteriophage, and transposases, and provided clues about the differential distribution of metabolic processes, phage-host interactions, and evolutionary processes along the depth continuum[24]. A more recent survey effort using new pyrosequencing technologies compared > 70 different marine metagenomic datasets, revealing statistically significant gene content differences among nine major biomes compared[25]. Finally, in a harbinger of the utility of cell-directed metagenomic approaches, the genome content of an uncultivated nitrogen-fixing cyanobacterium population ("UCYN-A") recovered via flow cytometry was recently reported[26]. The genomic 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 II. The big surprise was that critical genes in CO_2 fixation and oxygenic photosynthesis that are found in all other known free-living cyanobacteria, were absent in nitrogen-fixing cyanobacterium UCYN-A[26] ! The metagenomic data indicate that these cyanonbacteria are not oxygen-

generating photoautotrophs. This study provides an excellent example of leveraging metagenomics to identify the metabolic capabilities of yet-uncultivated 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 unanticipated presence of a bacteriorhodopsin-like photoprotein gene (proteorhodopsin) in an abundant marine bacterioplankon group (SAR86)[27]. Biophysical and functional characterization of this proteorhodopsin gene product confirmed its ability to function as a light-driven proton pump[27]. Following metagenomic surveys revealed the high abundance and global distribution of these rhodopsins in marine planktonic bacteria and archaea[20,21,28,29,30,31,32,33,34,35]. Later genome sequencing of cultivated marine isolates then confirmed the ubiquitous and widespread distribution of rhodopsin genes among commonly cultivated as well as the more fastidious marine bacteria[13,36,37] (Table 1). In a very similar way, novel phylogenetic types of aerobic, anoxygenic photosynthetic bacteria were observed in marine plankton via metagenomics[38], an observation that was later confirmed via strain isolation studies[39,40].

The predictive power of metagenomics was also demonstrated in the observation of genes associated with ammonia-oxidation in *Archaea*, a character previously known to be found in just a few bacterial groups. Two concurrent metagenomic studies[20,41] reported that a specific clade of *Crenarchaea* seemed to possess the genes diagnostic for chemolithotrophic ammonia oxidation. Soon thereafter, 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 a near-complete genome sequence from an uncultured crenarchaeon extended prior studies beyond a single gene in the pathway, and suggested specific functional differences between the archaeal and bacterial ammonia–oxidizing metabolic pathways[18,43].

These and other examples have clearly indicated the value gained from integrating and

cross-comparing metagenomic and culture-based studies. Indeed, the deficiencies of each approach to a large extent are compensated by the strengths of the other. On the one hand, phenotype, as well as metabolism and physiology, are mainly inferred from laboratory culture experiments. On the other hand, 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. The new perspectives enabled by integration of metagenomics, cultured-based studies, and environmental surveys intersect in a knowledge domain previously not readily accessible to microbiologists (Figure 1). More specifically, integration of cultivation-dependent and cultivation-independent approaches in part bridge the gaps between genomics, population genetics, biochemistry, physiology, biogeochemistry and ecology (Figure 1). Integrative studies employing of cultivation and metagenomic perspectives will undoubtedly be pursued more deliberately and more frequently, in future collaborative microbiological inquiries. Current and evolving plans for human microbiome studies represent a good case in point[44].

"Real time" microbial process experiments using nucleic acid sequences as analytes

Development of metagenomic methods has been important for expanding the repertoire of known microbial genes, their environmental distributions and allelic diversity. While the associated bioinformatic analyses help to generate new hypotheses, other means are required to test and verify *in silico* hypotheses and conclusions in the real world. Major gaps exist between simply describing the naturally occurring microbial "parts list", to understanding the functional properties, the multi-scalar responses and interdependencies that connect microbial and abiotic microbial ecosystem processes. To expand our understanding of the ocean's microbial "parts list", and its context within microbial ecosystem dynamics, new experimental approaches will be required. Experimental approaches that can leverage massively parallel sequencing technologies, or that can otherwise link information from pre-existing sequence datasets with experimental observations in natural assemblages, seem particularly promising.

Several approaches are available that have potential to link microbial community DNA

sequences with specific microbes and their activities in the environment. One method uses the thymidine analogue bromodeoxyuridine (BrDU) to tag actively growing substrate-responsive cells. After labeling, the BrDU-labeled DNA of actively growing cells 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 experimentally tracking specific microbial groups incorporating labeled organic or inorganic compounds into living tissues. Stable isotope tracers have been successfully used to identify methanotrophic archaea at cold, localize nitrogen-fixing symbionts in host tissues, and verify autotrophic metabolism in planktonic *Crenarchaea*. A novel approach having potential to directly link DNA sequence information with substrate-specific incorporation is stable isotope probing (SIP), where nucleic acids labeled with "heavy" isotope are physically isolated via buoyant density centrifugation, and subsequently sequenced[46]. Both stable isotope and BrDU-labeling strategies have potential to couple *in situ* physiological responses with metagenomic sequence data.

Application of gene expression technologies to track microbial sensing and response in the environment represents an exciting recent development. In this approach, bacterial and archaeal total RNA is extracted from microbial assemblages, converted to cDNA and directly sequenced (Figure 2). Early studies began by preparation and analysis of randomly primed cDNA clone libraries via Sanger-based capillary sequencing, to survey abundant transcripts from a coastal seawater sample[47]. Advances in technologies like pyrosequencing, which sidesteps the need for clone libraries, have enabled analyses of larger datasets from more rapidly collected, smaller volume (hundreds of ml) samples from marine bacterioplanton[48]. Pyrosequencing of both genomic DNA and cDNA from the same sample facilitates normalization of transcript abundance to the corresponding gene copy number of the community nucleic acid pool[48] (Figure 2).

Several new insights have resulted from early high throughput, pyrosequence-based microbial community transcriptomic studies[48]. Not surprisingly, genes associated with key metabolic pathways of open ocean microbial species (including photosynthesis, carbon fixation, and nitrogen acquisition for example) were highly expressed in the photic zone at 75 m depth in the North Pacific Subtropical Gyre. Both genomic and

transcriptomic datasets exhibited high coverage of some dominant community members like *Prochlorococcus*, in which hypervariable genomic regions displayed some of the highest transcript abundances. While many of the planktonic microbial community transcripts were similar to previously predicted genes found in ocean metagenomic surveys, a large fraction (~ 50%) appeared unrelated to predicted protein sequences in available databases[48]. The transcriptomic datasets in such studies contain several different categories of RNA species, including ribosomal RNAs, messenger RNAs and recently recognized small RNAs some of which play important role in regulation of gene expression. Each of these molecular species recovered from total microbial assemblage RNA pool – rRNA, mRNA, and sRNA – has potential to provide significant new insight on the phylogenetic, functional, and regulatory properties and dynamics of natural microbial communities.

Development of microbial community transcriptomic methods is enabling a new research agenda in microbial ecology, that utilize sequence data as an analyte in experimental field studies. The approach enables the measurement of microbial assemblage gene expression in microcosms, mesocosms or natural samples, as a function of environmental variability over time (Figure 3). The environmental variation examined can be natural (for example, tracking changes in gene expression as a function of the diel cycle), or applied (for example, monitoring changes in gene expression following nutrient emendation). By tracking genes responsive to specific environmental perturbations, it should soon be possible to track environmental perturbations that are first observable as changes in gene expression in resident microbial populations, but that later may lead to shifts in community composition. Quantifying the variability and kinetics of gene expression in natural assemblages has potential to provide a fundamentally new perspective on microbial community dynamics. Can expression patterns provide clues as to the functional properties of hypothetical genes ? What are the key community responses to natural or anthropogenic environmental perturbation ? Are there fundamental community-wide regulatory responses common to disparate taxa ? Are certain taxa or metabolic paths more or less responsive to particular environmental changes ? Are specific changes in gene expression indicative of downstream changes in community composition ? These and other questions can now be more directly

addressed by applying these newly developing experimental approaches.

Information management from genes to ecosystems

One of the major challenges in emerging metagenomic and 'metatranscriptomic' studies are the sheer size of the datasets, and the new methods and tools that are needed to deal with their magnitude. Size matters. These datasets raise new issues with respect to data management, computational resources, sampling and analytical strategies, and database architectures. An encouraging development has been recognition in the research community of the need to establish clear standards for metadata submission and reporting, so that primary sequence data can be related across relevant environmental parameters. The Genomic Standards Consortium (GSC) are promoting schemes reminiscent of the MIAME standards for microarray data (http://www.mged.org/Workgroups/MIAME/miame.html), that would capture metadata associated with genomes (Minimum Information about a Genome Sequence, MIGS), and metagenomic data (Minimum Information about a Metagenome Sequence, MIMS)[49,50]. For archived datasets, such metadata field standardization and reporting will be critical. As mentioned above, we are entering a new era in microbial ecology and biology, that will increasingly employ high-throughput sequencing data as an analyte in experimental protocols. Coordination of experimental reports from such inquiries will be important, and MIAME-like standards for such reporting (Minimum Information about a high-throughput SeQuencing Experiment – MINSEQE) have recently been proposed as well (http://www.mged.org/minseqe/). Even "simple" annotation, archiving, and accessing of the new sequence data types and experiments, along with associated and relevant metadata, poses significant challenges for the biological community. These challenges are now beginning to be addressed by the development of new types of metagenomic databases[51,52,53], analytical strategies and statistical approaches (see 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 still

exists a significant mismatch with respect to integrating "bottom up" reductionist molecular approaches, with "top down" integrative ecosystems analyses. Molecular datasets are often gathered in massively parallel ways, but acquiring equivalently dense microbial and biogeochemical process data[54] is not currently as feasible. This 'impedance mismatch' (e.g., inadequate (or excessive) ability of one system to accommodate the input from another), is one of the larger hurdles that will have to be overcome for more realistic, integrative analyses that interrelate datasets spanning from genomes to biomes.

The road ahead

While the microbial "parts list" of genes and genomes that are contained in microbial assemblages is growing rapidly, their functional and ecological relevance is not as currently well constrained. DNA sequence data and bioinformatic analyses alone fall short of describing which gene suites and metabolic pathways are actually being expressed and utilized within any given environmental context. How do community composition, gene content, and variability influence biogeochemical function, turnover rates, and ecosystem process? How important is functional redundancy and allelic diversity to community function and stability? How does the process of succession play out, from initial environmental change, to individual cellular sensing and response, to community compositional shifts? Can particular functional properties or gene categories be assigned specific probabilities, with respect to their likelihood of lateral gene transfer and genomic fixation ? Can suites of genes and their variability be correlated with larger scale biogeochemical and ecological patterns and processes ? What are the functional properties and roles of as-yet uncharacterized proteins, that share little or no homology with functionally annotated proteins or protein-encoding genes ? How representative are the activities and responses of microbial isolates in the laboratory, compared to their physiological and metabolic behavior in the environment ? New approaches will be required to address these and other questions currently being raised in the course of both new cultivation efforts and metagenomic surveys.

New strategies to bridge the gaps between microbial genomics, metagenomics, biochemistry, physiology, population genetics, biogeochemistry, oceanography and ecosystem biology need to be further developed and explored. Integrative and transdisciplinary interactions will be key in future inquiries, since microbial diversity, metabolism, and biogeochemistry are all intertwined at multiple temporal and spatial scales. One central hypothesis driving metagenomic inquiry is that the network instructions for community metabolic processes, biogeochemical function, and ecological interactions are encoded in the collective microbial genomes, and expressed by the environmental responses and kinetics of gene regulation. The instantiated network instructions eventually weave their way into ecosystem process (Figure 4). Microbial metabolic diversity itself, interacting with environmental variation, drives the lion's share of biological matter and energy flux in the sea. Via time series efforts[55] and mesocosm studies[56], the temporal domain, from hours, to days, to weeks, to years for microbial variability are beginning to be investigated. Currently, efforts to integrate microbial diversity and process data with quantitative models that incorporate physical oceanography and biogeochemistry are yet in their infancy[11,24,54,56,57,58,59]. It seems clear however that momentum is building, and just as direct observations of microbial diversity, variability and process will inform models, the models may soon begin to drive field-oriented questions, surveys, experiments and measurements. Together observation, experiment and theory can provide, verify, and integrate new information from genomics and metagenomics, microbial physiology and biogeochemistry, and ecology. A clearer picture of emergent properties comprising the microbial systems that drive energy and matter flux in ocean ecosystems should emerge. While the challenges for integrating efforts across multiple disciplinary and conceptual boundaries are formidable, the marching orders and critical need to advance a more transdisciplinary understanding of the microbial ocean are clear. The challenges for such integration are many and substantial, yet the potential rewards offer a greatly improved qualitative and quantitative perspective of the living ocean system, from genomes to biomes.

Acknowledgements. I thank my current and former students, colleagues and coworkers for generously sharing their ideas, insights, enthusiasm, and inspiration. This work is supported by grants from the NSF, DOE, The Gordon and Betty Moore Foundation, and

the Agouron Institute. This article is a contribution from the National Science Foundation Science and Technology Center, the Center for Microbial Oceanography: Research and Education (C-MORE).

Table 1. Proteorhodopsins found in large genome fragments and cultures

Origin	Taxonomic affiliation	*PR photosystem ?	Source/Reference
Cultivated Strains			
Methylophilales sp.	Betaproteobacteria	+	60
Rhodobacterales sp.	Alphaproteobacteria	+	GBMF
Vibrio angustum	Gammaproteobacteria	+	GBMF
Photobacterium sp.	Gammaproteobacteria	+	GBMF
Marine gammaproteobacteria	Gammaproteobacteria	+	37
Pelagibacter ubique	Alphaproteobacteria	-	13
Rhodospirillales sp.	Alphaproteobacteria	+	GBMF
Marinobacter sp.	Gammaproteobacteria	+	GBMF
Vibrio campbelli	Gammaproteobacteria	+	GBMF
Dokdonia sp.	Bacteroidetes	brp only	GBMF
Polaribacter dokdonensis	Bacteroidetes	brp only	36
Psychroflexus sp.	Bacteroidetes	brp only	36
Polaribacter irgensii	Bacteroidetes	brp only	GBMF
Flavobacteria sp.	Bacteroidetes	brp only	GBMF
BACs or fosmids			
EBAC31A08	Gammaproteobacteria	-	27
HOT2C01	Unknown	-	30
ANT32C12	Gammaproteobacteria	-	30
eBACHOT4E07	Gammaproteobacteria	-	31
EBAC20E09	Gammaproteobacteria	-	31
MED13K09	Unknown	+	32
MED18B02	Alphaproteobacteria	-	32
MED35C06	Gammaproteobacteria	-	32
MED42A11	Alphaproteobacteria	-	32
MED46A06	Alphaproteobacteria	+	32
MED49C08	Gammaproteobacteria	-	32
MED66A03	Alphaproteobacteria	+	32
MED82F10	Unknown	+	32
MED86H08	Alphaproteobacteria	-	32
RED17H08	Unknown	+	32
RED22E04	Unknown	+	32
HF70_39H11	Euryarchaea	-	33
HF10_3D09	Euryarchaea	-	33
HF70_19B12	Euryarchaea	-	33
HF70_59C08	Euryarchaea	-	33
HF10_05C07	Proteobacteria	-	34
HF10_45G01	Proteobacteria	-	34
HF130_81H07	Gammaproteobacteria	-	34
EB0_39F01	Alphaproteobacteria	+	34
EB0_39H12	Proteobacteria	-	34
EB80_69G07	Alphaproteobacteria	-	34
EB80_02D08	Gammaproteobacteria	-	34
EB0_35D03	Proteobacteria	-	34
EB0_49D07	Proteobacteria	-	34
EBO_50A10	Gammaproteobacteria	-	34
EB0_55B11f	Alphaproteobacteria	-	34
EBO_41B09	Betaproteobacteria	+	34
HF10_49E08	Planctomycetales	+	34
HF10_12C08	Alphaproteobacteria	-	34
HF10_19P19	Proteobacteria	+	35
HF10_25F10	Proteobacteria	+	35

* PR Photosystem + indicates the presence and co-location of genes required for retinal (beta carotene synthesis and cleavage) and proteorhodpsin biosynthesis. “brp only” indicates the presence and genetic linkage of the proteorhodopsin and beta carotene cleavage genes only.

Figure Legends.

Figure 1. Integration and intersection of traditional disciplines and metagenomics. The large circles represent the fundamental elements of study – genes, organisms and the environment. General areas of investigation associated with each element are indicated in the text. The dual intersections between the elements identify their disciplinary overlaps including genomics, 'gene ecology' (metagenomics), and ecology. The red cross-hatched area identifies the "sweet spot" where information from cultured-based studies, and environmental studies, and metagenomics can be combined.

Figure 2. Transritpome sequencing protocol for aquatic microbial assemblages. Cell biomass is collected are then processed to produce genomic DNA, or cDNA from total RNA, as previously reported[48]. Samples for RNA are collected in smaller volumes (< 1 liter) and filtered as rapidly as possible (~ 10 minutes). After RNA amplification and conversion to cDNA, both cDNA and genomic DNA extracted from the same assemblage are sequenced and compared.

Figure 3. Quantifying microbial response to environmental change with environmental transcriptomics. An example of field experiments now made possible by leveraging tandem metagenomic and metatranscriptomic pyrosequencing. Incubations are established with aquatic communities in large volume microcosms (1). The untreated sample controls for intrinsic incubation effects as well as natural diel variation in gene expression. The different experimental treatments could measure a variety of physical or environmental perturbations, including the effects of light, nutrients, temperature, etc. A hypothetical sampling scheme for microbial assemblage RNA and DNA is shown (2). Microbial assemblage DNA and RNA subsamples are then subjected to pyrosequencing and analyses (3), as described in Figure 2.

Figure 4. The network instructions encoded in microbial genomes drive ecosystem process. A cartoon showing the connections between microbial assemblage genomic information, and the collective ecological interactions and community metabolism that in part regulate and sustain biogeochemical and ecosystem process.

BOX 1. Evolving genomic technologies

The inventory of genomic and metagenomic data now available for marine microbes is expanding rapidly for a variety of reasons. Firstly, acquisition of whole microbial genome sequences from cultivated microbial strains continues to become much faster and cheaper, so genome sequences are accumulating rapidly, with thousands now currently in the "pipleline". With respect to marine microorganisms, hundreds of bacterial and archaeal whole or draft genome sequences are currently available in the public databases. In addition, nucleic avid sequences recovered directly from total microbial assemblages are fast outstripping available whole genome sequence data. The drivers for this include increasing awareness of the utility of such data, a few major expeditions that have contributed large volumes of shotgun sequence data advancing technologies[61] that are democratizing acquisition of large amounts of sequence data by individual investigators.

In addition to the size of metagenomic datasets, the heterogeneity of data types and environments sampled is also expanding dramatically. Original datasets mainly included Sanger-based shotgun sequence data of cloned DNA captured in small insert clone libraries (~ 3kb), or longer genome fragments (40 -100 kbp) in BACs or fosmids. More recently, pyrosequencing techniques[61] that do not required DNA clone libraries (eliminating associated labor and cost overheads), have rapidly evolved from initial read lengths of 100 bp, to average reads of ~250 bp, and most recently is yielding average read lengths approaching 450 bps in size. Other "next generation" technologies that involve sequencing by synthesis but generate very short reads (~25bp) may prove useful in metagenomics as well, if sufficient long-read reference databases are available. Not far on the horizon are new technologies that will enable even higher throughput, longer read, single-molecule sequencing[62,63]. These new advances are game changers, with respect to the volumes of data that can and will be collected and bioinformatic infrastructure that will be required for analyses and syntheses to occur.

On the horizon, single cell genome sequencing using Multiple Displacement Amplification (MDA) techniques coupled with new sequencing technologies also appear very promising for gaining better genomic access to uncultivated or rare

microorganisms[64,65,66]. While these techniques have great potential, there remain significant challenges[65,66]. Chief among these are extraneous DNA 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[66]. With considerable effort partial draft genomes can be produced from single cells, but currently not without extraordinary efforts to reduce contamination and normalize for uneven coverage[64,67]. Nevertheless, incremental improvements in single genome sequencing in the future are likely to enable to recovery of more partial draft genomes from yet-uncultivated *Bacteria* and *Archaea*. These are expected to both provide and derive benefit from the more 'traditional' metagenomic approaches currently in common use.

Box 2. Caveats, complications, and conundrums in metagenomic methods

The technical constraints of microbial sampling, the changing landscape of sequencing technologies, and the sheer complexity and size of the datasets, all present significant challenges and uncertainties for interpreting and comparing microbial community genomic data. A few of the larger challenges are listed below – many others are not due to the space constraints of this short review.

Sampling cells and molecules: 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 influenced by the type of the microbial community, the environment that it occupies, the spatial scale sampled, population densities, and the presence contaminating substances. Relevant questions include: Do the cells need to be purified away from a soil, sediment or rock matrix ? To reduce sample complexity, will the cells be size separated from larger eukaryotic species ? Do the cells need to be concentrated before DNA extraction ? These and other concerns about sampling are central to downstream interpretation of resultant datasets. In addition to cell sampling, the methods used to recover and sequence microbial community DNA are 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. Problems with such biases however have been largely overcome by pyrosequencing[61] and other "next generation" sequencing technologies that sidestep the need to clone individual DNA molecules.

Functional gene predictions and annotation. Even preliminary tasks of gene characterization including calling open reading frames (ORFs), identifying taxonomic origins, and inferring functional properties, are nontrivial enterprises in analyses of metagenomic datasets. Complicating factors include short sequence read lengths, sequence quality, the absence of gene-linkage context, extremely large datasets, and uneven coverage. A number of strategies for metagenomic ORF prediction[22,71,72],

phylogenetic assignment[73,74], and functional predictions[22,75,76] have recently been developed. Improvements and new approaches to these fundamental tasks continue to evolve. In an example of progress in this area, a recent study combining homology searches and gene neighborhood analyses succeeded in specific functional gene predictions for 76% of the 1.4 Mbp examined[77]. These advances, in tandem with customized databases for metagnomics databases[51,52,53], promise to improve current capabilities for gene identification and annotation of metagenomic datasets.

Comparative metagenomic analyses. Statistical approaches for comparison of metagenomic datasets have only recently been applied, and their development is at an early stage. The size of the datasets, 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 so far have included gene enrichment probability estimates in three-way comparisons[75], bootstrap re-sampling methods that evaluate gene abundance confidence intervals deviating from the median in pair-wise sample comparisons[78], canonical discriminant analyses that identify those genes most influencing distributional variance[25], and canonical correlation analyses that inter-relate metabolic pathway occurrence with multiple environmental variables[79]. So far however, only very disparate sample types (dead whales, silage, acid mine drainage biofilms, coral bioflims, animal guts, etc) have been the subject of much statistical scrutiny. It will very 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 occurring between microbiomes of human individuals[44]. As the availability of datasets and comparable metadata fields continues to improve, quantitative statistical metagenomic comparisons are likely to increase in their utility and resolving power.

Literature cited

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. *Nat 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).

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. *Archives of Microbiology* **157** (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).

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 U S A* **103**, 18296-18301 (2006).

19 Allen, E. E. *et al.* Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci U S A* **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).
[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 N2-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).
[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, doi:493 [pii] (2003).
[30] de la Torre, J. R. *et al.* Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc Natl Acad Sci U S A* **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 U S A* **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 U S A* **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).
[42] Konneke, M. *et al.* Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**, 543-546 (2005).
[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 U S A* **105**, 3805-3810 (2008).
[49] Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**, 541-547 (2008).
[50] Garrity, G. M. *et al.* Toward a standards-compliant genomic and metagenomic publication record. *Omics* **12**, 157-160 (2008).
[51] Markowitz, V. M. *et al.* IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* **36**, D534-538 (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*, 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 Research II* **48**, 1449-1470 (2001).
[56] Karl, D. M. Microbial oceanography: paradigms, processes and promise. *Nat 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, l. 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 U S A* **103**, 13104-13109 (2006).

[60] Giovannoni, S. J. *et al.* The small genome of an abundant coastal ocean methylotroph. *Environ Microbiol* **10**, 1771-1787 (2008).
[61] Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).
[62] Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* 133-138 (2009).
[63] Korlach, J. *et al.* Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc Natl Acad Sci U S A* **105**, 1176-1181 (2008).
[64] 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).
[65] 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).
[66] Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7**, 19 (2007).
[67] 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 U S A* **104**, 11889-11894 (2007).
[68] Sorek, R. *et al.* Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**, 1449-1452 (2007).
[69] Béjà , O. *et al.* Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**, 516-529 (2000).
[70] 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* (2008).
[71] Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34**, 5623-5630, (2006).
[72] Krause, L. *et al.* Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics* **22**, e281-289 (2006).
[73] Krause, L. *et al.* Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* **36**, 2230-2239 (2008).
[74] von Mering, C. *et al.* Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**, 1126-1130 (2007).
[75] Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554-557 (2005).
[76] Dalevi, D. *et al.* Annotation of metagenome short reads using proxygenes. *Bioinformatics* **24**, i7-13 (2008).
[77] Harrington, E. D. *et al.* Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc Natl Acad Sci U S A* **104**, 13913-13918 (2007).

[78] Rodriguez-Brito, B., Rohwer, F. & Edwards, R. A. An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**, 162 (2006).

[79] Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* **106**, 1374-1379 (2009).

Figure 1.

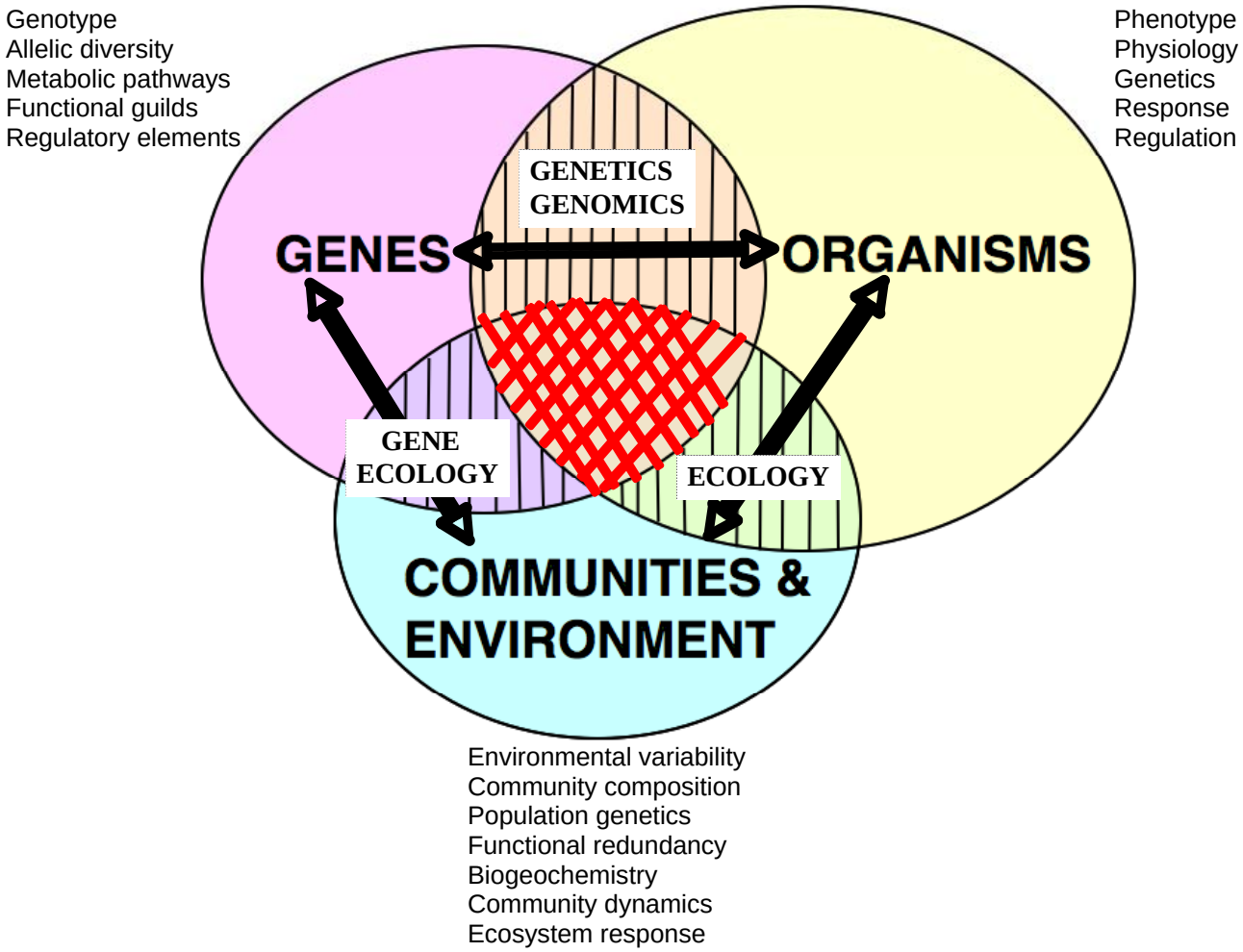

Genotype
Allelic diversity
Metabolic pathways
Functional guilds
Regulatory elements
Phenotype
Physiology
Genetics
Response
Regulation
GENETICS
GENOMICS
GENES
ORGANISMS
GENE
ECOLOGY
ECOLOGY
COMMUNITIES &
ENVIRONMENT
Environmental variability
Community composition
Population genetics
Functional redundancy
Biogeochemistry
Community dynamics
Ecosystem response

Figure 2.

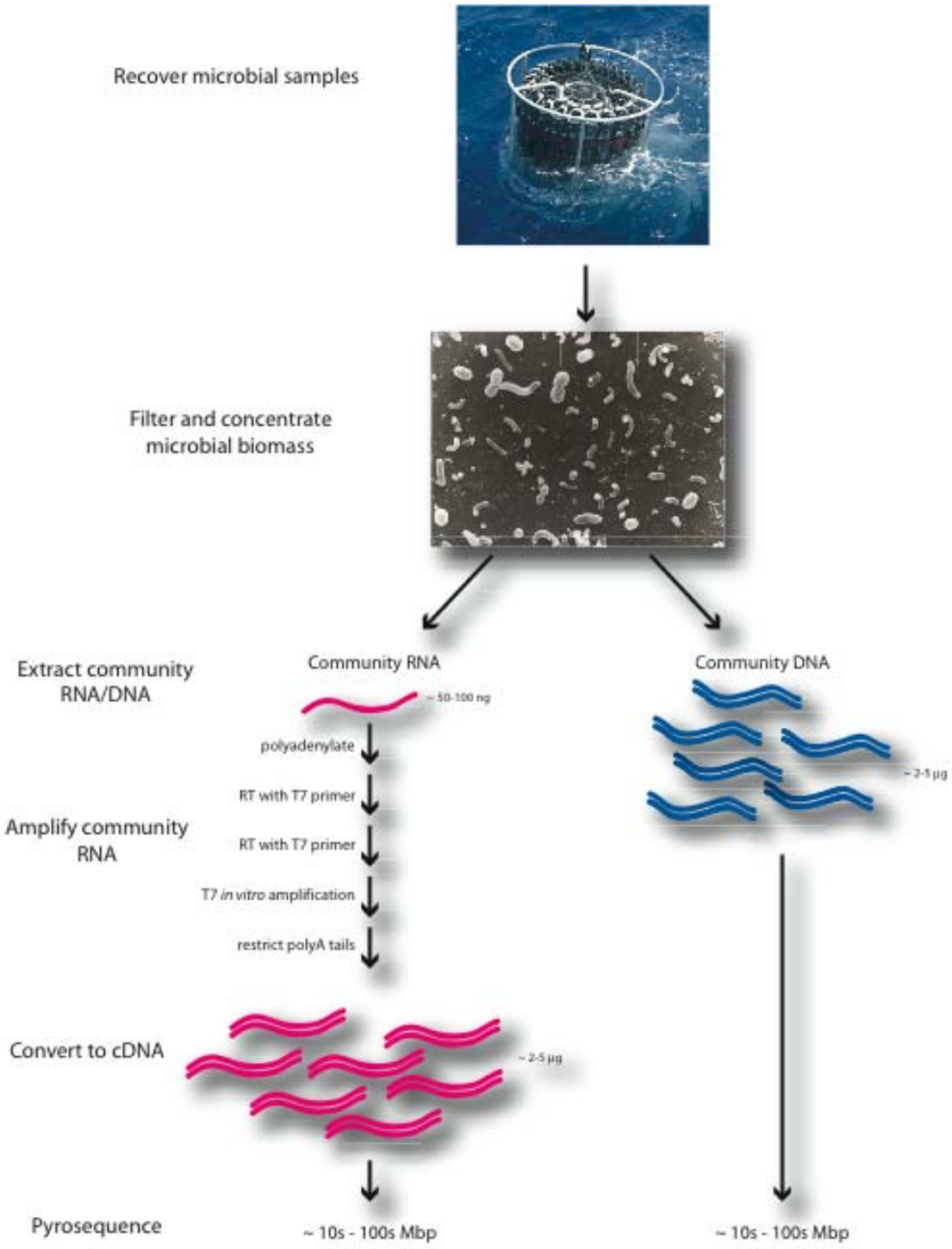

Recover microbial samples
Filter and concentrate microbial biomass
Extract community RNA/DNA
Community RNA
~ 50-100 ng
Community DNA
~ 2-5 µg
polyadenylate
RT with T7 primer
Amplify community RNA
RT with T7 primer
T7 in vitro amplification
restrict polyA tails
Convert to cDNA
~ 2-5 µg
Pyrosequence
~ 10s - 100s Mbp
~ 10s - 100s Mbp

Figure 3.

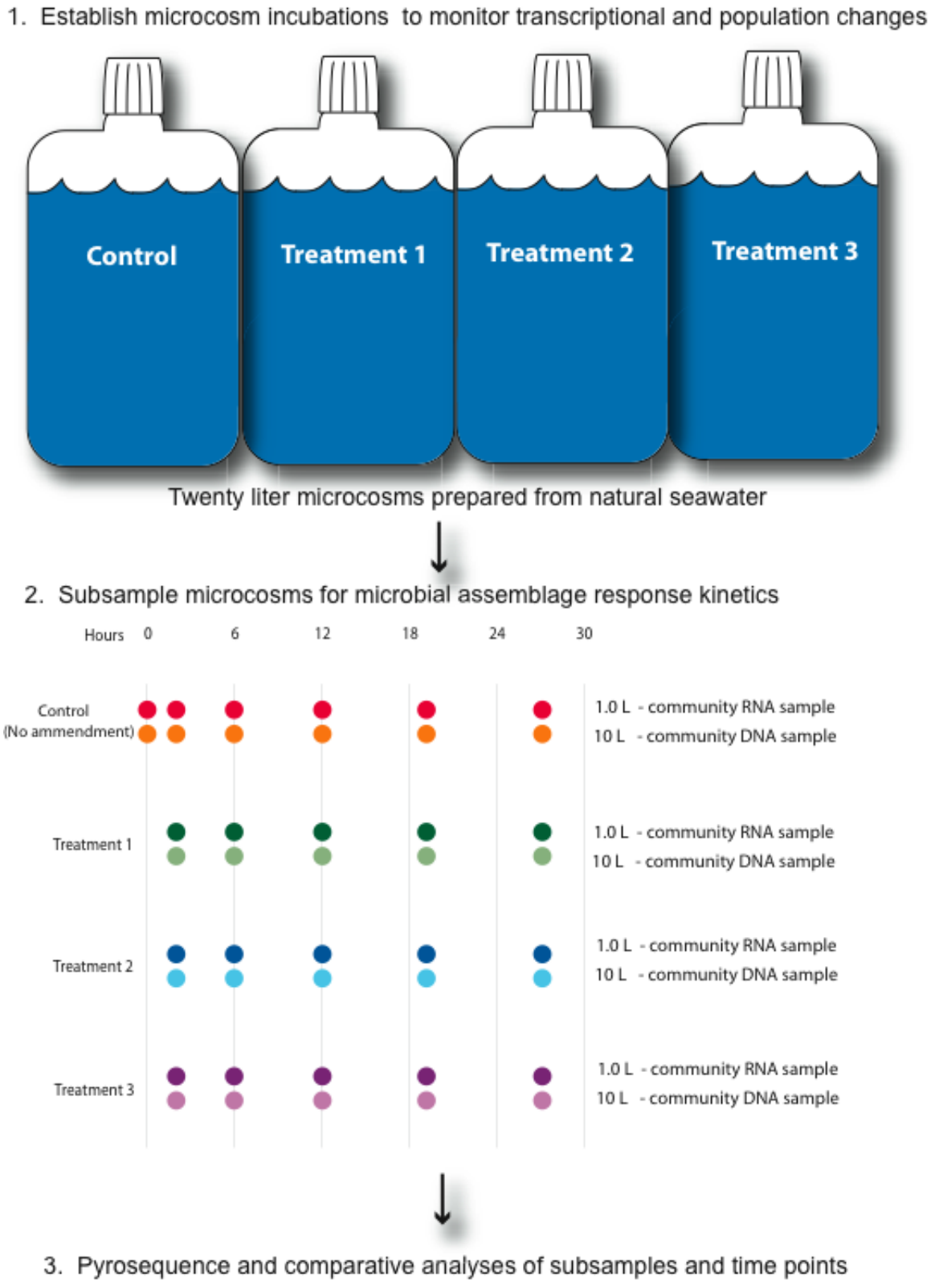
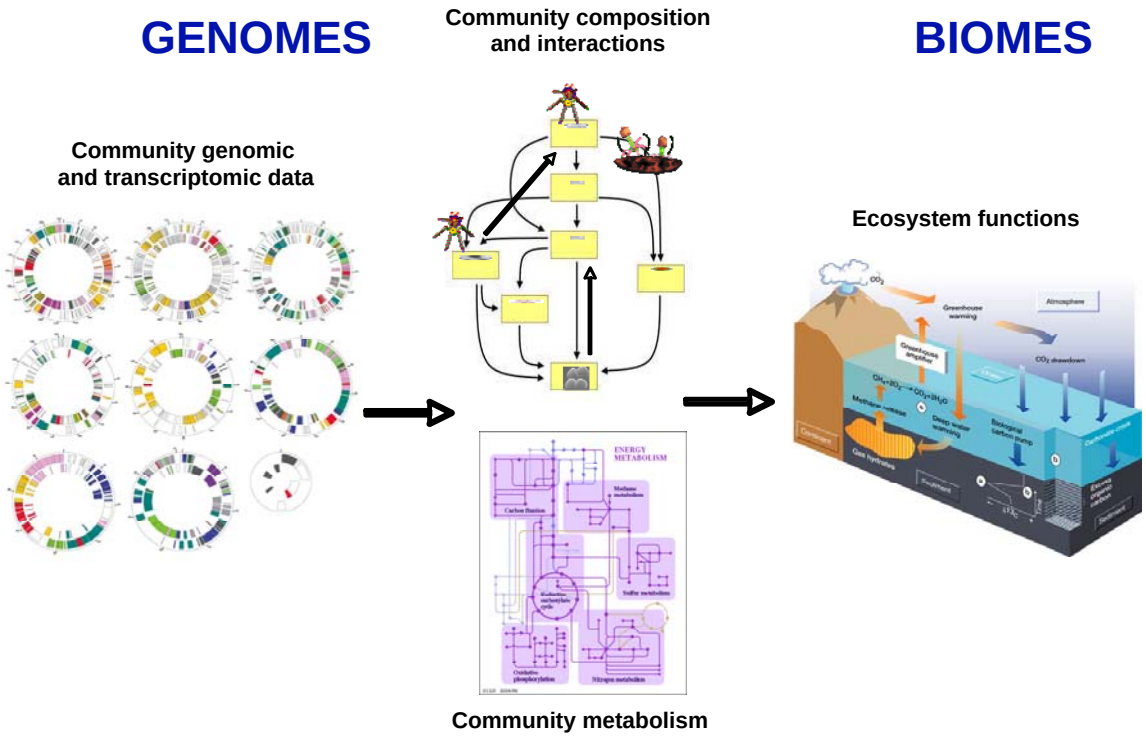

1. Establish microcosm incubations to monitor transcriptional and population changes
Control
Treatment 1
Treatment 2
Treatment 3
Twenty liter microcosms prepared from natural seawater
2. Subsample microcosms for microbial assemblage response kinetics
Hours 0 6 12 18 24 30
Control (No ammendment)
Treatment 1
Treatment 2
Treatment 3
1.0 L - community RNA sample
10 L - community DNA sample
3. Pyrosequence and comparative analyses of subsamples and time points

Figure 4.


GENOMES
Community composition and interactions
BIOMES
Community genomic and transcriptomic data
Ecosystem functions
Community metabolism