Massachusetts Institute of Technology
Woods Hole Oceanographic Institution

Joint Program in Oceanography/ Applied Ocean Science and Engineering

1930

DOCTORAL DISSERTATION

Development of a "Genome-Proxy" Microarray for Profiling Marine Microbial Communities, and its Application to a Time Series in Monterey Bay, California

by

Virginia Isabel Rich

September 2008

MIT/WHOI
2008-14

Development of a "Genome-Proxy" Microarray for Profiling Marine Microbial Communities, and its Application to a Time Series in Monterey Bay, California

by

Virginia Isabel Rich

Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

and

Woods Hole Oceanographic Institution
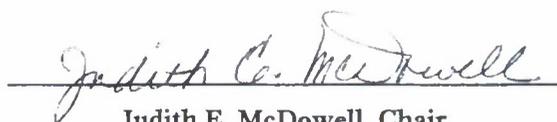Woods Hole, Massachusetts 02543

September 2008

DOCTORAL DISSERTATION

Approved for Distribution:

Judith E. McDowell

Judith E. McDowell, Chair

Department of Biology

Paola Malanotte-Rizzoli

Paola Malanotte-Rizzoli
MIT Director of Joint Program

James A. Yoder

James A. Yoder
WHOI Dean of Graduate Studies

Development of a "Genome-Proxy" Microarray for Profiling Marine Microbial
Communities, and its Application to a Time Series in Monterey Bay, California.

By

Virginia Rich

B.A, University of California at Berkeley, 1998

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
and the
WOODS HOLE OCEANOGRAPHIC INSTITUTION

September 2008

Signature of the Author

Joint Program in Biological Oceanography
Massachusetts Institute of Technology
Woods Hole Oceanographic Institution
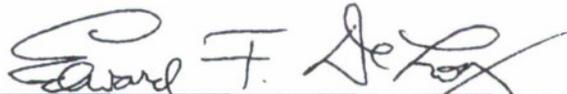August 1st, 2008

Certified by

Thesis Co-Supervisor
Edward F. DeLong
Massachusetts Institute of Technology

Certified by

Thesis Co-Supervisor
George Somero
Stanford University

Accepted by

Edward F. DeLong
Chair, Joint Committee for Biological Oceanography
Woods Hole Oceanographic Institution

Development of a "Genome-Proxy" Microarray for Profiling Marine Microbial Communities, and its Application to a Time Series in Monterey Bay, California.

by

Virginia Rich

Submitted to the Department of Biology and to the Woods Hole Oceanographic Institution in summer 2008, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biological Oceanography.

ABSTRACT

This thesis describes the development and application of a new tool for profiling marine microbial communities. Chapter 1 places the tool in the context of the range of methods used currently. Chapter 2 describes the development and validation of the "genome proxy" microarray, which targeted marine microbial genomes and genome fragments using sets of 70-mer oligonucleotide probes. In a natural community background, array signal was highly linearly correlated to target cell abundance (R^2 of 1.0), with a dynamic range from 10^2-10^6 cells/ml. Genotypes with ≥~80% average nucleotide identity to those targeted cross-hybridized to target probesets but produced distinct, diagnostic patterns of hybridization. Chapter 3 describes the development an expanded array, targeting 268 microbial genotypes, and its use in profiling 57 samples from Monterey Bay. Comparison of array and pyrosequence data for three samples showed a strong linear correlation between target abundance using the two methods (R^2=0.85-0.91). Array profiles clustered into shallow versus deep, and the majority of targets showed depth-specific distributions consistent with previous observations. Although no correlation was observed to oceanographic season, bloom signatures were evident. Array-based insights into population structure suggested the existence of ecotypes among uncultured clades. Chapter 4 summarizes the work and discusses future directions.

Thesis Co-supervisors: Edward DeLong and George Somero
Titles: Professor of Civil and Environmental Engineering, at MIT, and David & Lucile Packard Professor in Marine Sciences, at Stanford University.

Acknowledgments

Table of Contents

Chapter 1

Community profiling methods in microbial ecology.

Microorganisms drive global biogeochemical cycles, and are major numerical and biomass components of every habitat on the planet. Microbial ecology seeks to understand the relationships between microbial communities and their surrounding environment. However, these communities are complex, microscopic, and difficult to observe, and it remains an ongoing methodological and conceptual challenge to track individual taxa within communities and to profile communities as a whole. A range of methods are available for microbial community profiling, spanning differing degrees of phylogenetic resolution, specificity, sensitivity, ease of use, cost of adoption, and cost per sample. In this chapter, I present an overview of the major profiling methods (Figure 1), to place the new profiling tool developed and used in this thesis into context. This review does not address methods for functionally profiling communities or linking specific taxa to ecosystem functions, but rather focuses solely on methods used to assess community composition. These methods address questions about who is there, rather than what they are doing.

For the purposes of this review, it will be assumed that the microbial community sample has been prepared or obtained in a way that primarily captures the microbial component of the biota (since the larger cells can swamp or obscure the microbial signal, for many of the methods discussed here). For example, marine samples are often first passed through a 1.6-2.7 µm nominal pore-size pre-filter to remove larger eukaryotic cells (along with particle-attached or larger microbes) and then collected onto a 0.2um filter, allowing non-particle-associated viruses to pass through. Soil or sediment samples can be sieved to remove roots or invertebrates, while symbiont communities are dissected away from host tissues. Based on the habitat and conditions of the sample, the chemical details of subsequent treatments are then tailored to optimize efficiency, e.g. of DNA extraction.

Once a microbial community sample is obtained, there are four distinct

strategies for profiling the community composition of the sample:

1. Chemical fixation and whole-cell characterization: Cells remain essentially intact. Fixation strategies vary based on the cell types and background particles involved. Once a sample is chemically fixed, community members can be analyzed as whole cells based upon their light scattering and autofluorescent emission properties. Fixation can be followed by fluorescence *in situ* hybridization (FISH), flow cytometry (FCM), and/or fluorescence-activated cell sorting (FACS). These methods provide the most direct observations of organisms, and will be discussed further below, but can rarely discriminate beyond a handful of members or cell-types within complex microbial assemblages.

2. Lipid Extraction: A sample's lipids are extracted, subjected to specific chemical manipulation, and structurally profiled using high-pressure liquid chromatography (HPLC), mass spectroscopy (MS), and/or other methods. Lipid profiling has been a central tool in paleomicrobiology and in associated paleoclimate studies (e.g. reviewed in Summons *et al.*, 1996), for extinct communities in which there are no longer intact cells, and whose DNA may be absent or excessively degraded. In these cases, lipid composition is used to make inferences about phylotype presence. It has also been used to generate overall profiles for extant communities, (e.g. Vestal and White, 1989, Summons *et al.*, 1996, Schutter and Dick, 2000, Ritchie *et al.*, 2000, Hinojosa *et al.*, 2005, Moore-Kucera and Dick, 2008, Jiménez Esquilín *et al.*, 2008), and to examine particular groups with characteristic lipids (e.g. annamox bacteria, Schmid *et al.*, 2005, archaea, Biddle *et al.*, 2006, Ingalls *et al.*, 2006, Thiel *et al.*, 2007) However, because of ambiguities in the robust assignment of particular lipid types to specific clades (e.g. Rashby *et al.*, 2007), and because cellular lipid content and composition can alter, like ribosome content, depending on cellular physiology (Vestal and White, 1989, Villanueva *et al.*, 2004), it is not especially common as a *profiling* technique for extant communities, although lipid profiling

11

combined with isotope characterization is used to infer links between identity and particular metabolisms (e.g. Ingalls *et al.*, 2006). Therefore it will not be further discussed in this review.

3. Cultivation. Cultivation typically allows the characterization of <1% of microbial taxa, particularly in oligotrophic environments (e.g. Staley and Konopka, 1985), and cultivars are often rare phylotypes with physiologies quite different from those characteristic of the community (the standard culturing process often selects for eutrophic clades). However, some clades are amenable to cultivation and their presence and diversity in a sample can thus be partially profiled by cultivation (e.g. marine *Vibrio* spp., Thompson *et al.*, 2005). In addition, recent improvements in culturing, using conditions closer to those *in situ*, have led to a greater diversity of organisms becoming isolated or enriched (e.g. Connon and Giovannoni, 2002, Rappé *et al*, 2002, Cho and Giovannoni, 2004, Stingl *et al.*, 2008). However, cultivation alone has yet to effectively capture the complexity of a microbial community and even in the face of high-throughput optimizations remains a significant time commitment. Thus, while cultivation is critical for model-systems-based study of particular community members, it has limited use as a community profiling tool. Even for cultivation-amenable clades, cultured representatives rarely reflect *in situ* relative abundances, and thus further community characterization studies are required to accurately profile the original sample (Thompson *et al.*, 2005). For these reasons, cultivation will not be further described here (but see recent review by Giovannoni and Stingl, 2007).

4. Nucleic acid Extraction and characterization. The genetic code of a mixture of organisms, if it can be uniformly accessed[1], is a powerful means to

---

[1] Optimization of DNA extraction protocols to the particular community examined is critical. Particle association, sample chemistry (e.g. humic substances in soils, waxes in some plants, excessive mucus in marine invertebrate-associated communities), and cell surface properties (e.g., cell wall composition, spore coats) affect extraction efficiency may bias community representation.

survey a complex microbial community. Community DNA or rRNA extracted from a sample may require amplification before further analysis. Gene-specific amplification allows profiling of microbial diversity via a phylogenetically informative gene of interest, often the 16S rRNA gene or molecule. Amplification might be done quantitatively (e.g., quantitative PCR or RT-PCR) to assess relative abundances or non-quantitatively to generate sufficient copies of the gene of interest for further profiling analyses. In this latter case, amplified genes may be "fingerprinted" by methods ranging from terminal restriction fragment length polymorphism (tRFLP) to microarrays to sequencing. Alternatively, community profiling can occur without amplification by sampling the community genome directly through sequencing, either with or without cloning.

This range of commonly-used community profiling methods, from FISH through metagenomics (Figure 1), are described briefly in the context of profiling and compared below.

Fluorescence *in situ* Hybridization (FISH)

Fluorescence *in situ* hybridization (FISH) is widely used in microbial ecology (for a recent review, see Amann and Fuchs, 2008) and allows the enumeration and profiling of phylogenetic subsections of a community using fluorescence microcopy. Briefly, microbial cells from a natural community sample are chemically fixed and permeabilized, then hybridized to fluorescently-labeled oligonucleotide probes complementary to e.g. the 16S rRNA sequence of a clade of interest. Fluorescently labeled cells are then imaged and quantified using fluorescence microcopy (see e.g. Amann and Fuchs, 2008). Unlike the majority of community analysis techniques, FISH examines whole cells and thus provides additional community information about cellular morphology and spatial structuring.

A functional physiological component can be added to FISH by combining it with other methods. MAR-FISH (aka Micro-FISH, STAR-FISH) combines FISH

13

with microautoradiography by incubating a sample with radiolabeled substrate diagnostic for a metabolic pathway or biogeochemical process (Lee *et al.*, 1999). Further refinements include FISH-SIMS, which uses secondary-ion mass spectroscopy to detect both stable and radioactive isotopes, to a spatial resolution of 10-15µm (Orphan *et al.*, 2001). The resolution of such methods continues to improve, with FISH-NanoSIMS (50nm) allowing a new level of precision in hypothesis-testing (Lechene *et al.*, 2007).

In spite of the considerable strengths of FISH-related profiling methods, there are limitations: (1) variable fixation / membrane permeabilization across cell types, (2) detection sensitivity in natural samples, (3) variable probe specificity and accessibility, (4) low sample throughput and (5) high background and sample autofluorescence. (For a good methodological review of FISH, see Bottari *et al.*, 2006).

Fixation / permeabilization: Cell wall composition varies among microbes, and so fixation and permeabilization without lysis can be difficult to achieve across the range of cell types typically present in a community, and methods must be tailored to each investigation. Fixation and permeabilization is achieved with paraformaldehyde and ethanol, aided by enzymes (lysozymes and proteases), solvents, acids, and detergents (Amann and Fuchs, 2008). Not only can the efficacy of these steps vary greatly among cells, and thus affect target accessibility, but the composition of co-purified elements of the community habitat can also affect the efficiency of probe binding as well as sensitivity (Amann and Fuchs, 2008).

Sensitivity: Several factors affect FISH sensitivity. First, FISH probes target ribosomal RNAs, which are an active component of the ribosome, the cell's protein assembly machinery. The number of ribosomes in any given cell can vary significantly, from tens to tens of thousands of copies per cell. Small cells tend to have fewer ribosomes per cell (Kemp *et al.*, 1993, Morris *et al.*, 2002), and poor

14

growth conditions can result in decreased ribosomal content (Kemp *et al.*, 1998). The ability to detect specific organisms is highly dependent on the ribosome copy number, and thus for some environmental applications targeting small and/or nutrient-deprived cells, sensitivity may be a challenge.

In addition, the probe and fluorophore selection can affect sensitivity. Not all dyes are equally bright (e.g., Alexa Fluor 555 outperforms Cy3), and probe alterations can also help amplify signal. Multiple probes targeting the same taxon at different regions of its rRNA (e.g. for SAR11 - Morris *et al.*, 2002), or multiply-labeling single probes (Pernthaler *et al.*, 2002), can both be effective strategies. In the latter case, the accommodation of additional dye molecules requires more expensive polynucleotide probes instead of oligonucleotides, which consequently constrains target resolution to higher-level taxonomic levels (Amann and Fuchs, 2008). An additional option for increasing probe sensitivity is CARD (CAtalyzed Reporter Deposition) -FISH, which links the horseradish peroxidase (HP) enzyme to the oligo probe instead of a fluorophore. This enzyme then converts fluorescently-labeled tyramide molecules in the reaction mixture into their fluorescently-active state, in which they also bind locally to the cell, allowing signal amplification up to 20 times the intensity seen with monolabel FISH probes (Schönhuber *et al.*, 1997, Pernthaler *et al.*, 2002). The enzyme molecule is 40-80 times larger than a fluorophore, however, and thus good entry of the HP-oligo conjugate into the cell can be challenging and requires good permeabilization (e.g. initial embedding of the cells in agarose followed by stronger chemical treatment than in standard permeabilization protocols). The bulkiness of the CARD-FISH probe is particularly problematic when evaluating densely-associated cells (Schönhuber *et al.*, 1997), and in these cases becomes *less* sensitive than classical FISH. However, in planktonic cells, the increased sensitivity of CARD-FISH has allowed the simultaneous labeling of both ribosome and mRNA from particular genes of interest (Pernthaler and Amann, 2004). In this manner, one can simultaneously assay for "who is present" and

"are they active", within a complex microbial community. As sensitivity increases, FISH has moved from detecting high-copy ribosomal RNAs to mRNAs, and has achieved hybridization and visualization to single-copy genes in the genome (RING-FISH, Zwirglmaier et al., 2004; RCA-assisted FISH, Smolina et al., 2007, Marayuma et al., 2005).

Lastly, rather than increasing the absolute signal, sensitivity can be improved by reducing the noise and background fluorescence. Sample treatment plays an important role in this, since humics and other contaminants can autofluoresce and/or interfere with probe binding. In addition, some probe structural design modifications can decrease background fluorescence caused by non-specific binding (e.g. "molecular beacon" probes, which improve signal-to-noise ratios and thus sensitivity, e.g. Lenaerts et al., 2007).

Specificity: As with other DNA-based hybridization methods, specificity is an important consideration in FISH. Recent general modifications to probe structure (e.g. peptide nucleic acid probes, described in Bottari et al., 2006, Amann and Fuchs, 2008) offer increased binding affinity and/or stability, which can increase specificity (and sensitivity). As with all nucleic-acid probes, FISH probe specificity relies on the size and quality of the reference database used to design them. Many commonly-used domain and group-specific probes were designed when the 16S databases were significantly smaller, and thus have an expanding false-negative and -positive rate (for an updated analysis of common FISH probes specificity see Amann and Fuchs, 2008). Ideally, probes should be re-evaluated before use using current databases and tools (e.g. probeBase, the online database and toolkit for designing 16S-rRNA probes and primers, Loy et al., 2007). Probe design and optimization are not restricted to cultivated clades, since not only do many clades have sufficient database representation due to 16S environmental surveys, but such environmental sequences can be cloned and then used within their heterologous host for optimization of hybridization conditions (clone-FISH, Schramm et al., 2002). Finally, in order to add

confidence to the interpretation of FISH specificity, multiple hierarchic probes can be used simultaneously as internal cross-validation (Amann and Fuchs, 2008).

From a practical stand-point, as a community profiling method FISH is inexpensive once set up (i.e. requires a good fluorescence microscope but not exorbitant reagents), and is ideal for research questions that target a small number of monophyletic groups with good cell permeability, particularly where spatial structure and cell size observations are relevant. Sample preparation requires a moderate to advanced level of technical expertise in order to obtain good results (fixation, permeabilization and hybridization protocols all need to be tuned for different communities), and takes hours to days depending on the sample. Once methods are optimized, FISH's limit of detection can reach ~0.1% of the community (Amann and Fuchs, 2008, Woebken *et al.*, 2007). Although sample scanning and analysis are becoming more automated (Daims *et al.*, 2006; Alonso and Pernthaler, 2005; Cottrell and Kirchman, 2003), FISH is not a high-throughput method. In summary, FISH is an important community profiling tool when examining division-level community structure (using previously developed probes) or focusing in on a limited number of specific microbial groups (the number of different species simultaneously resolvable is currently around seven, Amann and Fuchs, 2008). However, given the complexity of natural microbial communities and time required for probe development, FISH is not meant to comprehensively profile more than a small fraction of the microbial community at a refined taxonomic scale. Profiling at higher taxonomic levels – e.g. the level of alpha-, beta-, or gamma- proteobacteria – can miss significant differences at lower taxonomic resolution. Also, the application of FISH cannot easily be standardized to a variety of samples, due to differences in fixation and hybridization.

Flow cytometry and fluorescently-activated cell-sorting

17

Flow cytometry (FCM) and fluorescence-activated cell sorting (FACS) use optical properties of a sampled population to enumerate and/or separate different optical grouping of cells. Basic FCM relies solely upon the inherent properties of the cells themselves. For example, FCM "signatures" result from the combined effects of cell size, shape, and internal structure on light scatter, and of the characteristic autofluorescence of naturally-occurring cell pigments (e.g., chlorophyll or phycoerythrin). While FCM simply counts and records cells based on these properties, FACS uses the same properties to sort different populations of cells.

A major focus of flow cytometric studies has been photosynthetic members of communities (for a recent discussion of phytoplankton flow cytometry see Dubelaar et al., 2007) because of their ease of detection due to pigment autofluorescence. Many photosynthetic clades can be distinguished based on differential pigment composition and cell size. A notable example is the FCM-signature-based discrimination of the ocean cyanobacteria Prochlorococcus from its co-occurring sister group, Synechococcus, by photosynthetic pigments (divinyl chlorophyll a vs chlorophyll a and phycoerythrin) and cell size, and from larger co-occurring picoeukaryotic phytoplankton (Chisholm et al., 1988, Waterbury et al. 1984). Even without the additional probe- and dye-based discriminatory abilities of flow cytometry, its suitability for profiling photosynthetic microbes ensures its value as a tool particularly for aquatic microbial ecologists, and it has been used widely and successfully in a number of studies (e.g. Seymour et al., 2005, Johnson and Zinser et al., 2006, Mary et al., 2006, Thyssen et al., 2008).

FCM profiling of non-photosynthetic groups is more challenging because they possess less native FCM-signature information. Without fluorescence, small cells are difficult to image based solely on light scatter unless significant instrument modifications are made for small-particle detection. As a result, methods developed for epifluorescent microscopy have been transferred to FCM applications, for example DNA stains (e.g., Hoescht, DAPI, SYBR) may be used

to intercalate with cellular DNA and cause cells to fluoresce (when excited at the correct wavelength). Such bulk DNA staining can allow total microbial counts (e.g. in Kuypers *et al.*, 2005), the delineation of gross microbial groups (e.g., high and low-DNA cells, Gasol *et al.*, 1999), and the examination of cell physiology (e.g. LIVE/DEAD stains, Berney *et al.*, 2007), it does not allow much resolution for community profiling.

Rather than bulk staining, clade-specific FISH tagging may be combined with FCM to enumerate particular phylogenetic groups from a community. FISH-FACS has begun to allow the enrichment of target populations from complex communities for further analysis. Although absolute separation of probe-targeted cell types has not yet been obtained, enrichment for targeted cells has been successful (e.g. type I and II methanotrophs enriched from 4.7% to 50% and 1.2% to 47.5%, respectively, Kalyuzhnaya *et al.*, 2006; but only ~2-fold for targeted *CFB* and no enrichment for targeted β-proteobacteria in Sekar *et al.*, 2004). Combining FCM with FISH involves many of the limitations and challenges associated with FISH (especially the negative effects of contaminants not removed during sample preparation), while removing the spatial structural observational power of FISH, but it does allow the enumeration of the whole cells of targeted groups in a high-throughput manner.

In addition, advances in FCM, such as equipment miniaturization, are bringing down costs and enabling novel field deployments (Gruden *et al.*, 2004, Yang *et al.*, 2006). Perhaps pinnacle among these field efforts are the "FlowCytoBot" (Olson and Sosik, 2007; Sosik and Olson, 2007) and "Cytosub" (Thyssen *et al.*, 2008) which use robotics and autonomous sampling devices to enable *in situ* real-time FCM. The FlowCytoBot can be deployed for more than a month, and performs both *in situ* flow cytometry and automated image-based taxonomic identification of larger cells.

Overall, as a profiling method, FCM and FACS are relatively quick,

reproducible, and inexpensive (once a flow cytometer/sorter is available; although most core facilities have them available because of their use in medical research). For community profiling, however, the approach has not been well developed for standard and comprehensive surveys. Depending on the population being targeted and the identification method, FCM can be a relatively straightforward process or a technologically very sophisticated one that requires substantial expertise and validation.

Nucleic acid-based profiling

Since Pace *et al.* (1985) pioneered the use of ribosomal rRNA gene as a genetic marker for studying the diversity of microbes in natural systems, this new window has provided unprecedented views of the "uncultured majority" and vastly expanded our understanding of microbial diversity. The remainder of this chapter is devoted to reviewing the major culture-independent molecular methods used to profile microbial communities. These methods fall into two fundamental classes: those that rely upon amplifying a single target gene from community DNA or rRNA, and use this gene to "fingerprint" the community through any of a number of methods, and those that directly examine the community DNA without gene-specific amplification (i.e. metagenomics).

Single-gene surveys

One way to profile a community is by surveying a phylogenetically-informative marker gene within that community. Commonly-used phylogenetic markers include genes involved in translation (16S rRNA, 23S rRNA, the internal transcribed spacer (ITS) region between the two, and ribosomal proteins), transcription, (e.g. transcription factors, and RNA polymerases component genes such as rpoB), and DNA replication and repair (DNA pol, recA), and other core cellular functions (e.g. the chaperone gene dnaK), as well as some functional genes that are considered phylogenetically robust (i.e., show little or no evidence of horizontal gene transfer), e.g. the *pmoA* gene of methanotrophs (McDonald *et*

al., 2008).

<u>Caveats of DNA amplification</u>

Single-gene investigations of community profiles use the polymerase chain reaction (PCR) to amplify the target gene from environmental DNA. However, this amplification has several caveats. First, PCR reactions can create both errors (heteroduplex and chimeric products, as well as polymerase error) and biases (skewing of the relative proportions of sequence types). Both significantly effect downstream diversity estimates (Acinas *et al.*, 2005, Thompson *et al.*, 2002, Polz and Cavanaugh 1998). Second, given the complexity of unknown microbial communities in the wild, primer specificity (as described above) may not be uncertain, i.e. primer sets may not amplify as comprehensively as they're assumed to. For example, "universal"-primer-based surveys missed an entire high-level taxonomic group, the kingdom Nanoarchaeota (Huber *et al.*, 2002). Similarly, other primer sets continue to miss lineages (e.g. among the archaea, Teske and Sørensen, 2007, and among the planctomycetes, Derakshani *et al.*, 2001, Köhler *et al.*, 2008).

Several methodological adjustments have been suggested to ameliorate the problems of specific-primer-directed PCR. First, the number of amplification cycles should be kept to a minimum to decrease bias (Suzuki and Giovannoni, 1996) and to minimize chimera formation and polymerase errors (Acinas *et al.*, 2005). Second, pooling replicate PCR amplification reactions helps compensate for early-cycle drift that leads to bias (Acinas *et al.*, 2005, Polz and Cavanaugh, 1998). Third, reactions should be ramped as quickly as possible from denaturing to annealing temperatures (Acinas *et al.*, 2005). Fourth, a "reconditioning" approach minimizes heteroduplex and chimera formation, by employing a second low-cycle amplification using a dilution of the first amplification with excess primer (Thompson *et al.*, 2002). Finally, as discussed above for FISH probes, it is important that primers (particularly at the domain level) be continuously

reassessed in light of the ever-expanding size of environmental sequence databases, and several combinations of primers may be used if the goal is to maximize the diversity of sequences recovered

Once amplified, phylogenetic marker gene amplicons can be used for community fingerprinting methods or sequenced.

Quantification during amplification

Quantitative PCR (qPCR) is widely used in microbial ecology as a means of directly enumerating the abundance of specific clades (via conserved sequences common to all clade members) in environmental samples. To date, many of these studies have examined specific food contaminants and pathogens, but qPCR has played a role in microbial ecology as well (reviewed in Zhang and Fang, 2006). Three studies of particular note apply such methods to ocean systems, and provide examples of the range of target genes used. First, Suzuki *et al.* (2000) was one of the first marine qPCR studies of the marine picoplankton, and in surveys of *Prochlorococcus*, *Synechococcus*, and *Archaea* in Monterey Bay showed good concordance between 16S-targeted qPCR and other methods. Second, Johnson and Zinser *et al.* (2006) targeted the ITS region by qPCR to describe *Prochlorococcus* ecotype abundance across a basin-wide transect in the Atlantic Ocean. Remarkably, the qPCR results accounted for most of the FCM-detected *Prochlorococcus* populations in these ocean samples. A third study used qPCR of a functional marker gene (*amoA*) to quantify the relative abundance of putative archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre (Mincer *et al.* 2007). qPCR is highly reproducible and sensitive, relatively inexpensive, high throughput, and is a valuable tool for profiling particular community members. Significant primer optimization is required, and robust primer design relies on comprehensive environmental sequence information to ensure that the breadth of desired native diversity is targeted. In addition, single to several taxa can be profiled using

qPCR, but because of design and optimization issues and finite amounts of sample DNA this method is not practical for profiling many clades or taxa simultaneously.

Fingerprinting of amplified genes

Most profiling studies examine amplified phylogenetic marker gene using one or more fingerprinting method. These methods may be used as an end in themselves or as the first or complementary step in an investigation. Fingerprinting methods separate amplicons based on sequence differences by assaying some consequence of those sequence differences - e.g., restriction sites, denaturation, or amplicon length. In general, these methods are useful for detecting structural changes in community composition between samples. Depending on the method and primer sets, these techniques may be executed along a scale from coarse to fine phylogenetic resolution. While inexpensive and relatively quick, these methods require substantial development and validation to connect particular fingerprints with particular clades. Furthermore, the dynamic range is often limited; beyond a narrow range, measurement of abundance differences between samples may be qualitative rather than quantitative. Many of the methods described below are applied to a variety of genes, although some are specific to the 16S rRNA gene or operon, and for profiling purposes the 16S gene remains the most common target.

Fingerprinting by differences in denaturation and annealing

DGGE & TGGE: Denaturing or temperature gradient gel electrophoresis. DGGE & TGGE (reviews in Muyzer and Smalla, 1998, and Nocker et al., 2007, respectively) separate small (generally less than 800 base pairs (bp)) PCR products by dissociation differences caused by sequence heterogeneity. Amplicons are electrophoretically separated on an acrylamide gel that contains a parallel denaturing gradient, generated either chemically (DGGE e.g. by urea-formaldehyde) or with temperature (TGGE). As amplicons move through the gel

23

towards the cathode, they enter increasingly denaturing conditions, causing bubbles of denatured DNA to form in the double-stranded amplicons. This greatly increases the molecules' surface area, which significantly retards their movement through the gel. Sequence heterogeneity among the amplicons results in difference points of denaturation, and thus different patterns of migration. Some DGGE & TGGE protocols incorporate a GC-clamp (usually ~ 40bp) on one primer in the PCR step, which will then act to hold a mostly-denatured amplicon together at one end as it moves through the gel (since G-C bonds are more stable than A-T ones, so a long G-C string will be slow to denature) and thus prevent or delay complete dissociation, which would complicate banding and interpretation. However, GC clamps can cause decreased PCR efficiency and increased likelihood of artifacts, and are not necessary if gentler denaturing gradients are used (Nocker *et al.*, 2007).

Optimal denaturing conditions are first identified by running amplicons in multiple wells through a denaturing gradient perpendicular to electrophoresis, and then the optimized range can be used parallel to electrophoresis to generate the DGGE banding. Bands may be visualized by staining with DNA dyes, or primers can be fluorescently label to improve visualization sensitivity of denatured amplicons (e.g. 10-fold, Moeseneder *et al.*, 1999). As with all fingerprinting methods, the phylogenetic specificity of DGGE and TGGE depends on the specificity of the primers used for amplification, although the methods are inherently limited in resolution due to their gel separation and visualization steps. They are generally used to identify bulk shifts in community composition, although they can be used effectively to track specific changes in simpler communities (Nocker *et al.*, 2007), and there is continual refinement of group-specific DGGE primer sets (e.g. updated marine bacterial clade primers, Mühling *et al.*, 2008). Under typical conditions, the limit of detection of DGGE and TGGE is target groups that represent at least ~1-2% of the community (Nocker *et al.*, 2007). The method is inherently quite sensitive to sequence variations among

amplicons, and bands can be excised and sequenced to link bands to particular genotypes. However, as with many fingerprinting methods, one band or fingerprint type cannot be assumed to derive from a single phylogenetically coherent sequence clade. In addition, tailoring denaturing conditions to the particular amplicons under study can be time-consuming, and it can be difficult to compare data robustly between runs and labs because of gel variability; though see "Additional Considerations" below.

DHPLC: Denaturing high-performance liquid chromatography. PCR products are separated using temperature and chemical denaturation. Products loaded onto an HPLC-cartridge in a solution of acetonitrile and triethylammonium acetate, TEAA. The TEAA converts to TEA+, an amphiphilic molecule, which interacts with the DNA at its charged end, and the HPLC cartridge material at its nonpolar end, thereby tethering the amplicons in place. The amplicons are sequentially eluted from the cartridge by increasing the temperature and acetonitrile concentration. Elutants are quantified by a UV detector that measures absorbance at 260nm, or amplicons can be fluorescently labeled via their PCR primers. A fraction collector can be joined to the HPLC to collect eluted fragments for further characterization. DHPLC does best at separating smaller fragments below about 500bp, but can work on larger e.g. 1500bp molecules at decreased sensitivity. Separation parameters need to be carefully tailored for the targets of interest, and this can be extremely time-consuming.

CDCE: Constant denaturant capillary electrophoresis. CDCE was developed for, and remains primarily used for, genetic screening of mutations (Khrapko *et al.*, 1994), but has been applied to community profiling in microbial ecology (Thompson *et al.*, 2004). PCR products are loaded onto a polyacrylamide capillary with constantly denaturing conditions (chemical- or temperature-based). Amplicons denature dynamically to differing degrees and rates based on their sequences, causing them to travel at different speeds, and elute at different times from the capillary. By tagging PCR primers with a

fluorophore, product elution can be measured by laser detection. CDCE has high sensitivity to single base pair differences, and there is a quantitative relationship between fluorescence intensity of eluted fragments and their relative abundance. Eluted amplicons can be collected for further analysis.

SSCP: Single strand conformation polymorphism. PCR amplicons are uniformly denatured first, and then are run on an acrylamide gel or on a capillary sequencer as single-stranded molecules. The ssDNA molecules folds back on themselves creating internal secondary structure. This secondary structure will result in differential migration through a matrix, and allow for separation. A challenge of SSCP is the difficulty of keeping ssDNA from re-annealing to its complement during gel loading and running. However, one of the two primers can be tagged with a 5' phosphate group, which allows selective targeting of one of the two strands by lambda exonuclease (Nocker *et al.*, 2007; Schwieger and Tebbe, 1998). Interpretation of results can be complicated by the fact that single ssDNA sequences can fold in multiple ways, which if they have similar energetic favorability can result in a single sequence type being represented by several bands.

Fingerprinting by amplicon restriction site heterogeneity

ARDRA: Amplified ribosomal DNA restriction analysis. The 16S rRNA genes are PCR-amplified from bulk community DNA, and cloned. Clones are then restriction digested and the fragments are separated on a gel, visualized by gel staining with a DNA dye (e.g. ethidium bromide), and the banding pattern is interpreted as a low-resolution proxy for phylogeny. Multiple restriction enzymes are required in order to differentiate among lineages (Moyer *et al.*, 1996), and enzyme choice has a significant effect on the scale of resolution (Zeng *et al.*, 2007). Generally, clones are also sequenced to validate interpretation of banding patterns. Using higher taxonomic-level 16S rRNA gene primers for amplification leads to in low resolution sample comparisons, while more taxonomically-specific

primers can be used to investigate particular clades of interest, as ARDRA is able to discriminate at the species level (e.g. among bioluminescent marine bacteria, Kita-Tsukamoto *et al.*, 2006). ARDRA is inexpensive and technically straightforward, however has a relatively low sensitivity and variable resolution, and as with all gel-image-based methods, comparing profiles confidently over time or between labs can be challenging. It is used for overall community profiling (e.g. Polz *et al.*, 1999), examining specific clades of interest within communities (e.g. Kita-Tsukamoto *et al.*, 2006), or directing investigations of 16S clone libraries in order to optimize the cost-benefit between clone sequencing and adequate description of community diversity (e.g. Sun *et al.*, 2008), or for indirect community profiling by distinguishing among isolates (e.g. Michel *et al.*, 2007).

TRFLP: Terminal restriction fragment length polymorphism. TRFLP (discussed in Hartmann and Widmer, 2008) begins with gene-specific PCR amplification using a fluorophore-conjugated primer. Amplicons are restriction digested, and fragments are separated by size using capillary electrophoresis, and visualized in Genescan mode. Restriction site distribution can be a phylogenetically informative character, and methods are tailored by initial *in silico* experiments using existing databases (e.g. testing the specificity of a particular primer set /restriction enzyme combination against all of RDP, Marsh *et al.*, 2000, Kent *et al.*, 2003). As with ARDRA, the use of multiple enzymes can help refine resolution of interpretation, and data analysis must be done carefully (Osborne *et al.*, 2006). Caveats include incomplete restriction digestion, and a slowing of fragment migration due to unwieldy dye molecules (although not all dye molecules have a significant effect on mobility) (Nocker *et al.*, 2007). TRFLP is typically more sensitive than DGGE (e.g. five times as sensitive, Tiedje *et al.*, 1999) due to fluorophore-based visualization rather than gel staining, although it may be less sensitive that LH-PCR (discussed below) because of partial restriction digestion (e.g. in an ITS analysis using tRFLP and LH-PCR, LH-PCR was both more sensitive and higher resolution, producing more distinct bands

27

than tRFLP, Mills *et al.*, 2003). TRFLP analyses of a number of genes are common in microbial ecology studies (e.g. Bertilsson *et al.*, 2007, Goffredi *et al.*, 2008, Morris *et al.*, 2005, and Horz *et al.*, 2005, Appendix 3) because of the method's relative ease and high reproducibility.

Fingerprinting by amplicon length

LH-PCR: Length-heterogeneity-PCR. LH-PCR (reviewed in Mills *et al.*, 2007) examines a subsection of the 16S rRNA gene, spanning some subset of its variable regions, and amplicons (uncloned) are then distinguished based on length heterogeneity. Since relatively small amplicons are created (generally several hundred base pairs), small differences in length can be resolved. One primer used for amplification is fluorescently linked, to allow relatively precise size assessment of amplicons using capillary sequencers in Genescan mode. The area under the Genescan peak is used as a metric for the abundance of a particular fragment length class. LH-PCR has been used in a number of community profiling studies (e.g. Suzuki *et al.*, 1998, Ritchie *et al.*, 2000, Mills *et al.*, 2003, Brusetti *et al.*, 2006, Sekar *et al.*, 2006). As with ARDRA, primers can be taxonomically tuned to allow the fingerprinting method to focus on particular clades (e.g. LH-PCR targeted to bovine gut commensals, Bernhard and Field 2000). LH-PCR can also be used on intergenic regions of other operons, and has been for the *amoC-amoA* operon (Norton *et al.*, 2002).

ARISA: Automated ribosomal intergenic spacer analysis. Rather than fingerprinting the 16S rRNA gene, ARISA uses the amplified, uncloned intergenic transcribed spacer (ITS) region between the 16S and 23S rRNA genes in the *rrn* operon. The ITS can be amplified from across broad taxonomic range by using conserved primers within each of the highly conserved flanking gene, though primers can be tailored to specifically target clades of interest. The ITS evolves at a faster rate than the rRNA genes, providing a finer scale of phylogenetic resolution and allowing discrimination up to ~98% rRNA sequence identity

28

(Brown *et al.*, 2005). Initially, ARISA amplicons were separated on polyacrylamide gels and visualized with silver staining, but later incorporated fluorescently-labeled primer-based visualization using capillary system (Fisher and Triplett, 1999). The amplicons generated using ARISA are generally longer (typically over a thousand bases) so the resolution of length may not be as precise as in LH-PCR (depending on the primers used in LH-PCR), although a larger variety of fragments is produced, potentially allowing finer-scale resolution. Again, Genescan peak area is used as a proxy for the abundance of the fragment source clade(s). ARISA has been used for a number of high phylogenetic-resolution community profiling studies, some with large numbers of samples, draw sophisticated correlations between detailed community structure and environmental parameters (e.g. from oceanic drifter samples, Hewson *et al.*, 2006a, and at several Microbial Observatories; Fuhrman *et al.*, 2006, Hewson *et al.*, 2006b, Kent *et al.*, 2007).

ITS-LH-PCR is a variant of ARISA, achieving yet a finer level of phylogenetic resolution. It involves a second, parallel step of restricting ITS amplicons by targeting the tRNA-alanine genes that commonly occur within ITS regions (Suzuki *et al*, 2004), and then estimating the size of the restricted fragment. This form of ITS-LH-PCR provides a higher degree of phylogenetic specificity among those clades with a tRNA-alanine in their ITS, as compared to standard ITS-LH-PCR/ARISA. ITS-LH-PCR has been used as a library-screening method (Suzuki *et al*, 2004) to assess community diversity captured in a large-insert clone library.

There are caveats involved in the interpretation of ITS patterns observed with ARISA. First, not all species have linked 16S and 23S rRNA genes (e.g. in Planctomycetes, Liesack and Stackebrandt, 1989, Menke *et al.*, 1991). Second, many organisms have *multiple* linked 16 and 23S rRNA genes in their genomes, the copies of which may or may not be identical. A recent study of 155 fully-sequenced bacterial genomes showed the average number of rRNA (*rrn*)

operons per genome to be 4.8 (range 2-15), with 2.4 unique ITS length variants per genome and 2.8 unique ITS sequence variants (Stewart and Cavanaugh, 2007). Thus, although gene conversion does act to homogenize multiple *rrn* operons in a genome, substantial heterogeneity persists (Stewart and Cavanaugh, 2007). Third, there appears to be preferential PCR-amplification of shorter ITS templates (Fisher & Triplett, 1999) causing the relative abundance of variants to be skewed (beyond other potential PCR biases; see "Amplification" section).

ARISA is relatively easy, and like LH-PCR its profiles are digitally extracted, and run alongside markers, and so can be compared easily between runs and labs. However, it is "destructive" fingerprinting and amplicons cannot be sequenced after they are measured.

Fingerprinting - Additional considerations: Lineage differences in target gene copy number, and intragenomic diversity of multicopy genes, can have potentially confounding effects on fingerprint-based single-gene community profiling. For example, with a range of 2-15 *rrn* operons per genome among 155 bacterial genomes (Stewart and Cavanaugh, 2007), 16S- or 23S-based diversity could overestimate organismal diversity quite considerably. In addition, for many of above fingerprinting methods, it can be difficult to compare between labs and environments. And for all of the above methods, it is not always straightforward to convert fingerprint data into ecologically-meaningful metrics. A recent unified set of metrics has been proposed (Marzorati *et al.*, 2008), developed for DGGE but also applicable to other fingerprint data, and summarized here as being of particular interest. This conversion of fingerprints to environmental metrics has three steps: 1. Generation of a range-weighted richness, *Rr*, describing the relative complexity of the fingerprint given the degree of separation applied; in the case of DGGE this would involved the number of bands (N) and the denaturing gradient the span of given bands (D_g, e.g. 35% to 40% urea and formamide), such that $Rr = (N^2 \times D_g)$. 2. Community dynamics, *Dy*, where profiles

are available for the same community over time (which they usually are, since they are used to identify shifts in community composition). This uses a moving window analysis to look at the Pearson correlation between subsequent time points, and calculates the consecutive percent change of the community at consecutive times. This is then converted into a Dy value for that community averaging the moving window percent change values over time. Low Dy values represent communities that do not change quickly or substantially, and high Dy values indicate highly dynamic communities. 3. Functional organization, Fo, is a proxy for evenness of the fingerprints observed. It is plotted as the number of unique fingerprint elements observed along the x axis, and their contributions (by e.g. intensity) along the y axis, such that a 1:1 line would represent perfect evenness and skewing from that line indicates the relative unevenness of a community. The authors take this final Fo metric a step further to tie evenness to community resiliency, but regardless of the merits of that connection, the potential utility of these basic quantification metrics stands. A given environment can then be plotted as a point in three dimensions, and compared to other environments.

Fingerprinting summary: The important unifying caveats of these methods are that single observed fingerprints may not be phylogenetically coherent, and that different fingerprints using a given method probably do not reflect the same level of phylogenetic resolution. In addition, since most techniques are not universal in their coverage, they may miss significant subsets of the community. Also, many of them have limited dynamic range, which may preclude accurate relative abundances comparisons of different groups among samples. Fingerprinting methods are ubiquitously used in microbial ecology research as they provide quick, relatively inexpensive profiling information. However, care must be taken in their interpretation and validation. Additional overall conclusions about fingerprinting methods are that visualization by fluorophores leads to higher sensitivity than by gel staining, and methods that capture data *in silico*

(e.g. using Genescan mode of capillary sequencers) offer higher resolution and reproducibility.

Sequencing of Amplified Genes

The highest-resolution way to distinguish sequence variants among PCR amplicons is to sequence them. Because these community PCR amplicons are heterogeneous, they must be first cloned in order to separate each variant and then individually sequenced. Although it has been speculated that small-insert cloning introduces bias, (e.g., small sequences clone with higher efficiency, end sequences can effect modification reactions required in cloning, and expressed inserts may be toxic to the host), one recent study showed good congruence between chemically dissimilar cloning methods and concluded that in general, little bias is introduced using the common TOPO TA cloning method (in a ITS and rRNA-gene survey, Taylor *et al.*, 2007). Sequencing 16S rRNA gene clone libraries remains the gold standard for profiling communities, and there is a wealth of information about interpreting microbial diversity with this method (e.g. Schloss, 2008, Fierer *et al.*, 2007, Janssen, 2006, Bohannan and Hughes, 2003, Hughes *et al.*, 2001), and so it will not be discussed in depth here.

New high-throughput sequencing technologies, e.g., "454" or "pyrosequencing" (Margulies *et al.*, 2005), allow direct sequencing without the need to clone, separating templates instead using microfluidics, dilution, bead-binding, and isolation within tiny reaction wells for sequencing reactions. This technique therefore avoids potential amplification and cloning biases, and also generates up to 400,000 sequencing reads per run. Current limitations are that sequencing chemistry is expensive, and read lengths are significantly shorter than traditional Sanger sequencing (100-250 bases versus 750+ bases). These shortened read-lengths create a trade-off between robust phylogenetic assignment (e.g. see Krause *et al.*, 2008) and phylogenetic discrimination. However, some analyses indicate that short reads can still allow profiling via the

16S rRNA gene with some confidence, at varying levels of resolution, depending on the assignment method used (e.g. Liu *et al.*, 2007, Sundquist *et al.*, 2007), particularly when variable regions within the 16S gene are targeted (e.g. Sogin *et al.*, 2006, Huber *et al.*, 2007, Roesch *et al.*, 2007). In addition, technology improvements are producing ever-lengthening reads (e.g., 500 bases by the end of 2008 with 454 technology, and an estimated 20kb by Pacific Biosciences in 2010, Korlach *et al.*, 2008).

DNA microarrays:

In general terms, DNA microarrays are a hybridization platform, with DNA probes immobilized on a substrate (often a glass slide), used to query a mix of nucleic acids for complementary sequences. The query mix is labeled with a fluorophore such that its hybridization of to the immobilized probes can be visualized. A single array can contain tens of thousands of probes, deposited using robotic spotters or synthesized in place using photolithography. Since their development, microarrays have been used primarily in gene expression studies to compare expression across different cellular types or conditions. However as the technology has matured, microarrays have been applied to a widening range of biological questions, including microbial ecology. Microarrays are currently applied in microbial ecology to assay the presence and relative abundance of particular organisms or genes (for reviews see Lucchini *et al.*, 2001, Zhou, 2003, Gentry *et al.*, 2006, Wagner, 2007).

There are two broad categories of microbial microarrays. The first, representing the majority of microarrays, target particular genes. There are two types of gene-specific arrays. In the first, arrays target putative functional guilds (e.g. sulfate reducers, nitrogen fixers, etc.), either via functional genes in the pathway(s) of interest or 16S genes in cases where 16S identity correlates to conserved metabolism (Small *et al.*, 2001, Wu *et al.*, 2001, Cho and Tiedje, 2002, Koizumi *et al.*, 2002, Loy *et al.*, 2002, Bodrossy *et al.*, 2003, Taroncher-

33

Oldenburg *et al.*, 2003, Greene and Voordouw, 2003, Stralis-Pavese *et al.*, 2004, Tiquia *et al.*, 2004, Rhee *et al.*, 2004, He *et al.*, 2007). The second type of gene-specific arrays are those which attempt to holistically profile a community, via its 16S diversity (e.g. Wilson *et al.*, 2002; Marcelino *et al.*, 2006; Palmer *et al.*, 2006; Brodie *et al.*, 2007; DeSantis *et al.*, 2007). For both types of gene-specific microarrays, PCR amplification of the targeted gene from the sample is typically (though not always) performed prior to or as part of the labeling reaction, such that labeled amplicons are hybridized to the arrays.

The second, less common category of arrays used in microbial ecology studies are community genome arrays (CGA) which use entire genomes as probes (e.g. Wu *et al.*, 2004, Wu *et al.*, 2006, Bae *et al.*, 2005), and evolved from earlier lower-throughput platforms whose use was dubbed "reverse sample genome probing" (RSGP; reviewed in Greene and Voordouw, 2003). Thus far, such arrays have relied on axenic cultures or isolates in order to generate the required genome probes. However, it has been suggested (Zhou, 2003; Greene and Voordouw, 2003) that environmental genomic surveys and large-insert clone libraries could instead be used to identify and generate genome fragment probes.

It is in this context that the genome-proxy array has been developed, and is described in this thesis. The genome proxy array is a hybrid of the two major categories of arrays currently used in microbial ecology, in that it targets genomes and genome fragments through many individual gene-specific oligonucleotide probes. In many respects it is conceptually more like a multi-species "comparative genome hybridization" (CGH) array. The multiple oligonucleotide probe design allows a finer-scale resolution than using entire genome fragments, and allows related cross-hybridizing sequences to be distinguished based on their hybridization patterns. In the genome proxy array, sets of 70-mer oligonucleotide probes (generally n=20 per genotype) were designed to different genomes and genome fragments derived from microbial assemblages found in the ocean, one of the most comprehensively characterized

and genomic sampled environments on earth. The majority of these targets (roughly two-thirds) were sequenced from in-house large-insert environmental genomic libraries, captured from the same sites under investigation. This array platform is described further in Chapter 2.

Microarrays have certain inherent limitations, as a technology based on hybridization. Microarrays can generally only provide information about what is represented on the array, or closely related sequences (although see Wang *et al.* 2002), and are thus fundamentally different from metagenomics in their ability to profile communities, and are more akin to methods like FISH. In addition, arrays provide relative rather than absolute quantification (although correlations between array signal and absolute abundance can be strong). The strength of microarray profiling is in the simultaneous tracking of many *distinct* organisms or genes, unlike FISH, FCM, or Q-PCR, and at a higher and more reliable level of resolution than fingerprinting methods.

Community Genomics, aka Metagenomics

There are two distinct methods for obtaining metagenomic data from a community. Environmental DNA can be cloned into small- (e.g. shotgun) or large-insert libraries (e.g. fosmid and BAC), and some or all of the clones can be sequenced, either in a random or a targeted (i.e. based on screening of the library for clones of interest) approach. Alternatively, new methods have allowed environmental DNA to be sequenced directly without cloning.

Environmental genomic libraries: Small-insert environmental genomic clone libraries(or shotgun clone libraries) have been used in a number of different environments to capture small genomic DNA fragments from the numerically dominant members of natural microbial assemblages (e.g. Tyson *et al.*, 2004, Venter *et al.*, 2004, Tringe *et al*, 2005, Strous *et al.*, 2006, Rusch *et al.*, 2007, Yooseph *et al.*, 2007, Kurokawa *et al.*, 2007). The relative success of small-insert metagenomic studies is directly related to the complexity of the community, the

35

amount of sequence obtained, and the goals involved. To date, these libraries have been used to reconstruct near complete "population genomes" from low complexity communities (e.g. Tyson *et al.*, 2004, Strous *et al.*, 2006), and also used for gene-centric approaches in more complex environments (e.g. Yutin *et al.*, 2007). In both cases, these data are used to look at the distribution and diversity of organisms and provide insight into their metabolic and functional capabilities.

Large-insert environmental genomic clone libraries typically employ a bacterial artificial chromosome (BAC, insert size ~20-160kb) or a fosmid (insert size ~35-40kb) to capture large genomic fragments from a cross-section of individual microorganisms from the environment. Such large-insert libraries have been constructed from a number of habitats and extensively screened for clones of interest (e.g. Rondon *et al.*, 2000, Béjà *et al.*, 2002a, Zeidner *et al.*, 2003, de la Torre *et al.*, 2003, Treusch *et al.*, 2004, Béjà *et al.*, 2000, Sabehi *et al.*, 2005, Frigaard *et al.*, 2006, Neufeld *et al.*, 2008). 16S- or ITS-profiling of libraries has also been proxy for profiling of the communities themselves (e.g. Suzuki *et al.* 2004, Martin-Cuadrado *et al.,* 2007, Neufeld *et al.*, 2008). End-sequencing of large-insert libraries has been used to describe the taxonomic and metabolic profile of the community (e.g. DeLong *et al.*, 2006, Appendix 4, Martin-Cuadrado *et al.,* 2007), while full-sequencing of particular clones has been used to investigate genomic context for particular groups or processes of interest (e.g. Béjà *et al.*, 2002a, Zeidner *et al.*, 2003, Hallam *et al.*, 2006, Frigaard *et al.*, 2006, Martinez *et al.*, 2007, McCarren *et al.*, 2007, Neufeld *et al.*, 2008).

Metagenomics without cloning: Cloning of unamplified total DNA (shotgun cloning) and subsequent sequencing is being eclipsed by highly-parallel, clone-free sequencing technologies, such as pyrosequencing (described above). Such clone-free sequencing avoids cloning biases, and is cheaper per base pair obtained. Currently this approach suffers only from short read lengths, but they are quickly increasing, see above. There have been a number of pyrosequencing

studies in microbial ecology (e.g. Mou *et al.*, 2008, Dinsdale *et al.*, 2008b, Wegley *et al.*, 2007) with perhaps the most advanced to date including a simultaneous comparison of 45 microbiomes and 42 viromes (Dinsdale *et al.*, 2008a), and an ocean microbial metagenome versus meta-transcriptome study (Frias-Lopez & Shi *et al.*, 2008). Although these studies offer previously unobtainable insights into microbial communities, made possible by the sheer depth of sequencing involved, care must be taken in comparing studies using unamplified DNA with those amplifying DNA prior to sequencing. Multiple displacement amplification (MDA), using the phi29 polymerase, has been used in a number of pyrosequencing studies (e.g. in some but not all of those reported in Dinsdale *et al.*, 2008b), and its biases in complex mixed community samples have not yet been described.

The metagenomic approach, by bypassing the preconceptions (e.g. about particular known sequence and metabolic diversity) inherent in the design and implementation of other profiling methods, offers microbial communities the clearest opportunity to "tell the story" of what is important in their world. In addition to the potential biases of pre-sequencing amplification, however, a major caveat of both cloning-based and cloning-free metagenomics is that the bulk of sequence space remains unexplained and undefined, with the majority of metagenome reads representing sequences of unknown function.

Community Profiling Conclusions

A wealth of profiling tools are available for characterizing microbial communities, and each has its strengths and weaknesses. As the price of sequencing continues to fall it will replace other methods, as it has already begun doing. In the interim, and to direct sample choice for community sequencing efforts, alternative profiling methods will continue to be useful, and remain widely applied in the field. Furthermore, some of these methods have uses other than just community profiling (e.g. FACS and FISH) and so will remain important tools

for years to come.

Presentation of the Genome Proxy Array in this Thesis

This thesis describes the development, testing, and application of a new microarray-based tool for community profiling. It builds upon a pre-existing knowledge of communities of interest derived from the sequencing of clones from environmental large-insert clone libraries, and of cultured isolates from related habitats. Like other indirect tools for microbial community profiling (Rohwer, 2007), the genome-proxy microarray is expected to be mostly obsolete within five years and entirely obsolete in 10, as sequencing costs decrease and massive sequencing is feasible for high-resolution spatial and temporal studies of microbial communities, in research labs at all funding tiers.

Chapter 2 of this thesis has two sections. The first places the genome proxy array in the context of marine microbial research, and expands upon the applications of other microarrays to microbial ecology. The second section describes the design and validation of a prototype of the genome proxy array, in a published paper.

Chapter 3 is a manuscript in preparation describing the design, development, and validation of the expanded genome proxy array, and its application to a time series in Monterey Bay.

Chapter 4 summarizes the work and outlines the next directions for the use of this tool during its remaining lifetime of relevance.

Appendices include the methods developed for this array platform, a primer on array design, and two papers I have been involved in during my PhD whose scope pertains to the topics covered in Chapter 1.
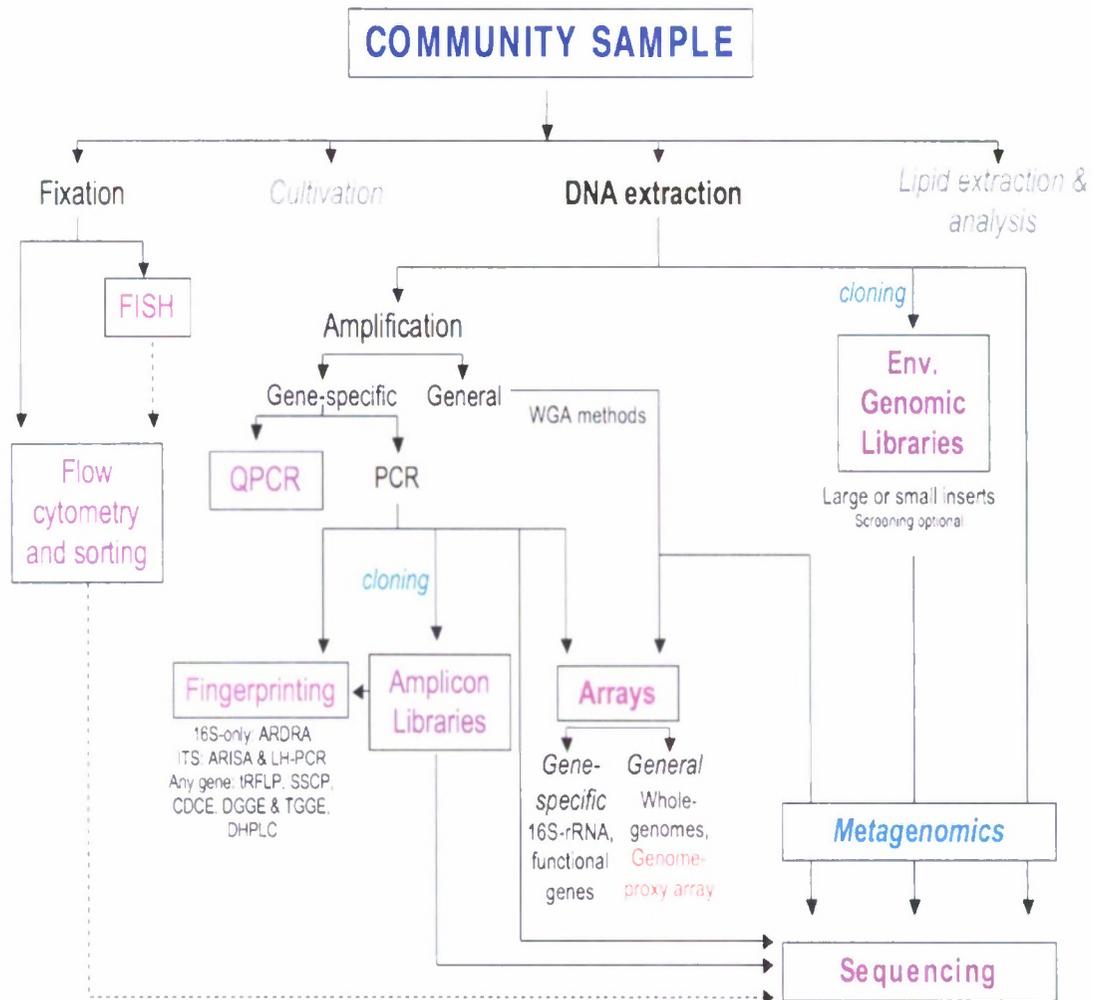
Figure 1. Community profiling methods in microbial ecology. A community sample may be treated in a number of ways during profiling, as presented in this review (greyed sections are not covered in depth). Dashed lines indicate less common links between methods.

Chapter 2

Development and Validation of a Prototype Genome Proxy Array.

2a. The case for the genome proxy array: motivation and context.

2b. Rich, V.I., K. Konstantinidis, and E.F. DeLong, 2008. Design and testing of 'genome-proxy' microarrays to profile marine microbial communities. *Environmental Microbiology*. 10(2): 506-521.

I. The Case for Microarrays in Marine Microbial Ecology

As details of the marine microbial communities have been revealed at ever-finer scales, the complexity and variability of marine microbial diversity continue to astonish (e.g., Acinas & Klepac-Ceraj et al., 2004, Thompson et al. 2005, Rocap et al., 2003). In tandem, the importance of these communities to global biogeochemistry has become increasingly evident (e.g. Howard et al., 2006, Karl et al., 2007, Moran and Miller, 2007, Kuenen et al., 2008). This deepening knowledge of, and respect for, marine microbial communities, particularly as scientific and societal concern over global change grows, has led to increasing research efforts devoted to further understanding their composition, dynamics, and functional capacities, and how observed genomic variability relates to functional variability at the level of organism and system.

Historically, the majority of marine microbial community studies have focused, of necessity, on specific phylogenetic or functional groups, or taken a coarser-grained higher-taxonomic-level approach. Recently new methods have allowed the scope of microbial community investigations to broaden without coarsening and to encompass entire co-occurring microbial communities and their metabolic potential. Despite the deluge of community genomic sequence information brought by the field of marine microbial metagenomics (e.g. Venter et al., 2004, DeLong et al., 2006 (Appendix 4), Rusch et al., 2007, Yooseph et al., 2007, Martin-Cuadrado et al., 2007, Mou et al., 2008, Dinsdale et al., 2008) our understanding of marine microbial ecology remains far from complete.

Thus far marine metagenomic studies have been snapshots of communities, and have been enormously informative but do not reveal community dynamics. Microbial ecology is at a cross-roads, with a need for repeated community-level sampling to better understand both the variability of the tools themselves (notably, metagenomics is single sequence datasets randomly sampled from large pooled sample), and the actual variability of communities across time and

space, under both natural and perturbed conditions. For the next several years, community genome sequencing will remain prohibitively costly for such widespread use for ecological studies. These investigations require inexpensive, high-throughput methods for assessing microbial diversity and function. Microarrays are a high-throughput tool that permit the highly-parallelized simultaneous query of many targets.

II. Previous and Current Microarray Applications in Microbial Ecology

Over the last 5-10 years, microarrays have become an increasingly common tool in microbial ecology and span increasingly-diverse designs and goals (recently reviewed in Gentry *et al.*, 2006, Wagner *et al.*, 2007). The majority of microbial microarrays designed for environmental use can be broadly grouped into three categories: functional gene arrays (FGAs), 16S arrays ("Phylochips"), and community genome arrays, with the last being least common.

1. Functional Gene Arrays. FGA probes may be either PCR products or oligonucleotides. Sequence information may be pulled directly from the environment of interest by PCR-amplification of the relevant functional gene(s), and then using the amplicons directly as probes, or cloning and sequencing them and designing oligonucleotides to target their diversity (e.g. Taroncher-Oldenburg *et al.*, 2003). In general, FGA studies include an amplification of the sample DNA prior to hybridization, using primers specific to the functional gene(s), to enrich for the sequences being targeted. FGAs have been developed for a number of important microbially-mediated biogeochemical transformations, such as methanotrophy (Bodrossy *et al.*, 2003, Stralis-Pavese *et al.*, 2004, Cébron *et al.*, 2007, Gebert *et al.*, 2008) and marine nitrogen cycling (Taroncher-Oldenburg *et al.*, 2003, Moisander *et al.*, 2007, Ward *et al.*, 2007, Zhang *et al.*, 2007).

43

The most ambitious FGA to date is the "GeoChip" platform designed by Jizhong Zhou and colleagues (reviewed in He *et al.*, 2007, with design details in Wu *et al.*, 2001, Rhee *et al.*, 2004, Tiquia *et al.*, 2004, He *et al.*, 2005, and Liebich *et al.*, 2006). Rather than targeting a single functional gene or pathway, the GeoChip targets most major nutrient uptake and transformation pathways that have been identified in cultivated organisms, with >24,000 probes targeting >10,000 genes, in more than 150 functional groups (see table 1 at end). These targeted groups span the carbon, nitrogen and sulfur cycles, as well as metal and organic pollutant transformations. Variants of the GeoChip have been used to investigate microbial responses to anthropogenic soil contamination from organic pollutants and metal (Rhee *et al.*, 2004) and radioelements (He *et al.*, 2007). In addition, the GeoChip has been used to assess microbial *activity*, by hybridization to amplified community RNA from a uranium-contaminated soil (Gao *et al.*, 2007).

2. Phylochips. Phylochips typically target 16S rRNA sequences, and typically have hierarchically-nested sets of probes tailored to differing levels of phylogenetic specificity. Hybridization usually occurs to PCR-amplified 16S rRNA genes from an environment, but may also use extracted unamplified 16S rRNA.

There are three major types of phylochips, based on their target breadth and selection. Phylochips may be designed to particular monophyletic functional guilds, to particular phylogenetic clades, or to a broad swath of phylogenetic diversity. The first type is valuable when investigating microbial metabolisms in environmental community settings, and complements the functional genes approach described above. Here, all phylotypes *known* to be responsible for a given process are targeted via their 16S sequences. Applications of this guild-specific Phylochip approach include sulfate-reducers (Small *et al.*, 2001; Koizumi *et al.*, 2002; Loy *et al.*, 2002) and hydrocarbon-degrading microbes (Koizumi *et*

al., 2002).

The second class of Phylochips target monophyletic clades, regardless of their functional homogeneity. Examples of these arrays include the "RHC-PhyloChip", targeted to *Rhodocyclales* (Loy *et al*, 2005), the "ECC-Phylochip" targeted to *Enterococcus* spp. (Lehner *et al.*, 2005), a marine *Vibrio*-specific array (employing both 16S and 23S probes, Marcelino *et al*, 2006), and an array for rhizosphere *Alphaproteobacteria* (Sanguin *et al.*, 2006).

The third class of Phylochips target many phylogenetic clades, such as a broad suite of pathogenic bacteria (via *gyrB* rather than 16S; Kostić *et al.*, 2007), or sediment genotypes (Peplies *et al.*, 2006, el Fantroussi *et al.*, 2003, Eyers *et al*, 2006, Neufeld *et al.*, 2006). Two such generalist-Phylochips warrant special mention for their remarkable breadth. The Brown and Relman Labs have developed a 16S-based array for profiling gut microbiota, the current iteration of which has ~3,100 species-level probes and ~6,000 group-level probes (Palmer *et al.*, 2007; the prototype version targeted 229 species and 130 higher nodes, Palmer *et al.*, 2006). This array has been used with great success in a ground-breaking study of the development of the gut microbiome in human infants (Palmer *et al.*, 2007). A second generalist Phylochip has been developed for environmental microbes by the Andersen Lab, and targets ~9,000 distinct OTUs (see Table 2 at end) (Brodie *et al.*, 2006, DeSantis *et al.*, 2005) and has been applied to the study of microbiota in urban aerosols (Brodie *et al.*, 2007), subsurface soils and waters (DeSantis *et al.*, 2007), as well as cystic lung fibrosis patients (V. Klepac-Ceraj, pers. comm.). A similarly large-scale phylochip has subsequently been published and used for oral microbial communities (Huyghe et al, 2008).

3. Community Genome Arrays (CGAs). In contrast to using 16S probes as a hook for target genotypes of interest, another approach is to target entire genomes or genome fragments. Such arrays have been developed for several habitats by using genomes of cultivated organisms and environmental isolates. CGA probes are entire genomes. rather than the oligonucleotides or PCR products typical of phylochips and FGAs. Community genome arrays have been successfully applied to explore dynamics of cultivars and isolates from soils, river and marine sediments (Wu *et al.* 2004, 2006), kimchi fermentation (149 lactic acid bacterial genomes were monitored during fermentation, Bae *et al.* 2005). The CGA approach was developed out of earlier lower-throughput community genome hybridization efforts, which had been used to explore oil fields, salt marshes, and acid mine drainage sediments (reviewed in Greene and Voordouw, 2003).

Community genome arrays generally have species-level specificity, and may even be specific to the targeted strain, depending on hybridization conditions. However, because each probe represents an entire genome, the cross-hybridization of different strains with similar overall identities to the target strain cannot be distinguished from one-another. In addition, with strain-level specificity sometimes difficult to achieve, the resolution of CGAs may be lower than the ideal for ecology studies, since recent genome research shows that closely related genotypes can have significant ecological differences (e.g., Thompson *et al.* 2005, Coleman *et al.* 2006). Most importantly, however, CGA as described uses cultivated organisms or isolates as targets, both of which are unlikely to represent groups that are abundant *in situ*.

To overcome this last limitation, the community genome array approach could work equally-well with cloned and captured genome fragments from the environment. This alternate design, of first conducting extensive genomic surveys of the environment of interest, and then designing an array using that sequence information, was described in the literature several years ago years

(Zhou, 2003; Greene and Voordouw, 2003), but has not yet been realized for the purposes of microbial ecology.

Environmental genomic libraries have previously been combined with microarray technology, but with the goal of screening the libraries rather than exploring the environment (Sebat *et al.*, 2003, Park *et al*, 2008). The library inserts themselves serve as probes on a microarray, which is then used to query pure cultures, enriched treatments, and natural communities. The results identify specific clones for further investigation, such as end- or full-sequencing; for example, if querying enrichments, the results can provide information on the potential importance of a given clone in a specific process. This application of microarrays – using the environment to look back and define a library – is in some ways the inverse of the community genome proxy approach.

4. The Virochip: Lastly, although viral ecology has not been the focus of this overview, a virally-targeted microarray platform has been developed that is conceptually distinct from the microbial arrays described above, and represents a highly successful application of microarray technology to complex natural communities. Furthermore, this alternative design was the inspiration for the genome proxy array described in this thesis. Wang *et al.*'s "virochips" (2002), like CGAs, start from the knowledge of entire genomes' sequences. The crucial difference is that instead of immobilizing the entire viral genomes, oligonucleotide probes for each open reading frame (ORF) are created. Using custom software (ArrayOligoSelector, Bozdech *et al.*, 2003) sets of 70-mer probes for each viral genome were targeted to sequences of high conservation among the viral genome database, on the theory that these conserved probes would be the most likely to pick out novel undescribed virotypes as well, particularly among quickly-evolving viruses.

The first iteration of the virochip was specific to viruses involved in head

47

colds, and was initially tested on RNA from infected tissue culture cells. The virochip was able to clearly identify the presence of targeted viruses from the pure viral cultures of tissue culture cells. In addition, related viruses could be detected through their conserved regions, showing distinctive hybridization patterns to some subset of the probes of their closest relatives – what Wang *et al.* (2002) called a "viral barcode". Next, the Virochip was used to examine natural community samples, by adding a random-amplified PCR step to their protocol to obtain cDNA pools from the nasal lavages of purposefully-infected and healthy control patients. The virochips were successfully able to distinguish which of the test viruses were present in the infected patients. Finally, sick patients with unknown and in some cases multiple viruses related to those targeted were tested, and the virochip could identify both the targeted and the novel viruses. The phylogenetic affiliation of these novel viruses could be determined based on their patterns of hybridization, their "barcodes", and was confirmed by RT-PCR with family-specific primers. Thus, even within complex natural samples, the virochip could distinguish related serotypes of a particular virus, and also place completely unknown samples in their phylogenetic context. Although nasal lavages are considerably less complex than a typical soil or aquatic microbial community, this research shows that by using not one but a whole suite of probes for a given species, with varying levels of specificity, a maximal amount of information and resolution can be obtained.

5. Limitations of Array Platforms. There are several important caveats in contemplating the use of microarrays in microbial ecology, related to their inherent limitations and also to the methods associated with their use.

First, arrays provide only relative rather than absolute quantification; different probe, even designed to the same hybridization parameters *in silico*, can behave differently (e.g. Kreil *et al.*, 2006). Thus while the correlation between

target abundance and array signal is often very high for a given probe, it may be vary considerably between probes. This effect can be ameliorated but not eliminated by the use of multiple probes per target. (This variability in probe sensitivity can be seen in the behavior of the genome proxy array. For the *Prochlorococcus* MED4 probe set, the signal correlation to cell concentration had an R^2 of 1.0, as seen in Figure 5a of Rich, Konstantinidis and DeLong, 2008. However, across all targeted genotypes, the correlation of array intensities to pyrosequencing-inferred target abundance ranged from an R^2 of 0.85 – 0.91, as seen in Figure 2 of Rich *et al.*, in prep., Chapter 3). The conclusion is that arrays are most robust for assessing the relative changes in *each* target between samples, and somewhat less accurate at quantifying the relative abundances *between* targets.

Second, microarrays can generally only provide information about what is represented on the array. Some platforms such as the larger Phylochips, the Virochip and the genome proxy array allow significant and meaningful cross-hybridization to genotypes not explicitly represented by the array, but are still limited to sequences related to those targeted. This "you can only see what you look for" drawback is not limited to arrays, and can have significant consequences to our ecological interpretations. Any method that brings a filter to our observation potentially excludes important data. For example, FGAs and guild-specific Phylochips rest upon the completeness of our understanding of the process of interest, at the time of the design of the array. This completeness may be flawed in several major ways. Probes in the above types of arrays examine only the already-recognized participants in the process of interest, and their close relatives. Recent discoveries in microbial ecology have proven that our picture of even things as basic as phototrophy may be much narrower than what is common – let alone present – in Nature (e.g., Béjà *et al.*, 2000b). Not only may there be as-yet undiscovered genes and pathways mediating processes we are trying to map, but the connection between 16S identity and coherence of

organism function is poorly understood. Organisms with highly similar 16S identities may have quite dissimilar overall gene contents and consequently occupy distinct niches and/or play different ecosystem roles. Finally, this potential myopia is exacerbated when using specific-primer-directed PCR in the creation and design of the probes or preparation of the target. Not only does PCR have potential reaction-based errors and biases (e.g. Thompson *et al.*, 2002), but also primers may not amplify as comprehensively as they're assumed to (for example, "universal"-primer-based surveys missed an entire high-level taxonomic group, the kingdom Nanoarchaeota; Huber *et al.*, 2002).

III. The Genome Proxy Array.

Over the last eight years, a number of large-insert (fosmid and BAC vector) genomic libraries from marine picoplankton (Béjà *et al.*, 2000a) collected from a variety of ecologically-relevant depths in two oceanic habitats (the coastal waters of Monterey Bay, and the oligotrophic open ocean at Station ALOHA in the North Pacific Subtropical Gyre) have become available. These libraries have been surveyed to characterize their phylogenetic and functional gene content (Béjà *et al.*, 2002a; Zeidner *et al.*, 2003; Suzuki *et al.* 2004; DeLong *et al.*, 2006; Hallam *et al.*, 2006), thousands of clones have been end-sequenced (DeLong *et al.*, 2006, Appendix 4), and many clones have been fully sequenced (Stein *et al.*, 1996, Béjà *et al.*, 2000b, Béjà *et al.*, 2002a, Béjà *et al.*, 2002b, de la Torre *et al.*, 2003, Sabehi *et al.*, 2004, Coleman *et al.*, 2006, Frigaard *et al.*, 2006, Grzymski *et al.*, 2006, Martinez *et al.*, 2007, McCarren *et al.*, 2007). This wealth of environmental genomic data provides an unprecedented window into marine microbes, the majority of which remain uncultivated, and their communities.

By designing a microarray with existing genomic survey data from the target

ecosystem, a broader sampling of the ecosystem's microbial diversity can occur. This has two important benefits: (i) the diversity being assessed is intimately linked to potential function, since a large piece of each organism's genome is known and many genes along that piece are being targeted, and (ii) because a large section of target sequence is known, multiple probes can be designed to target each genome, allowing more conclusive identification of community members, and meaningful cross-hybridization to their relatives.

From the sequence data in available environmental genomic libraries from Monterey Bay and Station ALOHA, I designed 70-mer oligonucleotide probes and created a prototype marine picoplankton microarray. The prototype "genome proxy" array targeted thirteen sequenced environmental large-insert clones and the full-sequenced marine cyanobacterium *Prochlorococcus* MED4. The development and testing details are described in the published paper Rich, Konstantinidis and DeLong, 2008, which comprises the second section of this chapter.

Chapter 2b

Design and testing of 'genome-proxy' microarrays to profile marine microbial communities.

Authors: Virginia Rich, Konstantinos Konstantinidis, Ed DeLong

Design and testing of 'genome-proxy' microarrays to profile marine microbial communities

Virginia I. Rich,[1] Konstantinos Konstantinidis[2][1] and Edward F. DeLong[1,2]*

[1] *The MIT-WHOI Joint Program in Biological Oceanography, and* [2] *The Department of Civil and Environmental Engineering, MIT, 48-427, 15 Vassar St., Cambridge, MA 02139, USA.*

Summary

Microarrays are useful tools for detecting and quantifying specific functional and phylogenetic genes in natural microbial communities. In order to track uncultivated microbial genotypes and their close relatives in an environmental context, we designed and implemented a 'genome-proxy' microarray that targets microbial genome fragments recovered directly from the environment. Fragments consisted of sequenced clones from large-insert genomic libraries from microbial communities in Monterey Bay, the Hawaii Ocean Time-series station ALOHA, and Antarctic coastal waters. In a prototype array, we designed probe sets to 13 of the sequenced genome fragments and to genomic regions of the cultivated cyanobacterium *Prochlorococcus* MED4. Each probe set consisted of multiple 70-mers, each targeting an individual open reading frame, and distributed along each ~40–160 kbp contiguous genomic region. The targeted organisms or clones, and close relatives, were hybridized to the array both as pure DNA mixtures and as additions of cells to a background of coastal seawater. This prototype array correctly identified the presence or absence of the target organisms and their relatives in laboratory mixes, with negligible cross-hybridization to organisms having ≤ ~75% genomic identity. In addition, the array correctly identified target cells added to a background of environmental DNA, with a limit of detection of ~0.1% of the community, corresponding to ~10^3 cells ml^{-1} in these samples. Signal correlated to cell concentration with an R^2 of 1.0 across six orders of magnitude. In addition, the array could track a related strain (at 86% genomic identity to that targeted) with a linearity of $R^2 = 0.9999$ and a limit of detection of ~1% of the community. Closely related genotypes were distinguishable by differing hybridization patterns across each probe set. This array's multiple-probe, 'genome-proxy' approach and consequent ability to track both target genotypes and their close relatives is important for the array's environmental application given the recent discoveries of considerable intrapopulation diversity within marine microbial communities.

Introduction

Microarrays are currently applied in microbial ecology to assay the presence of particular organisms or genes in the environment (for a recent review, see Gentry *et al.*, 2006). Both functional gene arrays and phylogenetic arrays have been developed for several systems and guilds, including sulfate reducers (Small *et al.*, 2001; Koizumi *et al.*, 2002; Loy *et al.*, 2002), methanotrophs (Bodrossy *et al.*, 2003, Stralis-Pavese *et al.*, 2004), hydrocarbon degraders (Koizumi *et al.*, 2002) and microbes involved in the nitrogen cycle (Wu *et al.*, 2001; Taroncher-Oldenburg *et al.*, 2003; Tiquia *et al.*, 2004), among others (Cho and Tiedje, 2002; Greene and Voordouw, 2003; Rhee *et al.*, 2004, He *et al.*, 2007).

In addition, several larger 16S microbial microarrays have been developed to widen the phylogenetic scope of diversity investigations (e.g. Wilson *et al.*, 2002; Marcelino *et al.*, 2006; Palmer *et al.*, 2006; Brodie *et al.*, 2007; DeSantis *et al.*, 2007), employing a hierarchical probe design (i.e. some probes specific to class, others to order, etc.) critical for robust interpretation. This approach ideally builds upon a thorough survey of the major rRNA gene diversity in an environment prior to the array design, as was undertaken for the human gut microflora prior to the construction of a rRNA-targeted gut-census array (Palmer *et al.*, 2006). Such phyloarrays provide a high-throughput approach to determining community composition, while functional gene arrays have the potential to map a community's functional capabilities. Few array platforms have yet emerged to bridge the two, however; the links between 16S identity and functional capacity remain unclear, as two organisms with highly similar 16S rRNA sequences may have distinct ecological capabilities and

genetic content (e.g. Rocap et al., 2003; Konstantinidis and Tiedje, 2005).

Another approach to designing microarrays would be to target genome fragments that represent both the phylogenetic and functional breadth of a community, to permit finer-scale tracking of populations. This requires genomic surveys of the environment of interest (as suggested by Greene and Voordouw, 2003; Zhou, 2003) to provide the native genomic sequence information for probe design. While several arrays have been designed using some sequences derived from the community of interest (for example with environmental isolates' DNA, Greene and Voordouw, 2003;Wu et al., 2004; or 1 kb inserts from a groundwater cosmid library, Sebat et al., 2003), a more comprehensive use of uncultivated genomic data for arrays has not yet been reported.

We describe here the design and implementation of a prototype oligonucleotide microarray targeting environmentally occurring bacterial and archaeal genotypes, which were characterized through the recovery and analyses of large genomic fragments from marine plankton. A number of large-insert genomic libraries have previously been constructed from marine picoplankton collected from different depths in several oceanic habitats: the coastal waters of Monterey Bay (Béjà et al., 2000a), the oligotrophic open ocean at the Hawaii Ocean Time series (DeLong et al., 2006), and Antarctic coastal waters (Béjà et al., 2002a). These libraries have been surveyed to characterize their phylogenetic and functional content (Béjà et al., 2002a; Zeidner et al., 2003; Suzuki et al., 2004; DeLong et al., 2006; Hallam et al., 2006), tens of thousands of clones have been end-sequenced (DeLong et al., 2006), and hundreds of clones have been fully sequenced (Stein et al., 1996; Béjà et al., 2000b; 2002a,b; de la Torre et al., 2003; Sabehi et al., 2004; Coleman et al., 2006; Frigaard et al., 2006; Grzymski et al., 2006; McCarren and DeLong, 2007; Martinez et al., 2007).

The 'genome-proxy' array targeted ecologically relevant marine microbes through sets of probes designed to these genome fragments, which served as 'proxies' for the genomes of these uncultivated, unsequenced microbes. The array's specificity and sensitivity were tested against laboratory mixes, and to cells added to natural seawater samples at a variety of concentrations, under various hybridization conditions.

Results

Array design

The prototype microarray targeted 13 BAC or fosmid genome fragments (20–160 kb) from both bacteria and archaea (Table 1), recovered from a variety of marine

Table 1. Targets of the microarray.

GenBank accession number	Clone/organism name	Phylogenetic affiliation (by 16S placement or top BLAST hits)	Interesting gene content	Clone size	Number of probes	Reference
AY372455	HOT_02C01	Proteobacteria; Alphaproteobacteria	Proteorhodopsin	42	20	de la Torre et al. (2003)
EU221239	EB000_55B11	Proteobacteria; Alphaproteobacteria Rhodobacter-like	Proteorhodopsin	40	20	This publication
AE008921	EB000_60D04	Proteobacteria; Alphaproteobacteria Roseobacter-like	Bacteriochlorophyll Superoperon	104	20	Béjà et al. (2002b)
AY458637	EB750_02H05	Proteobacteria; Alphaproteobacteria; SAR11	rRNA operon	40	20	Suzuki et al. (2004)
EF089400	EB000_41B09	Proteobacteria; Betaproteobacteria	Proteorhodopsin + carotenoid biosynthesis cluster (crt)	44	20	McCarren and DeLong (2007)
AY458645	EB080_L12H07	Proteobacteria; Betaproteobacteria Nitrosomonas	rRNA operon	25	20	Suzuki et al. (2004)
AF279106	EB000_31A08	Proteobacteria; Gemmaproteobacteria; SAR86-II	Proteorhodopsin rRNA operon	129	40	de la Torre et al. (2003)
AY619685	HOT_04E07	Proteobacteria; Gammaproteobacteria; SAR86-I	Proteorhodopsin rRNA operon	71	20	Sabehi et al. (2004)
AE008919	EB000_65D09	Proteobacteria; Gammaproteobacteria	Bacteriochlorophyll Superoperon	60	20	Béjà et al. (2002b)
AY458650	EB750_10A10	Proteobacteria; Gammaproteobacteria	Form II RuBisCo (cbbM)	45	20	Suzuki et al. (2004)
U40238	ORE_4B7	Crenarchaeota; Marine Group I	rRNA operon	41	20	Stein et al. (1996)
AF393466	ANT_74A4	Crenarchaeota; Marine Group I	rRNA operon	40	20	Béjà et al. (2002a)
AF268611	EB000_37F11	Euryarchaeota; Marine Group II	rRNA operon	60	20	Béjà et al. (2000a)
EU221238	EF100_57A08	Euryarchaeota; Marine Group II	Portion of 16S Rrna	40	20	This publication
BX548174	Prochlorococcus MED4	Cyanobacteria	N/A	3 × 80 kb regions	3 × 20	Rocap et al. (2003)

Library name	Collection location	Collection depth (m)	Date of collection	Vector type	Library reference
EB000	Monterey Bay: 36.7°N, 122.4°W; near Station M2	3	3/17/99	BAC	Béjà *et al.* (2002b)
EB080	Monterey Bay, 36°45.50N, 122°02.10W	80	7/23/99	BAC	Suzuki *et al.* (2004)
EF100	Monterey Bay, 36°45.50N, 122°02.10W	100	2/21/02	pEFlFos	Suzuki *et al.* (2004)
EB750	Monterey Bay, 36°41.13N, 122°02.37W	750	4/11/00	BAC	Suzuki *et al.* (2004)
HOT	Station ALOHA, 22.75°N, 158°W	0	12/11/01	pIndigoBAC536	de la Torre (2003)
ANT	Coastal waters near Palmer Station, Anvers Island, Antarctica	0	8/96	pFos1	Béjà *et al.* (2002a)
ORE	Oregon coast (44°02.729N, 124°57.309W)	200	8/29/92	pFos1	Stein *et al.* (1996)

habitats (Table 2), as well as the cyanobacterium *Prochlorococcus* MED4. These clones were originally sequenced because of the presence of taxonomic marker or specific functional genes. This array consisted of sets of 70 bp oligonucleotides targeting each genome or genome fragment (Fig. 1), dispersed along the target sequences with no more than one probe per gene, and excluding rRNA genes as targets. The probes were selected solely based on theoretical thermodynamic properties and GC content (~40%); that is, probe selection did not focus on specific genes or regions, but simply produced the 'optimal' probes for each genome proxy based on the probes' predicted hybridization properties. rRNA genes were excluded, because this probe design approach, which avoids sequence alignments and considerations of RNA secondary structure, would be unlikely to result in useful rRNA probes. Furthermore, rRNA probes of traditional design could not be included on the array because their appropriate hybridization conditions would be very different from those of this array's probes.

Array specificity

When hybridized to mixtures of cloned environmental genomic DNA targeted by the array, the array produced signal from the correct probe sets, with no appreciable cross-hybridization to other probe sets. For example, a mix of DNA from clones ORE_4B7, EB750_10A10, EB080_L12H07 and HOT_02C01 produced above-background signals from only the corresponding probe sets on the array (data not shown). When equal amounts of each clone DNA were hybridized, the mean signal for each genotype was not equivalent, reflecting microarrays' relative – rather than absolute – quantification abilities due to variability in probe hybridization signal (e.g. Kreil *et al.*, 2006). The use of multiple probes to target many genes from each organism helped to normalize probe-to-probe heterogeneity, by averaging across all probes in a set (as described below). The evenness of probe response across each genotype's set was also used to evaluate the relatedness of hybridizing DNA (see below).

To more precisely define the array's phylogenetic range and specificity, it was tested against DNA from *Prochlorococcus* MED4 and related strains, spanning the known range of *Prochlorococcus* phylogenetic diversity (Fig. 2a and Table 3). The majority of tests used four strains: MED4, the strain explicitly targeted by the array; MtT9515, the only cultivated sister strain to MED4 within the high-light clade II (clade definitions *sensu* Rocap
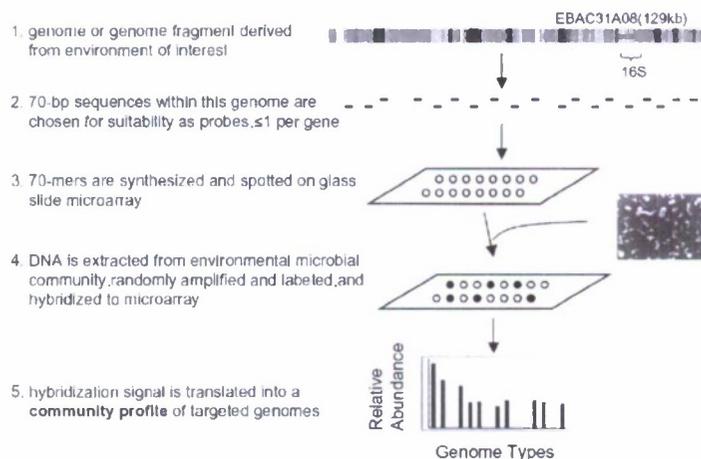
Fig. 1. Overview of array design and use.

1. genome or genome fragment derived from environment of interest

EBAC31A08(129kb)

16S

2. 70-bp sequences within this genome are chosen for suitability as probes,≤1 per gene

3. 70-mers are synthesized and spotted on glass slide microarray

4. DNA is extracted from environmental microbial community,randomly amplified and labeled,and hybridized to microarray

5. hybridization signal is translated into a **community profile** of targeted genomes

Relative Abundance

Genome Types

a.) Prochlorococcus phylogeny:

tree modified from Rocap et al. 2002.

MED4
MIT9515
} "High light" clade I

MIT 9211
NATL2A } "Low light" clade I
NATL1A
SS120 } "Low light" clade IV
MIT9313

MIT9312 } "High light" clade II

0.1 substitutions per position

b.) Relative signal of each strain to array probes:

- MED4 DNA only
- MIT9515 DNA only
- MIT9312 DNA only
- MIT9313 DNA only

Mean Normalized Intensity

Genotype Targeted:

c.) Hybridization pattern across all MED4 probes:

Normalized Intensity

Each Individual MED4 probe

d.) Targeted regions' position along MED4 genome:

1   2*   3

Mean Normalized Intensity

Region 1   Region 2   Region 3

Fig. 2. Specificity tests with *Prochlorococcus* strains.
a. The *Prochlorococcus* strains tested span the clade's phylogenetic diversity (ITS-based tree modified from Rocap et al., 2002).
b. The mean normalized signal across the *Prochlorococcus* MED4 probe set was highest when the array was hybridized to pure MED4 DNA. Hybridization separately to other *Prochlorococcus* strains produced decreasing signal as the phylogenetic distance increased. The hybridization data shown here and in (c) and (d) were from 1.8 ng of each strain's DNA.
c. The evenness of the signal across the probe set also decreased with increasing evolutionary distance.
d. The 'genome proxy' test: the location of the three 80 kb targeted regions of the *Prochlorococcus* MED4 genome, and the relative genomic identity between strains MED4 and MIT9313 (also see Table 3 for these three regions' relative genomic identity to other strains), and a representative example of the signal across the three probe sets designed to each region when hybridized to MED4 and related strains MIT9312 and MIT9313. Region 1 is 0.80 kbp along the MED4 genome, region 2 is 1.29–1.38 Mbp along, and region 3 is 1.58–1.66 Mbp along. The asterisk indicates that region 2 spans IsI5 island of inter-strain genomic variability identified by Coleman and colleagues (2006).

et al., 2002); MIT9312, from the high-light clade I; and strain MIT9313, from the low-light clade IV. DNA from low-light strains NATL2A, MIT9211 and SS120 was also tested.

As expected, the set of probes targeting *Prochlorococcus* MED4 all produced signal when hybridized to pure MED4 DNA (Fig. 2b and c). When other *Prochlorococcus* strains were each hybridized separately to the array, both

Table 3. *Prochlorococcus* strain relatedness.

| Strain | 16S identity to strain MED4 | Overall ANI[a] to strain MED4 | ANI to MED4 | | |
			Region 1	Region 2[b]	Region 3
MED4	100%	100% (1658)[c]	100% (80)	100% (80)	100% (78)
MIT9515	99.9%	86% (1433)	86% (78)	85% (46)	87.5% (78)
MIT9312	99.1%	78.5% (1422)	78% (79)	77% (37)	79% (78)
MIT9313	97.9%	64.5% (403)	64.5% (20)	65% (6)	64% (28)

a. ANI is average nucleotide identity, calculated per Konstantinidis and Tiedje (2005).
b. Region 2 spans the ISL5 genomically variable island described in Coleman and colleagues (2006).
c. In parentheses are the number of non-overlapping 1000 bp fragments with BLAST-based identity.

the mean signal and the evenness of signal across the MED4 probes decreased as the phylogenetic distance to MED4 increased (Fig. 2b and c). Consistently across many hybridizations, MED4 probes showed strongest signal when hybridized to strain MED4, moderate signal to strain MIT9515 (86% genomic identity to MED4), lower signal to strain MIT9312 (78.5% genomic identity), and no significant signal to strain MIT9313 (64.5% genomic identity). More distantly related strains NATL2A, MIT9211 and SS120 produced no appreciable signal (data not shown). Furthermore, the probe sets targeting environmental clones did not show appreciable cross-hybridization signal against *Prochlorococcus* (Fig. 2b). While the overall *Prochlorococcus* signal intensity decreased from MED4 to MIT9515 and MIT9312, the distribution of signal across the individual probes in the MED4 set also became less even. For example, in the hybridization shown in Fig. 2c, the coefficient of variation (CV) among the probes increased from 0.79 for MED4 to 2.40 for MIT9312; across the positive control probes in the same two hybridizations, the CV was 1.01 and 0.97 respectively.

To test the effects of hybridization stringency on the specificity and signal of the MED4 probes, *Prochlorococcus* strains were hybridized at a range of conditions (data not shown). In general, lowering stringency by decreasing the temperature (65°C, 60°C, 55°C to 50°C), or increasing the salt concentration in the hybridization buffer (3× SSC, 0.2% SDS to 3.5× SSC, 0.3% SDS), produced a decrease in the array's dynamic range (i.e. less signal difference between low and high concentrations of a given strain), and poorer discrimination among related strains. The protocol giving the best dynamic range and discrimination among strains (65°C and 3× SSC, 0.2% SDS buffer, see *Experimental procedures*) was used for the data reported here unless otherwise noted.

To test whether the specificity results for *Prochlorococcus* were comparable for other targeted clades, two genome fragments recovered from closely related phylotypes within the SAR86 clade of the gammaproteobacteria were represented on the array, and were tested for specificity. The clones HOT_04E07 from subclade SAR86-I (clade placement per Sabehi *et al.*, 2004) and EB000_31A08 from subclade SAR86-II are syntenic, 97.5% identical at their 16S genes, and share 72% genomic identity [calculated as average nucleotide identity (ANI), Konstantinidis and Tiedje, 2005]. When the array was hybridized separately to DNA from either of these two clones, the signal of the probes targeting the other was within the background signal (data not shown), as expected from the *Prochlorococcus* results. These results demonstrate that the specificity of the arrays can distinguish between closely related phylotypes of yet-uncultivated microorganisms.

Effect of designing probe sets to different regions of a target genome

To understand the equivalence of probe sets targeting different regions of the same organism's genome, we targeted three 80 kb 'genome-proxy' regions of the *Prochlorococcus* MED4 genome. One of the regions fell in a genomic 'island' where inter-strain variability is concentrated ('ISL5' in Coleman *et al.*, 2006). Shared gene content among strains was variable between the three regions, while the sequence identity (as ANI) of shared genes among strains was very similar between the regions (Table 3).

When hybridized to DNA from MED4 and related strains, the cumulative signal across the three regions' probe sets was not identical (Fig. 2d), as expected given probe-to-probe signal variability (e.g. Kreil *et al.*, 2006), and given the three regions' differences among strains. For example, between the target strain MED4 and the strain MIT9515, Region II shows 57.5% shared gene content (46 of 80 genes) and 85% genomic identity, while region III shows 100% shared gene content (78 of 78 genes) and 87.5% genomic identity. The hybridization signal for each region was calculated as the mean signal across all probes designed from that region. Despite the differences among genomic regions, the probe sets designed to all three regions were effective at identifying both targeted and related genotypes. Each region's probe set produced maximal signal to MED4, with decreasing signal to the other strains as phylogenetic distance increased (Fig. 2d). Across the three regions, this relative decrease in signal from MED4 to MIT9515 and MIT9312 was correlated more to relative genomic identity than to relative shared gene content (average Pearson correlation of 0.90 versus 0.70).

Array response to target cells in natural seawater

To test the array in a complex environmental context, we collected coastal seawater (lacking detectable *Prochlorococcus* cells by flow cytometry) and added *Prochlorococcus* cells from strains MED4, MIT9515, MIT9312 and MIT9313 over a range of concentrations from ~10^1 to 10^6 cells ml^{-1} (Fig. 3). The seawater was then filtered, and the DNA was extracted, amplified, labelled and hybridized to the array. The results in this background of environmental DNA agreed generally with earlier specificity results using DNA from laboratory cultures. MED4 probes showed strong signal when hybridized to strain MED4, moderate signal to strain MIT9515, and no significant signal to MIT9313 (Fig. 4a and b). In this environmental background, the relative signal from strain MIT9312 was markedly lower than observed in single-strain laboratory hybridizations (Fig. 4b versus Fig. 2b). Thus, the
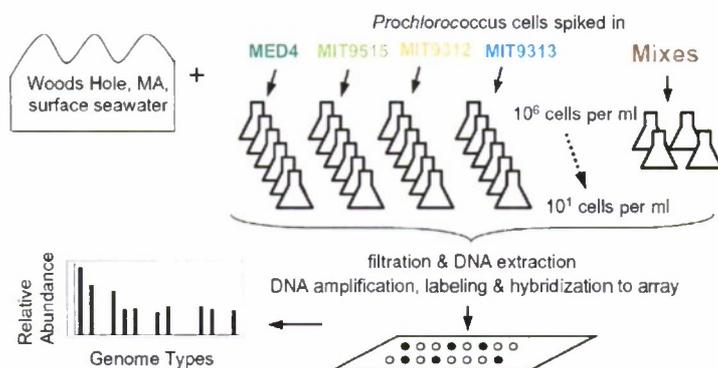
Fig. 3. Design of the *Prochlorococcus* addition experiment to a natural community. Coastal seawater was collected, and cells from *Prochlorococcus* strains MED4, MIT9515, MIT9312 and MIT9313 were spiked in at a range of concentrations from ~10^1 to 10^6 cells ml^{-1}.

operational phylogenetic breadth of the array for tracking related genotypes was near 86% genomic identity to the target.

To further explore the relationship between array signal and genomic identity to the target, in the absence of a cultivated strain bridging the relatedness between strains MED4 and MIT9515, we examined the subset of the MED4 probes with the highest identity to strain MIT9515. These probes had an average identity of 92.4% to strain MIT9515, while the average genomic identity between MED4 and MIT9515 is 86%. The signal across these probes from the MIT9515 hybridization was intermediate between the MED4 and MIT9515 signals (Fig. S3a and b). There was a clear linear increase in array signal with increasing genotype identity to target (R^2 of 0.9959, from 78.5% to 100% identity, Fig. 4b inset).

Correlation between cell numbers and signal

As the cell concentrations of the targeted strain MED4 increased across six orders of magnitude within a complex natural community, the mean signal intensity across the MED4 probe set increased linearly, with an R^2 of 1.0 (Fig. 5a). At the lowest cell concentrations the signal diverged from this linear relationship, so that the operational limit of detection was ~10^3 cells ml^{-1} of the target, which in these coastal water samples represented ~0.1% of the community.

To test whether this linearity would hold for tracking related, non-target genotypes, we examined the cumulative MED4 probe signal with varying cell concentrations of strain MIT9515 (86% genomic identity to the targeted strain MED4). As cell concentrations of strain MIT9515 increased in the background of environmental cells, the mean normalized intensity across the MED4 probe set increased linearly, with an R^2 of 0.9999 across six orders of magnitude. For this non-target strain, the limit of detection was around 10^4 cells ml^{-1}, representing approximately 1% of the community (Fig. 5a).

There was no appreciable correlation between cell concentrations of the more distantly related *Prochlorococcus* strains and mean normalized signal across the MED4 probes, across this concentration range (R^2 of 0.04 for strain MIT9312 and 0.31 for strain MIT9313; data not shown).

Array data metrics

The array data could be examined in two ways, either probe-by-probe across each probe set, or as overall organism signals. In addition, at a given hybridization stringency, different treatments of the data might result in different degrees of apparent cross-hybridization between strains, different *in silico* stringency. In order to determine which method of converting the individual probe signals into an overall organism signal gave optimal discrimination between, or optimal cross-hybridization among, strains, and gave optimal correlation to cell concentration, the data were analysed using different combinations of metrics. We focused primarily on the data from the *Prochlorococcus* addition experiment, as being the most informative and representative of environmental data sets. All data presented above were obtained using the optimized analysis, described below.

The analysis pipeline began by taking either the mean or median of replicate spots of each probe, minus the mean, median or Tukey Biweight of the negative control probe set. [The Tukey Biweight is the used in Affymetrix's MAS5 analysis methods to calculate the signal across sets of 11–20 oligonucleotide probes targeting a single open reading frame (ORF) (Affymetrix, 2002); it weights each value based on its proximity to the median, thereby reducing the effect of outliers.] To remove the effects of non-discriminatory probes, we tested what minimum per cent (Y%) of the probes in each probe set should be required to show signal (greater than 1× or 2× mean background, or greater than 1× or 2× mean negative control) for that probe set to be considered 'present'. Y

a.)

Mean Normalized Intensity

MIT 9313 { 10^1 ... 10^6 } MIT 9312 { 10^1 ... 10^6 } MIT 9515 { 10^1 ... 10^6 } MED4 { 10^1 ... 10^6 }

Mixes {
■ Mix 1: 10^5 MIT9312, others 10^3
■ Mix 2: 10^5 MIT9515, others 10^3
■ Mix 3: 10^5 MED4, others 10^3
■ Mix 4: all 10^5
}

Genotype Targeted

4B7, ANT74A4, EB037F11, EB041B09, EB055B11, EB060D04, EB065D09, EB0_31A08, EB75002H05, EB75010A10, EB80L12H07, EF10057A08, HOT2C01, HOT4E07, ProMED4, Positive Control, Negative Control

b.)

Prochlorococcus MED4 probeset only :

Mean Normalized Intensity

Signal vs. Identity to MED4:

Mean Normalized Intensity

$y = 321.32x - 252.31$
$R^2 = 0.9959$

"Virtual strain" 9515 data, but using subset of probes with higher identity

Nucleotide Identity to Target Strain MED4

Pro MED4
Genotype Targeted

could not equal 100 because some probes were poor performers, and because we wished to retain the signal from related, non-target genotypes. Next, a single intensity value for each probe set was calculated as the mean, median and Tukey Biweight of the probe signal across each probe set. Finally, each value was normalized for array-to-array brightness by the mean, median or Tukey Biweight of the positive control probe set.

At the optimized hybridization conditions, the combination of metrics that gave the best correlation between cell concentration and signal, and also produced cross-hybridization signal of the MED4 probes to the related

strain MIT9515, was the following: median among replicates, then the mean signal across probes, minus the mean of the negative control probes, normalized to the mean of the positive control probes, with at least Y = 45% of the probes required to produce signal greater than 2× mean negative control.

By lumping related genotypes together as a single signal, this combination of metrics had only ~10-fold difference limit of resolution between samples (e.g. in Fig. 5a, and 10^5 cells ml^{-1} of MIT9515 gave approximately the same signal as would ~2.5×10^4 cells ml^{-1} of MED4), and missed underlying changes in population structure.
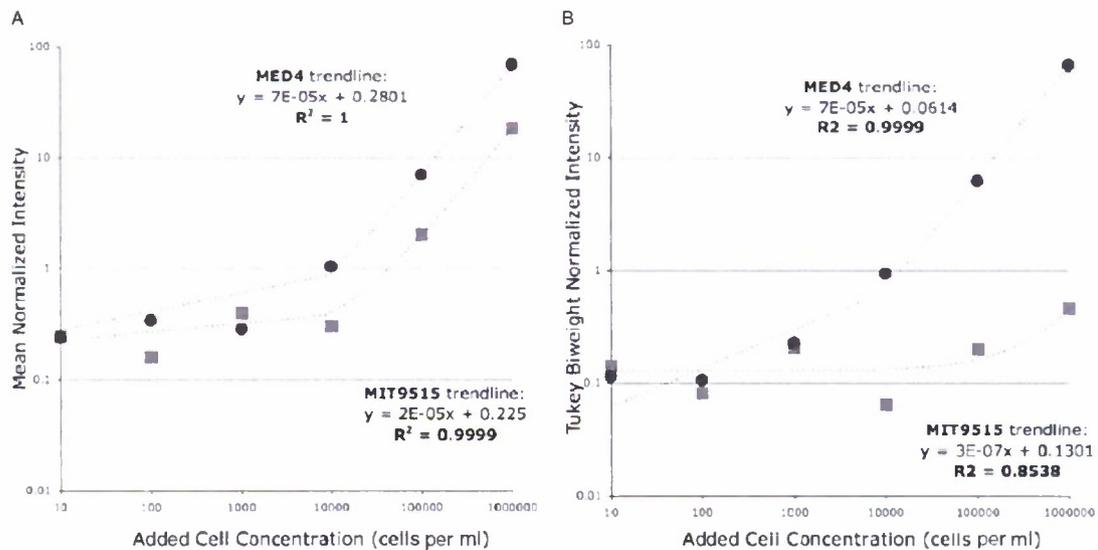
Fig. 5. Signal versus cell concentration.
a. As the MED4 cell concentration increased across six orders of magnitude, the mean normalized intensity across the MED4 probe set (filled circle) increased linearly with an R^2 of 1.0. The operational limit of detection is around 10^3 cells ml^{-1} of the target, which in these coastal water samples represents approximately 0.1% of the community. As cell concentrations of strain MIT9515 increased in the background of environmental cells, the mean normalized intensity (filled square) across the MED4 probe set increased linearly, with an R^2 of 0.99 across six orders of magnitude. For this non-target strain with 99.9% 16S rRNA identity and 86% overall genomic identity to the target, the limit of detection is around 10^4 cells ml^{-1}, representing approximately 1% of the community.
b. If the normalized Tukey Biweight signal across the probe set is used instead of the mean, the stringency of the hybridization is increased dramatically *in silico* such that there is much greater separation of the MED4 and MIT9515 signals.

Thus, a secondary metric was also used, the Tukey Biweight across probes in a set, in which non-target signal was dramatically reduced (Figs 5b and S1). For example, target signal to 10^6 cells ml^{-1} of MED4 decreased only 4% between the mean and Tukey Biweight metrics, while non-target signal decreased to 10^6 cells ml^{-1} of MIT9515 decreased 98% (Figs 5b and S1). To explore the effect of using the Tukey Biweight as the genomic identity to target increased, we also examined the 'virtual' strain composed of the subset of MED4 probes with the highest nucleotide identities to strain MIT9515 (92.4%, versus 86% between the strains overall, as described above). The Tukey Biweight reduced the signal from these probes, e.g. by 40% for the 10^6 cells ml^{-1} data, resulting in a signal intermediate between the MED4 signal and the whole probeset MIT9515 signal (Fig. S3b).

To better distinguish target from non-target (but related) signal, we also tested metrics for measuring the signal evenness across each probe set. (For example, 10^5 cells ml^{-1} of MIT9515 gave a different pattern of hybridization across the MED4 probe set than would ~2.5×10^4 cells ml^{-1} of MED4, despite their mean signal being identical; Fig. 5a versus Figs 2c and S2c,e). We calculated the Shannon, Simpsons and modified Simpsons evenness metrics borrowed from ecology, and the CV. The CV maximized the separation between the target strain MED4 and the related, detected strain MIT9515 (data not shown); however, none of these evenness metrics captured the sequential pattern of hybridization across probes. These measures could distinguish target from non-target populations, but could not track shifts between two related, non-target populations.

To further compare hybridization patterns between samples, we tested the Pearson correlations between the probe-by-probe signals of each probe set. The hybridization pattern was consistent within strains, across all concentrations above the limit of detection. The Pearson correlation of the probe-by-probe signal for the *Prochlorococcus* probes was significantly higher between any two hybridizations of the same strain, than it was between strains. For strains MED4, MIT9515 and MIT9312 (divergent strain MIT9313 was omitted because its cross-hybridization was near-background signal), at 10^3–10^6 cells ml^{-1}, Pearson correlation within strains was on average 0.73 (SD = 0.18), versus 0.44 on average across any two strains (SD = 0.18). These correlations were significantly different at P = 0.000 by a Student's equal-variance (satisfied by F-test = 0.95) two-tailed t-test. To test the effect of higher genomic identities, the Pearson correlations were also calculated using the higher-identity 'virtual strain' probe set. For these probes' patterns in the hybridizations to MED4, MIT9515 and MIT9312, the average within-strain correlation was 0.74 (SD = 0.18), while the between-strain correlation dropped to 0.49

(SD = 0.18), and these were significantly different at P = 0.000 (F-test = 0.95). Lastly, exact replicates of the same cell concentration produced a higher Pearson correlation, as might be expected. The same amount of positive control DNA was added (pre-amplification and labelling) to all *Prochlorococcus* addition experiments, so was represented by 27 replicates at the optimal hybridization conditions reported here. The hybridization patterns from the positive control probe set had an average Pearson of 0.90 (SD = 0.13) between replicates. Thus the hybridization pattern across a probe set can be compared between samples, to allow the discrimination of different populations of cells.

Array response to mixed populations of related genotypes

Mixtures of *Prochlorococcus* strains were also added to seawater samples to test the performance of the array when challenged with mixtures of the target and its relatives, in a community background. Specifically, four mixtures were tested: (i) 10^5 cells ml^{-1} of all four strains (MED4, MIT9515, MIT9312, MIT9313); (ii) 10^5 cells ml^{-1} of MED4 and 10^3 cells ml^{-1} of the other three strains; (iii) 10^5 cells ml^{-1} of MIT9515 and 10^3 cells ml^{-1} of the other three strains; and (iv) 10^5 cells ml^{-1} of MIT9312 and 10^3 cells ml^{-1} of the other three strains. The cumulative signal from each mix was essentially equivalent to the additive signal of the component populations; the presence of related genotypes did not interfere with the hybridization of strains that could bind the MED4 probes (Fig. 4a). Furthermore, by Pearson correlation, the mixed samples in which one genotype dominated gave patterns that were distinct from one another. The pattern produced by Mix 3 (dominated by MED4 cells) was different from that by Mix 2 (dominated by MIT9515 cells), with a Pearson of only 0.38 (Fig. S2). Lastly, by using the Tukey Biweight instead of the mean, the contribution of non-target cells was markedly reduced (Fig. S1).

Array-based observations of natural community microbes

In addition to the added *Prochlorococcus* cells, the coastal water samples from Woods Hole showed the consistent presence of several of the targeted genotypes in the natural microbial community. Probe sets designed to proteorhodopsin-containing environmental clones EB000_41B09, EB000_55B11, and EB000_31A08, and bacteriochlorophyll operon-containing clone EB000_60D04, all showed above-background signal at consistent levels across almost all of the 27 experiments hybridized at the optimal conditions (Fig. 4a). This signal persisted even when using the Tukey Biweight metric

(Fig. S1), which reduces signal from related non-target genotypes, suggesting that the genotypes present in the natural seawater were closely related to the targeted genotypes. Furthermore, the signal from each of these four probe sets showed high agreement in pattern among hybridizations. The average Pearson among all EB000_31A08 hybridizations was 0.74 (SD = 0.18), for EB000_60D04 the average was 0.84 (SD = 0.12), for EB000_55B11 it was 0.88 (SD = 0.09), and for EB000_41B09 it was 0.81 (SD = 0.17). This suggests that the genotypes present were similar among samples tested.

Discussion

The development of the genome-proxy microarray approach was motivated by the need for high-throughput tools to track marine microbes in a community context. The array described here represents a complementary approach to existing array platforms for microbial ecology. While previous microarrays used in microbial ecology have primarily targeted single functional or phylogenetic genes, the arrays described in this report target uncultivated microbes through 'genome proxies', fragments of native genomes captured from the environment. Each genome fragment was targeted with a set of probes, selected based on predicted hybridization characteristics and GC content. Each probe targeting a given genome could then 'vote' on the presence of the target organism, averaging across the variable individual probe responses. By targeting sections of genomes rather than single genes, this array was able to track clusters of related genotypes in the environment at a relatively high level of resolution, while simultaneously tracking a subset of each genotype's individual genes. These abilities distinguish the 'genome-proxy' array from single-gene arrays.

We characterized the phylogenetic specificity and experimental sensitivity of this array. It was able to detect targeted organisms in simple laboratory mixtures, and in complex backgrounds of environmental DNA. *Prochlorococcus* was used as the primary test clade (Fig. 2a) because of its relevance in marine plankton, and the availability of culturable strains, many with associated genomic and ecological information (e.g. Partensky *et al.*, 1999; Moore *et al.*, 2002; Scanlan and West, 2002; Rocap *et al.*, 2003; Bouman *et al.*, 2006; Coleman *et al.*, 2006; Johnson *et al.*, 2006; Zinser *et al.*, 2006; Garczarek *et al.*, 2007). Thus, the degree of cross-hybridization to the *Prochlorococcus* MED4 probes by DNA from other strains could be placed in the context of their genetic and ecological relatedness, providing a model for the array's phylogenetic specificity within an environmental context.

Under the hybridization conditions and analysis methods used, in hybridizations to both pure DNAs and to cells spiked into natural community samples, the array showed negligible cross-hybridization to distantly related genotypes [less than ~78.5% genomic nucleotide identity (ANI) to the target] (Figs 2b and 4b). Cross-hybridization signal from indiscriminate probes was further removed by requiring each probe set to show signal above a threshold value in a certain percentage of its probes. Closely related genotypes were consistently detected by the array (at 86% genomic identity to target) (Figs 2b and 4b). There was a strong correlation ($R^2 = 0.9959$) between the mean signal and the identity of the hybridized genotype to the target, above 78.5% genomic identity (Fig. 4b inset).

This ability to track both targets and their relatives represented both a benefit and a challenge. If the signal from all detected relatives of a given target were lumped together into a single signal for that target, then the array's limit of resolution would be ~10-fold change between samples, with cross-hybridization to relatives indistinguishable from up to 10-fold changes in target abundance, and underlying changes in population structure would be missed. However, the array's multiprobe design allowed for more nuanced analysis. Probe sets could 'vote' through either a permissive metric, the mean signal across the set, or a non-permissive metric, the Tukey Biweight signal across the set (e.g. Fig. 5a and b). Thus, from a single hybridization, the data could be interpreted broadly or stringently *in silico*, to cast the net narrowly for the targeted organisms or more broadly to include their close relatives, down to at least ~86% genomic identity.

The evenness and pattern of the signal across the probe set also provided important information and allowed discrimination of the target genotype from that of its relatives, and close relatives from one another (Fig. 2c). The probe-by-probe signal patterns of a given strain in different hybridizations were significantly more highly correlated to one another, regardless of cell concentration, than to the patterns of different strains. In mixtures of related genotypes, the pattern of the most abundant genotype dominated, such that shifts in population structure between mixes were evidenced by quite distinct hybridization patterns (Fig. S2). This feature of the array allowed it to track shifts in population structure between samples spiked with cells of single and multiple strains.

To understand the ecology of organisms over time, it is ideal to track not only their presence and absence but also their relative abundance, making it important to understand how the microarray signal related to the target organism's abundance. This array showed a highly linear relationship between cell concentrations and signal, even in an environmental background. This linearity held for both the targeted strain (MED4; $R^2 = 1.0$) and its relative (MIT9515; $R^2 = 0.9999$) when using the mean normalized intensity across probes (Fig. 5a). The limit of detection for

the targeted strain MED4 was approximately 10^3 cells ml^{-1}, or ~0.1% of the community, and 10^4 cells ml^{-1} for strain MIT9515, ~1% of the community. This limit of detection is equivalent to or below that reported for other recent environmental microbiology microarrays (e.g. Rhee *et al.*, 2004; Loy *et al.*, 2005; Gentry *et al.*, 2006).

Not only was the array able to track added target cells and their relatives in a complex background, but it also identified the likely presence of other targets in coastal waters, in a different oceanic province than those from which the target clones originated. The genome proxies whose probe sets produced signal in the Woods Hole water contain proteorhodopsin or bacteriochlorophyll operons, and represent putatively phototrophic organisms, which are predicted to occur in such a habitat (Béjà *et al.*, 2002b). The consistency of their array signal in samples from many aliquots of adjacent water, by both overall mean and Tukey Biweight signal, and by hybridization pattern (Fig. 4a), strongly suggested that each target was present in the community, and that its population structure did not vary significantly across the spatial scales spanned by these aliquots.

The design approach of using suites of probes to assay for each organism was a crucial feature of this array. The power of the multiprobe-per-target design approach has been employed by other array platforms, although their different goals have required distinct design and analysis strategies. For example, Affymetrix arrays use multiple short oligonucleotide probes (in some cases tiled at regular intervals) to assay each gene or gene product, but seek very high specificity, whereas our arrays seek to track both target sequences and their relatives, within a complex environmental background. Our goals are more comparable to that of the 'Virochip' microarray used for viral identification (Wang *et al.*, 2002) in clinical samples. However, its design employed viral genome alignments, with hierarchical probes selected to conserved and variable regions. In contrast, the approach described here made use of the higher degree of sequence conservation within microbes (compared with viruses). By selecting oligonucleotide probes based primarily on their hybridization kinetics and without requiring alignments to related sequences, we sidestep the problem of limited and differentially distributed sequence coverage in different habitats and of different clades.

The use of this genome proxy array raises the question as to whether an organism can be targeted based on a subset of its genome: Do probe sets designed to different genomic regions give substantially different results for the presence, absence or relative abundance of an organism? The environmental clones targeted by this array represent 20–160 kb sections of genome. Population genomic variability is unevenly dispersed along genomes, concentrated in hypervariable regions (e.g. as

in *Prochlorococcus*, Coleman *et al.*, 2006), such that some percentage of environmental genomic clones capture hypervariable genomic regions. If such regions were targeted, the resulting probe sets might be so genotype-specific that they would produce little cross-hybridization to close relatives. However, we do not anticipate this being a significant problem in the use and expansion of this array, for several reasons. First, the environmental clones that are sequenced and used for probe design tend to be 16S-containing clones, and 16S operons are not in hypervariable regions. Second, even when somewhat variable regions are captured and targeted, as in this microarray's second targeted region of the *Prochlorococcus* genome, which spanned the ISL5 island of inter-strain variability, the probe sets still cross-hybridize to related genotypes (Fig. 2d). Signal intensity was correlated more strongly with the identity of shared genes than with the overall shared gene content in a region. Thus, except in extreme cases of hypervariable island capture (which would likely be identifiable by gene content anomalies, i.e. high numbers of integrases, transposases and hypothetical genes, and therefore avoided), targeting environmental genomic clones as described here should allow the subsequent tracking of their relatives. Furthermore, this approach is robust to genomic rearrangements among strains, as it assays the presence or absence of sections of DNA rather than their relative positions.

Frequently observed in the oceans, highly similar but non-identical microbial genotypes tend to share a high degree of synteny and minimal nucleotide variation across their genomes (Coleman *et al.*, 2006; Rusch *et al.*, 2007). For example, only 3–5% variation in nucleotide identity in surface-ocean *Prochlorococcus* MIT9312-like sequences is usually observed in natural populations (Coleman *et al.*, 2006; Rusch *et al.*, 2007). Thus, the array's ability to track related genotypes suggests its suitability for identifying and tracking microbes at the relevant levels of sequence divergence found in native microbial populations. The empirical results with *Prochlorococcus* genome fragments, along with the SAR86 and pufLM-containing genotypes we detected in Woods Hole seawater, also support this conclusion.

Furthermore, this degree of specificity should allow the arrays to detect previously unrecognized ecotypes within uncultivated target lineages. Overall genomic identity is clearly more sensitive than 16S rRNA identity at discerning closely related populations, with organisms highly similar at the 16S level sometimes occupying quite distinct niche space (e.g. Jaspers and Overmann, 2004; Hahn and Pöckl, 2005; Johnson *et al.*, 2006). The microarray approach described here has the potential to track shifts in populations of closely related genotypes under changing environmental conditions.

In cases where a functional gene of interest is present on a targeted clone, these arrays also may be able to match the distribution and expression of the gene to that of its 'owner'. This tool has the potential to simultaneously assay both DNA and RNA from environmental samples, to track not only which targeted genotypes are present but also which are functionally active. This will improve our understanding of microbial activity and dormancy in different environmental conditions. It may also indicate when functionally important genotypes are missing from our targets, for example when the DNA and RNA signal for a given gene is high but that of its genome proxy overall is low. There is even the possibility of using the array elements as capture probe, to further characterize novel environmental sequences, as hybridized DNA and RNA can be recovered directly from arrays to be clones and sequenced (Wang et al., 2003).

We expect the 'genome-proxy' oligonucleotide microarray to be a useful tool for conducting high-throughput investigations of microbial distributions, community dynamics and functional activity. The multiprobe design strategy results in hybridization signal that is dependent on genomic identity to the targeted organism, across the region targeted. This allows not only the tracking of clusters of related genotypes in the environment, but also the distinction among related genotypes, by using the pattern and evenness of the signal across each probe set as a 'barcode' of each different genotype. This ability gives the array the potential to map shifts in population structure. In addition, this allows the array's use in geographically disparate but similar habitats, as considerable sequence divergence is tolerated. No sequence alignments are required, obviating the need for coverage of the phylogenetic space surrounding the targeted organism. Also, by using genome proxies rather than single genes to target organisms, there is additional 20–160 kb genomic context available, potentially expandable by locating contigs.

With these prototype arrays now validated, we are constructing an expanded microarray representing hundreds of genotypes from different depths in open and coastal oceans. These will be used to track microbial community and population changes in time-series datasets (with accompanying physical and chemical data) to provide a higher-resolution understanding of the dynamics of marine microbial communities.

Experimental procedures

Culturing and DNA extractions

Prochlorococcus strains MED4, MIT9515, MIT9312, NATL2A, MIT9211 and SS120 and MIT9313 were grown in 250 ml^{-1} l cultures of Sargasso seawater-based Pro99 medium (Moore et al., 2002), under continuous light conditions at 20°C. High-light strains (per Rocap et al., 2002) were grown at 35 μmol photon m^{-2} s^{-1} light intensity, while low-light strains were grown at 18–20 μmol photon m^{-2} s^{-1}. DNA was extracted according to a modified phenol-chloroform protocol (Steglich et al., 2003) and treated with RNase at 50 μg ml^{-1} final concentration for 37°C for 1 h, then re-extracted. DNA from the DeLong Lab library environmental clones used in this study was extracted from overnight cultures using either Qiagen miniprep kits (Qiagen, Valencia, California) or an AutoGenprep 960 (AutoGen, Holliston, Massachusetts) automated extraction robot, followed by treatment to digest *Escherichia coli* DNA with ATP-dependent exonuclease (Epicentre, Madison, WI) according to the manufacturer's instructions. DNA concentrations were measured using an ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, Delaware). The positive control *Halobacterium salinarum* NRC-1 DNA was purchased (#700922, ATTC, Virginia).

Additions of *Prochlorococcus* to natural seawater samples

Coastal seawater was collected from the Woods Hole, MA, town pier using a rinsed bucket and transported to MIT in a 50 l carbuoy; *Prochlorococcus* was undetected in this water by flow cytometry (per Moore et al., 1998), and total cell density was 4.15×10^6 by Sybr-stained flow cytometric counts. *Prochlorococcus* strains MED4, MIT9515, MIT9312 and MIT9313 were separately spiked into the natural samples, each to final cell concentrations ranging from ~10^1 to 10^6 cells ml^{-1}. Culture cell concentrations were measured using flow cytometry, and necessary dilutions of cultures were made with 0.2-μm-filtered Sargasso Sea water. Each aliquot of Woods Hole water with its spiked-in *Prochlorococcus* was filtered through a GF-A prefilter, then collected on a Supor-200 (#60300, Pall Corporation, Ann Arbor, MI) 0.2 μm filter, using a MasterFlex peristaltic pump system (Cole-Parmer Instrument Company, Vernon Hills, IL). Filtered volumes ranged from 250 ml to 1 l. Filters were immediately frozen. All additions and filtrations were made within 24 h of water collection.

Extractions were a modification of a filter extraction protocol described previously (Suzuki et al., 2001). Filters were transferred to 2.0 ml screw-top microcentrifuge tubes, and 242 μl of lysis buffer was added to each [lysis buffer: 40 mM EDTA, 50 mM Tris pH 8.3, 0.73 M sucrose, 1.15 mg ml^{-1} lysozyme (Sigma, #L-6876), 200 μg ml^{-1} RNase (Qiagen, Valencia, CA, #1018048), 0.2 μm-filter-sterilized]. Samples were incubated at 37°C for 30 min, rotating. In total, 13.5 μl of a Proteinase K solution [10 mg ml^{-1} (EMD, #24568-2) in 40 mM EDTA, 50 mM Tris pH 8.3, 0.73 M sucrose] was added, and SDS was added to a final concentration of 1%. Each sample was incubated at 55°C, rotating, overnight. The samples were then extracted with the DNeasy 96 Tissue kit (Qiagen, Valencia, CA), by a modification of the manufacturer's protocol. Each tube received 300 μl of Buffer AL (buffer AL/E without ethanol added), was vortexed, and incubated for 70°C for 10 min. Then 300 μl of 99% ethanol was added to each, they were vortexed, and pipetted onto the 96-well spin plate. The plate was sealed with the Airpore sheets (supplied with kit) and spun. All spins were carried out at 40°C, 4612 *g* in a Sorvall Legend RT centrifuge (Kendro

Laboratory Products, Newtown, CT). The plate was spun 10 min, 500 µl Buffer AW1 was added to each well, and the plate re-sealed and spun 5 min. In total, 500 µl Buffer AW2 was then added to each well, and the plate re-sealed and spun 5 min. To dry the plate, the column portion was then transferred to a new rack of elution microtubes RS (supplied with kit) and incubated for 15 min at 70°C. To elute, 200 µl Buffer AE preheated to 70°C was then added to each well, the plate was re-sealed, incubated 1 min and spun for 2 min. The elution was repeated with an additional 200 µl. The eluted DNA was then concentrated using Excela-Pure 96-well PCR purification kits (Edge BioSystems, Gaithersburg, MD), following the manufacturer's protocol. Each well was rinsed once with 100 µl nuclease-free water (#9937, Ambion, Austin, TX), then resuspended in 20 µl dilute TE (1 mM Tris pH 8, 0.1 mM EDTA pH 8), transferred to a clean 96-well plate, and stored at –20°C. Concentrations were measured by Nanodrop.

Microarray probe design

Microarray 70-mer probes were designed using the program ArrayOligoSelector (Zhu et al., 2003) with the following settings: target %GC = 40%, 1 probe/gene, with the ORFs for each genome fragment as both the input and the database file. The output candidate 70-mers were then sorted based on their %GC and those closest to 40% were chosen. In the case of more than the target number of probes having 40% GC, the subset with the lowest free energy of hybridization were selected as probes. Generally, 20 probes were selected per organism. *Prochlorococcus* MED4 was represented by 60 probes total, 20 each for three different 80 kb 'genome-proxy' regions: 0–80 kb, 1.29–1.37 Mbp, and 1.58–1.66 Mbp.

Using the same method, a set (*n* = 20) of positive control probes were designed to the genome of the halophillic archaeon *H. salinarum* NRC-1. Negative control probes (*n* = 28) were designed to a set of 49 random 1000-base sequences (Stothard, 2000). All probes sequences and specifications are available online in the Gene Expression Omnibus (GEO).

Probe and target comparisons in silico

Targets were compared *in silico* in several ways. Target relatedness was measured for closely related organisms (for example within the SAR86 or *Prochlorococcus* clades) by both 16S rRNA gene identity and ANI. 16S gene identity was calculated using the Distance Matrix (DNADist format, Jukes Cantor corrected) feature of the Ribosomal Database Project (Cole et al., 2007). Genomic identity of related genomes and genome fragments was calculated as ANI, as described by Konstantinidis and Tiedje (2005).

Microarray construction and hybridization

Oligonucleotides were synthesized (Illumina, San Diego, CA), suspended in 3× SSC to a concentration of 40 pmol µl⁻¹, and spotted on homemade poly-L-lysine-coated glass slides using a QArray 2 microarraying robot (Genetix, Hampshire, UK). Six replicates of each probe were spotted.

For the experiments shown, target DNA was amplified and labelled using A/B/C random amplification (Wang et al., 2003), with the modification that the initial reverse transcription step was omitted. Briefly, random-primed amplification was carried out in three reactions: round A used Sequenase to extend primer A (GTT TCC CAG TCA CGA TCN NNN NNN NN); round B used 20 rounds of PCR to amplify the resulting fragments, using primer B (GTT TCC CAG TCA CGA TC); and round C used 10 rounds of PCR to incorporate amino-allyl-deoxyuridine triphosphates (aa-dUTP). For the environmental samples, the amount of DNA into each reaction was normalized to represent 70 ml of filtered seawater. All A/B/C reactions were performed in triplicate and pooled. Amplification products were cleaned using a Microcon YM-30 and concentrated to 9 µl in nuclease-free water, and labelled with Cy3 by combining 8 µl aa-DNA, 2 µl 0.5 M NaHCO3 and 5 µl Cy3 dye (33 µg in DMSO), and incubating at room temperature in the dark for 1 h. Samples were cleaned in a Microcon YM-30, concentrated to 19 µl in TE, and 17.33 µl was added to hybridization buffer for final concentrations of 3× SSC, 0.2% SDS, 0.4 mg ml⁻¹ poly A, 0.02 M Hepes, pH 7, in a final volume of 25 µl. Samples were denatured 4 min at 100°C, then pipetted onto the arrays. Arrays were hybridized overnight in a heating oven (Model, 2000 Micro Hybridization Incubator, Robbins Scientific, Sunnyvale, CA), then washed, first vigorously for 30 s in 0.6× SSC, 0.03% SDS, and second in 0.06× SSC vigorously for 30 s then gently for 5 min. For the data shown in this paper, hybridizations were carried out at 65°C and washes were performed at room temperature.

Microarray data analysis

Hybridized arrays were scanned using an Axon Instruments 4000B scanner (Foster City, CA), and the data were normalized and filtered using perl scripts written for the purpose, by the following steps. (i) Signal intensities for each spot were calculated by subtracting the local background (mean F532 – median B532, as calculated by GenePix Pro 5.1 software, Axon Instruments). (ii) The median value across replicates was calculated for each probe. (iii) For each probe set, the number of probes greater than twice the mean negative control signal was calculated, before further processing. (iv) Filter I: Arrays with less than half their positive control probes exceeding twice the mean negative control signal were considered poor quality, low dynamic range, arrays and were excluded from further analysis. (v) Each probe signal was corrected for non-specific binding by subtracting the mean negative control spot signal. (VI) The data were then normalized for array-to-array variations in brightness by dividing each probe signal by the mean positive control signal. This positive control signal was the mean signal across the *H. salinarum* probes in each hybridization, with identical amounts of *H. salinarum* DNA having been added to each reaction prior to amplification and labelling. (VII) Filter II: In order for a genotype to be considered 'present', at least 45% of its probes had to exceed twice the mean negative control signal. (VIII) Finally, each genotype signal was calculated as either the mean or Tukey Biweight across its probe set.

The Tukey Biweight was calculated as follows (Affymetrix, 2002). For n probes in a given probe set, the individual probe values are $x_1, x_2, \ldots, x_n$, after earlier pre-processing steps. m is the median of these values for a given probe set. MAD = weighted median of these values = median $(|x_1 - m|, |x_2 - m|, \ldots, |x_n - m|)$. For each probe, its distance from the centre is calculated as $t_i = (x_i - m)/(5^*MAD + 0.0001)$; where $i = 1, 2, \ldots, n$. Weights for each probe value then are calculated by the bisquare function, $B(t) = (1 - t^2)^2$ for $|t| < 1$, or $B(t) = 0$ for $|t| = 1$. Then the Tukey Biweight (TBW) can be for the probe set as a whole across n probes with values $x_1, x_2, \ldots, x_n$: $TBW(x_1, x_2, \ldots, x_n) = (\Sigma_{i=1} B(t_i) (x_i)/[\Sigma_{i=1} B(t_i)]$.

In each experiment, four metrics of evenness were calculated for each probe set. These were the Shannon's index of evenness, the Simpson's index of evenness, the Simpson's modified index of evenness (all Magurran, 1988) and the CV. They were calculated as follows:

Shannon's index of evenness

$$E_{Shannon} = \left[-\sum p_i \ln(p_i) \right] / \ln(n)$$

As above, n = the number of probes in the probe set. $p_i = x_i/X$, where X is the summed signal across all probes in that set.

Simpson's index of evenness:

$$E_{Simpson} = \left[1/\sum p_i^2 \right]/n$$

Simpson's modified index of evenness:

$$E_{Simpson\,modified} = \left[1/\sum x_i(x_i - 1)/X(X - 1) \right]/n$$

Coefficient of variation

$$CV = \left[1/n * \sum (x_i - a)^2 \right]^{0.5}/a$$

Where a is the mean value of x across each probe set.

Finally, for the *Prochlorococcus* addition experiment only, outlier arrays were identified as having normalized mean positive control signal less than 25% of the average across the experimental series, and were excluded from further analyses.

For all experiments, pre-processed data were imported into Excel for visualization, and the raw data are available online at GEO.

Data deposition

The sequences of the environmental clones EB000_55B11 and EF100_57A08 have been deposited in GenBank under Accession Nos. EU221238-9. Microarray data are MIAME compliant and have been deposited in GEO under platform Accession No. GPL6012.

References

Affymetrix (2002) *Statistical Algorithms Description Document*. Technical Report. Santa Clara, CA, USA: Affymetrix.

Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., *et al.* (2000a) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902–1906.

Béjà, O., Suzuki, M.T., Koonin, E.V., Aravind, L., Hadd, A., Nguyen, L.P., *et al.* (2000b) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**: 516–529.

Béjà, O., Koonin, E.V., Aravind, L., Taylor, L.T., Seitz, H., Stein, J.L., *et al.* (2002a) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Appl Environ Microbiol* **68**: 335–345.

Béjà, O., Suzuki, M.T., Heidelberg, J.F., Nelson, W.C., Preston, C.M., Hamada, T., *et al.* (2002b) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**: 630–633.

Bodrossy, L., Stralis-Pavese, N., Murrell, J.C., Radajewski, S., Weilharter, A., and Sessitsch, A. (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environ Microbiol* **5**: 566–582.

Bouman, H.A., Ulloa, O., Scanlan, D.J., Zwirglmaier, K., Li, W.K., Platt, T., *et al.* (2006) Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. *Science* **312**: 918–921.

Brodie, E.L., DeSantis, T.Z., Moberg Parker, J.P., Zubietta, I.X., Piceno, Y.M., and Andersen, G.L. (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proc Natl Acad Sci USA* **104**: 299–304.

Cho, J., and Tiedje, J. (2002) Quantitative detection of microbial genes by using DNA microarrays. *Appl Environ Microbiol* **68**: 1425–1430.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., *et al.* (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* **35**: D169–D172.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768–1770.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U., *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

DeSantis, T.Z., Brodie, E.L., Moberg, J.P., Zubieta, I.X., Piceno, Y.M., and Anderson, G.L. (2007) High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. *Microb Ecol* **53**: 371–383.

Frigaard, N.U., Martinez, A., Mincer, T.J., and DeLong, E.F. (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439:** 847–850.

Garczarek, L., Dufresne, A., Rousvoal, S., West, N.J., Mazard, S., Marie, D., *et al.* (2007) High vertical and low horizontal diversity of *Prochlorococcus* ecotypes in the Mediterranean Sea in summer. *FEMS Microbiol Ecol* **60:** 189–206.

Gentry, T.J., Wickham, G.S., Schadt, C.W., He, Z., and Zhou, J. (2006) Microarray applications in microbial ecology research. *Microb Ecol* **52:** 159–175.

Greene, E.A., and Voordouw, G. (2003) Analysis of environmental microbial communities by reverse sample genome probing. *J Microbiol Methods* **53:** 211–219.

Grzymski, J.J., Carter, B.J., DeLong, E.F., Feldman, R.A., Ghadiri, A., and Murray, A.E. (2006) Comparative genomics of DNA fragments from six Antarctic marine planktonic bacteria. *Appl Environ Microbiol* **72:** 1532–1541.

Hahn, M.W., and Pöckl, M. (2005) Ecotypes ot planktonic actinobacteria with identical 16S rRNA genes adapted to thermal niches in temperate, subtropical, and tropical freshwater habitats. *Appl Environ Microbiol* **71:** 766–773.

Hallam, S.J., Konstantinidis, K.T., Putnam, N., Schleper, C., Watanabe, Y.-I., Sugahara, J., *et al.* (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum. Proc Natl Acad Sci USA* **103:** 18296–18301.

He, Z., Gentry, T.J., Schadt, C.W., Wu, L., Liebich, J., Chong, S.C., *et al.* (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J* **1:** 67–77.

Jaspers, E., and Overmann, J. (2004) Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Appl Environ Microbiol* **70:** 4831–4839.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M.S., and Chisholm, S.W. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311:** 1737–1740.

Koizumi, Y., Kelly, J.J., Nakagawa, T., Urakawa, H., El-Fantroussi, S., Al-Muzaini, S., *et al.* (2002) Parallel characterization of anaerobic toluene- and ethylbenzene-degrading microbial consortia by PCR-denaturing gradient gel electrophoresis, RNA-DNA membrane hybridization, and DNA microarray technology. *Appl Environ Microbiol* **68:** 3215–3225.

Konstantinidis, K.T., and Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102:** 2567–2572.

Kreil, D.P., Russell, R.R., and Russell, S. (2006) Microarray oligonucleotide probes. *Methods in Enzymology* **410:** 73–98.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J., *et al.* (2002) Oligonucleotide microarray for 16s rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol* **68:** 5064–5081.

Loy, A., Schulz, C., Lucker, S., Schopfer-Wendels, A., Stoecker, K., Baranyi, C., *et al.* (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the Betaproteobacterial order 'Rhodocyclales'. *Appl Environ Microbiol* **71:** 1373–1386.

McCarren, J., and Delong, E.F. (2007) Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. *Environ Microbiol* **9:** 846–858.

Magurran, A.E. (1988) *Ecological Diversity and Its Measurement.* Princeton, NJ, USA: Princeton University Press.

Marcelino, L.A., Backman, V., Donaldson, A., Steadman, C., Thompson, J.R., Preheim, S.P., *et al.* (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. *Proc Natl Acad Sci USA* **103:** 13629–13634.

Martinez, A., Bradley, A.S., Waldbauer, J.R., Summons, R.E., and Delong, E.F. (2007) Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. *Proc Natl Acad Sci USA* **104:** 5590–5595.

Moore, L.R., Rocap, G., and Chisholm, S.W. (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393:** 464–467.

Moore, L.R., Post, A., Rocap, G., and Chisholm, S.W. (2002) Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus. Limnol Oceanogr* **47:** 989–996.

Palmer, C., Bik, E.M., Eisen, M.B., Eckburg, P.B., Sana, T.R., Wolber, P.K., *et al.* (2006) Rapid quantitative profiling of complex microbial populations. *Nucleic Acids Res* **34:** e5.

Partensky, F., Hess, W.R., and Vaulot, D. (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63:** 106–127.

Rhee, S.-K., Liu, X., Wu, L., Chong, S.C., Wan, X., and Zhou, J. (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-Mer oligonucleotide microarrays. *Appl Environ Microbiol* **70:** 4303–4317.

Rocap, G., Distel, D.L., Waterbury, J.B., and Chisholm, S.W. (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68:** 1180–1191.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424:** 1042–1047.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., *et al.* (2007) The. *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5:** e77.

Sabehi, G., Beja, O., Suzuki, M.T., Preston, C.M., and DeLong, E.F. (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. *Environ Microbiol* **6:** 903–910.

Scanlan, D.J., and West, N.J. (2002) Molecular ecology of the marine cyanobacterial genera *Prochlorococcus and Synechococcus. FEMS Microb Ecol* **40:** 1–12.

Sebat, J.L., Colwell, F.S., and Crawford, R.L. (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl Environ Microbiol* **69:** 4927–4934.

Small, J., Call, D.R., Brockman, F.J., Straub, T.M., and Chandler, D.P. (2001) Direct detection of 16s rRNA in soil extracts by using oligonucleotide microarrays. *Appl Environ Microbiol* 67: 4708–4716.

Steglich, C., A.F.Post and W.R.Hess. (2003) Analysis of natural populations of *Prochlorococcus* spp. in the northern Red Sea using phycoerythrin gene sequences. *Environ Microbiol* 5: 681–690.

Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H., and DeLong, E.F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* 178: 591–599.

Stothard, T. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 28: 1102–1104. [WWW document]. URL http://posnania.cbio.psu.edu/sms/rand_dna.html

Stralis-Pavese, N., Sessitsch, A., Weilharter, A., Reichenauer, T., Riesing, J., Csontos, J., *et al.* (2004) Optimization of diagnostic microarray for application in analysing landfill methanotroph communities under different plant covers. *Environ Microbiol* 6: 347–363.

Suzuki, M.T., Preston, C.M., Charez, F.P., and DeLong, E.F. (2001) Quantitative mapping of bacterioplankton populations in seawater: field tests across an upwelling plume in Monterey Bay. *Aquat Microb Ecol* 24: 117–127.

Suzuki, M.T., Preston, C.M., Béjà, O., de la Torre, J.R., Steward, G.F., and DeLong, E.F. (2004) Phylogenetic screening of ribosomal RNA gene-containing clones in Bacterial Artificial Chromosome (BAC) libraries from different depths in Monterey Bay. *Microb Ecol* 48: 473–488.

Taroncher-Oldenburg, G., Griner, E., Francis, C., and Ward, B. (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Appl Environ Microbiol* 69: 1159–1171.

Tiquia, S.M., Wu, L., Chong, S.C., Passovets, S., Xu, D., Xu, Y., and Zhou, J. (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. *Biotechniques* 36: 664–672.

de la Torre, J.R., Christianson, L.M., Béjà, O., Suzuki, M.T., Karl, D.M., Heidelberg, J., and DeLong, E.F. (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proc Natl Acad Sci USA* 100: 12830–12835.

Wang, D., Coscoy, L., Zylberberg, M., Avila, P.C., Boushey, H.A., Ganem, D., and DeRisi, J.L. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* 99: 15687–15692.

Wang, D., Urisman, A., Liu, Y.-T., Springer, M., Ksiazek, T.G., Erdman, D.D., *et al.* (2003) Viral Discovery and Sequence Recovery Using DNA Microarrays. *PLoS Biol* 1: 257–260.

Wilson, K.H., Wilson, W.J., Radosevich, J.L., DeSantis, T.Z., Viswanathan, V.S., Kuczmarski, T.A., and Andersen, G.L. (2002) High-density microarray of small-subunit ribosomal DNA probes. *Appl Environ Microbiol* 68: 2535–2541.

Wu, L., Thompson, D., Li, G., Hurt, R., Tiedje, J., and Zhou, J. (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl Environ Microbiol* 67: 5780–5790.

Wu, L., Thompson, D.K., Liu, X., Fields, M.W., Bagwell, C.E., Tiedje, J.M., and Zhou, J. (2004) Development and evaluation of microarray-based whole-genome hybridization for detection of microorganisms within the context of environmental applications. *Environ Sci Technol* 38: 6775–6782.

Zeidner, G., Preston, C.M., Delong, E.F., Massana, R., Post, A.F., Scanlan, D.J., and Béjà, O. (2003) Molecular diversity among marine picophytoplankton as revealed by psbA analyses. *Environ Microbiol* 5: 212–216.

Zhou, J. (2003) Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* 6: 288–294.

Zhu, J., Bozdech, Z., and DeRisi, J. (2003) *Array Oligo Selector.* [WWW document]. URL http://arrayoligosel.sourceforge.net/

Zinser, E.R., Coe, A., Johnson, Z.I., Martiny, A.C., Fuller, N.J., Scanlan, D.J., and Chisholm, S.W. (2006) *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl Environ Microbiol* 72: 723–732.

Supplementary material

The following supplementary material is available for this article online:

Fig. S1. Adjusting the stringency, *in silico*.

a. The Tukey Biweight-normalized signal of all probes sets on the array, from one representative hybridization series (data not comparable between series because of different salt concentrations and temperatures used) of the experiment described in Fig. 4, whose mean normalized signals are shown in Fig. 5. Note that the other genotypes present by the mean are still present by the more stringent Tukey Biweight.

b. Focusing just on the data for the MED4 probe set (red bar in panel a). Using Tukey Biweight, cross-hybridization to related strains is virtually eliminated.

Fig. S2. The pattern of hybridization across the *Prochlorococcus* MED4 probe set for the cell addition experiment to a natural seawater community. (a) Approximately 10^5 cells ml^{-1} of each strain MED4, MIT9515, MIT9312 and MIT9313; (b) ~10^5 cells ml^{-1} of MED4 and ~10^3 cells ml^{-1} of MIT9515, MIT9312 and MIT9313; (c) ~10^5 cells ml^{-1} of MED4; (d) ~10^5 cells ml^{-1} of MIT9515 and ~10^3 cells ml^{-1} of MED4, MIT9312 and MIT9313; (e) ~10^5 cells ml^{-1} of MIT9515; (f) ~10^5 cells ml^{-1} of MIT9312 and ~10^3 cells ml^{-1} of MED4, MIT9515 and MIT9313; (g) ~10^5 cells ml^{-1} of MIT9312; and (h) ~10^5 cells ml^{-1} of MIT9313.

Fig. S3. Testing a 'virtual' strain with a higher identity to target strain MED4, created by using the 17 probes (of 60 total) with BLAST-based identities higher than 90% to strain MIT9515. Their ANI to MIT9515 was 92.4%.

a. Across a range of cell concentrations, the mean signal from these higher-identity probes is intermediate to that of the whole probe set-based signal of MED4 and MIT9515. Also, see inset to Fig. 4b for the correlation between signal and genomic identity.

b. The Tukey Biweight signal across these probes is also intermediate between the whole-set signals for MED4 and MIT9515.

This material is available as part of the online article from http://www.blackwell-synergy.com

a.)

Legend:
- MIT 9313: 10^1 ⋯ 10^6
- MIT 9312: 10^1 ⋯ 10^6
- MIT 9515: 10^1 ⋯ 10^6
- MED4: 10^1 ⋯ 10^6

Mixes:
- Mix 1: 10^6 MIT9312, others 10^3
- Mix 2: 10^6 MIT9515, others 10^3
- Mix 3: 10^6 MED4, others 10^3
- Mix 4: all 10^6

Tukey Biweight Normalized Intensity

Genotype Targeted

4B7, ANT74A4, EB037F11, EB041B09, EB055B11, EB060D04, EB065D09, EB0_31A08, EB75002H05, EB75010A10, EB80L12H07, EF10057A08, HOT2C01, HOT4E07, ProMED4, Positive Control, Negative Control

b.)

Prochlorococcus MED4 probeset only:

Tukey Biweight Normalized Intensity

Pro MED4
Genotype Targeted

Supplementary Figure 1

71

a.) ~10⁵ cells/ml each of MED4, MIT9515, MIT9312 and MIT9313

b.) ~10⁵ cells/ml of MED4, and ~10³ cells of MIT9515, MIT9312 and MIT9313

c.) ~10⁵ cells/ml of MED4

d.) ~10⁵ cells/ml of MIT9515, and ~10³ cells of MED4, MIT9312 and MIT9313

e.) ~10⁵ cells/ml of MIT9515

f.) ~10⁵ cells/ml of MIT9312, and ~10³ cells of MED4, MIT9515 and MIT9313

g.) ~10⁵ cells/ml of MIT9312

h.) ~10⁵ cells/ml of MIT9313

Supplementary Figure 2

a.)

MIT 9313 { 10¹ ... 10⁶ }   MIT 9312 { 10¹ ... 10⁶ }   MIT 9515 { 10¹ ... 10⁶ }   MED4 { 10¹ ... 10⁶ }

Mean Normalized Intensity

MIT9515 data, subset of probes with higher identity

Pro MED4
Genotype Targeted

b.)

Tukey Biweight Normalized Intensity

MIT9515 data, subset of probes with higher identity

Pro MED4
Genotype Targeted

Supplementary Figure 3

Chapter 3

Time-series investigation of a coastal microbial community in Monterey Bay, CA, using the "genome proxy" microarray.

Authors: Virginia Rich, John Eppley, Vinh Pham, Yanmei Shi, Ed DeLong

Author Contributions:

Please note, these are likely to evolve by the time of publication – e.g., I anticipate my co-authors will contribute to the manuscript writing, but at this time my co-authors (except E.F.D) have not read it, and so I take full responsibility for any errors.

VIR collected ~15% of the samples, performed all DNA extractions and array hybridizations, developed the experimental design, did the clustering analyses and the marker prediction, and wrote the manuscript.

JE and VIR did the PCA and CDA analyses

VP did 16S alignments and phylogenetic tree construction, and submitted sequences

YS and VIR did the array-versus-pyrosequencing comparisons

EFD conceived of the experiment and funded the research.

Abstract

Coastal marine microbial communities are dynamic assemblages, inhabiting spatially and temporally variable environments. To gain improved temporal, spatial and phylogenetic resolution of the microbial communities in Monterey Bay, we used an expanded "genome proxy array" (an oligonucleotide microarray targeting marine microbial genome fragments and genomes) to profile a total of 57 samples over 4 years. Samples derived from 0m (photic), 30m (base of the surface mixed layer), and 200m (subphotic) habitats were hybridized to the array, along with a single depth profile from Hawaii for comparative purposes. The updated array, which targeted 268 genotypes (vs. 14 in the prototype), was cross-validated using pyrosequence data from three samples. The taxa abundances measured by the two methods were highly correlated (linear regression with $R^2=0.85\text{-}0.91$ for the three samples). The strongest differences among sample profiles were observed between the shallow (0m + 30m) and deep (200m) samples, with a number of depth-specific taxa distributions driving these differences. Depth-specific array profiles were also evident in the Hawaii samples, although the photic zone taxa present were different between the two locations. Although Monterey Bay is dominated by strong seasonal upwelling, the sample profiles within each depth did not cluster based on sample "oceanographic season" (*sensu* Pennington *et al.*, 2007). However, the abundance of the most dominant genotypes did correlate to strong episodic upwelling events. Genotypes representing common marine photo- and heterotroph clades, the majority of which are uncultivated, were observed in both shallow and deep samples, including the ubiquitous *Pelagibacter* clade, SAR86, OM42, OM43, NAC11-7, CHAB1-5, SAR116, SAR324, SAR406, OM60, ZD0417, Arctic96BD-19, and the G1 and G2 marine archaea. Most showed strong depth-specific distributions consistent with their previously-documented 16S-clone library and FISH-based distributions. Nutrient concentrations were strongly correlated to overall array profile variance, driven by the strong oceanographic

differentiation of the three sample depths, and finer-scale within-depth analyses linked several diverged array profiles to correlated nutrient profiles. The population structure of deeper taxa was more variable than that of shallow taxa, and sporadic taxa were more variable than common taxa. Specific population shifts were evident in several abundant target taxa, with populations in some cases clustering by depth or oceanographic season and in others apparently ecologically neutral for the sample designations examined. This multi-year community survey showed the consistent presence of a core group of common and abundant targeted taxa at each depth in this location, higher variability among shallow than deep samples, and episodic occurrences of other targeted marine genotypes.

Introduction

Marine microbial communities have garnered much attention in recent years, as major active participants in biogeochemical cycling (Arrigo, 2005, Howard et al., 2006, Karl et al., 2007), and due to novel metabolic discoveries (e.g. Béjà et al., 2000b, Dalsgaard et al., 2003, Kuypers et al., 2003, Kolber et al., 2000), and metagenomic surveys beyond the scale of those undertaken in other habitats (Venter et al., 2004, Tringe et al, 2005, DeLong et al., 2006 (Appendix 4), Kennedy et al., 2007, Rusch et al., 2007, Yooseph et al., 2007, Wegley et al, 2007, Wilhelm et al., 2007, Dinsdale et al., 2008 a and b, Mou et al., 2008, Neufeld et al, 2008, Marhaver et al., 2008). The marine realm makes up >99% of the available habitat on the planet, with its inhabitants comprising the bulk of the planet's biomass and diversity. In spite of this importance and growing attention, the marine microbial world remains incompletely understood due to the technical challenges of studying its vast diversity and habitat space. As with most complex biological systems, marine microbial systems cannot yet be modeled, in that their ecological and evolutionary units and defining interactions are not

known (although see the promising nascent attempts with cyanobacteria of Follows *et al.*, 2007). The dynamism of these communities remains poorly mapped; the majority of information derives from spatiotemporal snapshots, or from studies focusing solely on a few groups, often at higher phylogenetic resolutions which may not correspond to ecologically-relevant biological units.

Interest in developing a time series perspective on marine microbial systems has been growing, however, and methods have allowed increasingly comprehensive and fine-scale investigations. Several marine Long Term Ecological Research (LTER) sites have incorporated microbial investigations, leading to new insights into community structure over time, correlations to environmental parameters, and responses to change (e.g. Karner *et al.*, 2001, Morris *et al.*, 2005). In this LTER context, particularly noteworthy marine microbial time-series investigations have occurred (although many at relatively coarse phylogenetic resolution), at the Hawai'i Ocean Time-Series (HOT) (Karner *et al.*, 2001, Campbell *et al.*, 1997), the Bermuda Atlantic Times Series (BATS), (Steinberg *et al.*, 2001, DuRand *et al.*, 2001, Morris *et al.*, 2005, McGillicuddy *et al.*, 2007), the San Pedro Ocean Time-Series (SPOT) (Fuhrman *et al.*, 2006), and the Monterey Bay Microbial Observatory (MBMO) within the Monterey Bay National Marine Sanctuary (Ward, 2005, O'Mullan and Ward, 2005, Mincer *et al.*, 2007).

A number of methods exist, each with strengths and weaknesses, for tracking microbial community members (see Chapter 1). Community genomic sequencing may be the optimal tool for exploring community composition because of its high information yield, but for now remains financially unfeasible for sampling-intensive investigations. We previously described the "genome proxy" array (Rich, Konstantinidis and DeLong, 2008) which used sets of 70-mer probes to target 14 genotypes (genome fragments and genomes). The array was designed to cross-hybridize to related genotypes at ≥ ~80% average nucleotide identity (ANI, as in Konstantinidis and Tiedje, 2005), which could be raised to ≥

~90% ANI by tuning the analysis *in silico*. In addition, related cross-hybridizing strains produced distinct hybridization patterns across their target probe set, which could reveal shifts in population structure across samples.

Here, we developed an expanded genome proxy array, and applied it to investigate the time series dynamics of the 268 targeted clades over a four-year period at Monterey Bay Station M1 (36.747° N, 122.022° W), a well-studied coastal environment characterized by strong seasonal upwelling. Photic (0m) and subphotic (200m) samples from 24 time points spanning ~4 years, and samples just below the mixed layer (30m) from 13 time points over ~1.5 years, were hybridized to the array. Array data were cross-validated by comparison to pyrosequencing data for three 0m samples. The array-based organism profiles for 57 samples were used to investigate: (i) genotype differences with depth, (ii) genotype differences between Monterey Bay's "oceanographic seasons" (*sensu* Pennington *et al.*, 2007) (iii) genotype differences associated with episodic upwelling events, (iv) correlations between hybridization profiles and nutrient concentrations (nitrate, nitrite, phosphate, silicate), (v) and correlations in the distribution of genotypes to one another.

Methods

Sampling and DNA Extractions

Samples were collected from Station M1 (36.747° N, 122.022° W) in Monterey Bay periodically (at approximately monthly intervals, with several longer gaps) between Julian Day (JD) 271 in 2000 and JD167 in 2004. 2L of seawater from each of eight depths (0, 20, 30, 40, 80, 100, 150 and 200m) were filtered through a 45mm GF-A prefilter (Whatman) and concentrated onto a 25mm Supor-200 0.2μm filter (Pall Corp, Ann Arbor, MI), using a MasterFlex peristaltic pump system (Cole-Parmer Instrument Company, Vernon Hills, IL).

Filters were stored dry in 2ml screw-cap tubes, immediately placed in a -20 degree Celsius freezer shipboard, and transferred on ice to a -80 degree Celsius freezer upon landfall.

All MB DNA extractions were performed simultaneously in 96-well format to minimize extraction variability, as in Rich, Konstantinidis and DeLong, 2008. DNA was extracted from all 0m and 200m filters available from 2000 JD271 through 2004 JD167, and all 30m samples available from 2000 JD271 through 2002 JD070. In this location, 0m is in the photic zone, 30m is generally below the mixed layer, and 200m is below the photic zone. Extracted DNAs were quantified spectrophotometrically (Nanodrop, Thermo Scientific) and stored at -80 degree Celsius until use. Yields averaged ~470 ng per liter of seawater for 200m samples (range 177-903 ng) and ~1460 ng per liter of seawater for 0m and 30m samples (range 484-3804 ng).

In addition to Monterey Bay samples, several community DNAs from the Hawaii Ocean Time series Station ALOHA were hybridized to the array. These samples were collected on cruise HOT179 in March of 2006 as described in Frias-Lopez and Shi *et al.* (2008), and include the 75m DNA sample used in that study. DNA was extracted as described in Frias-Lopez and Shi *et al.* (2008).

Oceanographic Data

Oceanographic data were kindly provided by Reiko Michisaki and Francisco Chavez of the Biological Oceanography Group at the Monterey Bay Aquarium Research Institute, who collected and processed it as part of the Monterey Bay time series program. Measurement methods were described in Asanuma *et al.*, 1999.

Arrays Design, Hybridization, and Data Processing

The expanded genome proxy array was designed as in Rich, Konstantinidis and DeLong, 2008, with a broader scope (268 target genotypes, as opposed to

the prototype's 14) and the addition of a co-spot oligo for spot alignment and gridding purposes (the "alien" sequence used in Urisman *et al.*, 2005). The targets were selected from fully-sequenced marine microbial genomes, publicly-available marine-derived BAC and fosmid clone sequences, and fully-sequenced clones from the lab's Monterey Bay and Hawai'i environmental BAC- and fosmid-based genomic libraries. Targeted genotypes are detailed in Table 1, summarized in Table 2, and presented in a schematic phylogenetic overview in Figure 1. Previously-unpublished sequences used for array design were submitted to Genbank under accession numbers XXX-XXX.

For each sample, at least three replicate arrays were hybridized. For samples in which one or more of the arrays showed significant surface peeling or excessive background fluorescence, additional arrays were hybridized. Hybridizations were performed as in Rich, Konstantinidis and DeLong, 2008, with the following modifications: Round A, B and C reactions were performed in 96 well plates for higher throughput, and cleaned through ExcelaPure 96-well plates (Edge Biosystems, Gaithersburg). 1 pmol of Cy5-labeled co-spot complement oligo was added to each hybridization for spot localization purposes (modified from Urisman *et al.*, 2005).

Data were pre-processed as in Rich, Konstantinidis and DeLong, 2008, with minor modifications. Briefly, poorly-performing arrays, defined as those with less than half the positive control probes brighter than the standard deviation of the negative control probes, were removed from further analysis. Within each remaining array, bad spots (those with areas of poly-L-lysine peeling or excessive background fluorescence) were manually flagged and removed from further analysis. Background-subtracted spot intensities were negative-control-subtracted and normalized to each array's mean positive control value, then replicate spots of a given probe were pooled across arrays and the median was taken as the value for that probe. For each organism, the mean or tukey biweight (TBW) across each probe set was taken, as in Rich, Konstantinidis and DeLong,

2008, with an improvement in the subsequent thresholding step for each organism, as follows. At least 40% of each organism's probes were required to be above the standard deviation of the negative control probe set (rather than above twice the mean negative control value, as previously), or else the organism was considered "absent" and its value set to zero. This was done to remove erroneous organism abundances due to uninformative single-gene cross-hybridizations.

Array platform design and hybridization data were deposited in the Gene Expression Omnibus, under GEO Accession numbers XXX and XXX-XXX, respectively.

Data Analyses

Clustering analyses of sample hybridization data were performed in GenePattern (Reich *et al.*, 2006), using hierarchical clustering (Eisen *et al.*, 1998) by Pearson correlations for both rows and columns, using pairwise complete-linkage, and without row or column centering. Marker Prediction was performed in GenePattern. Principal component analyses (PCA) was performed in both GenePattern and in R using the prcomp function. Canonical discriminant analyses (CDA) were performed in R with the candisc function. In order to keep the number of variables less than the number of responses (i.e., samples), CDA was performed using the top 28 principal components instead of all detected organisms. Correlations were calculated between environmental parameters or organism abundances and each plotted principal component or canonical discriminant axis. The relative values of the correlations were represented as vectors on the analysis graphs.

Array-vs-pyrosequencing Comparisons

Three samples were chosen for parallel pyrosequencing and array hybridization, based on their DNA yield. Approximately 3μg each of samples

82

2000 JD298, 2001 JD115 and 2001 JD135 were sequenced at the Schuster Lab pyrosequencing facility (Penn State University) on a 454 sequencer.

Sequence Clean-Up: To remove poor quality sequences, the length distribution of the raw pyrosequencing reads for each sample was plotted.   From the empirical cumulative density function (ecdf) plot, the lower and upper boundary lengths were estimated so that 95% of the read lengths fell between the boundaries (which varied for each sample: 71 and 305bp for 2000JD298, 65 and 255bp for 2001JD115, and 65 and 303 bp for 2001JD135). The outlying 5% of the reads were removed. Furthermore, reads with more than one "N" were also removed. This two-step process removed approximately 5.5% of the reads overall; for 2000JD298, 23917 out of 419684 reads (5.7%) were discarded, for 2001JD115, 19822 out of 365472 reads (5.4%) were discarded, and for 2001JD135, 22887 out of 414861 reads (5.5%) were discarded.

BLASTN parameters: To identify BLASTN parameters that would give the closest *in silico* similarity to the array's range of cross-hybridization, we used the genomes of *Prochlorococcus* MED4, MIT9515, and MIT9312, whose relative hybridization strength to the array's strain MED4 probes was measured previously (Rich, Konstantinidis and DeLong, 2008). The genomes were fragmented into overlapping (tiled) 100-bp fragments using a perl script (kindly provided by G. Tyson), and each set of fragments was BLASTed against the MED4 genome to compare, for varying parameters, the self-self results (MED4 to MED4, 100% identity), MIT9515 to MED4 (86% average genomic identity, calculated as in Konstantinidis and Tiedje, 2005), and MIT9312 to MED4 (78.5% average genomic identity). The following combinations of command-line BLASTN parameters were tested: 1)X150 q-1 r1 W7 FF, 2)X30 q-3 r1 W7 FF, 3)X30 q-5 r1 W7 FF, 4)X30 q-5 r2 W7 FF, and 5)X30q-7r2W7FF, among which the first parameter set yielded the best separation of MED4-MIT9515 and MED4-MIT9312 distribution of hits, and was subsequently used in downstream analyses.

Parsing parameters: BLASTN hits to a given target were parsed by bit score. However, because pyrosequencing reads range in lengths, and read length effects bit score, we investigated the correlation between read length and bit score for MIT9515 fragments versus MED4 and for MIT9312 fragments versus MED4. In addition to tiled 100-bp fragments, tiled 50-bp, 75-bp, and 125-bp fragments were also generated. Linear equations for bit-score (y-axis) versus read length (x-axis) were determined. The MED4-MIT9312 slope was smaller than that of MED4-MIT9515, due to the lower average identity involved at any given read length. Since cross-hybridization at or above the MIT9515-MED4 level of identity dominates the signal of the microarray (Rich, Konstantinidis and DeLong, 2008), the equation for that comparison was used to adjust the bit score cutoff to the read length for each individual read.

Monterey Bay pyrosequencing versus array comparison: Using the BLASTN parameters and parsing criteria optimized above, the pyrosequencing reads from each sample were BLASTed against all 268 genomes and genome fragments to which the array was targeted. Reads were assigned to (a.k.a. recruited to) one or more array targets, proportional to their bitscore, to mimic the cross-hybridization permitted by the array. Thus, if 1 read matched three targets using the criteria outlined above, then it would be assigned to the first of those targets as 1 * (bitscore1 / (bitscore1 + bitscore2 + bitscore3)), to the second as 1* (bitscore2 / (bitscore1 + bitscore2 + bitscore3)), etc.. The read-based abundance of each array target was then normalized to the length of the target query, and to the database size, and compared to the unthresholded array signal (that is, the signal for each organism before requiring at least 40% of its probes to be above the described threshold) of the same clone.

Results

Development of the Expanded Genome Proxy Array

The expanded genome proxy array targeted 268 organisms, through suites of probes (n=20 per target, in general) dispersed along genomes and genome fragments derived from marine habitats. Targeted organisms were selected to span known marine microbial diversity (Figure 1, Tables 1 and 2, and Figures S1-S5). For particularly diverse and abundant clades (e.g. the marine cyanobacteria *Prochlorococcus* and *Synechococcus*), representatives were chosen where possible from each major pelagic coastal and open-ocean lineage. Of the 268 organisms represented on the array, 42.5% were clones derived from HOT, 26.5% were marine microbial genomes isolated from a variety of locations, 19% were clones derived from Monterey Bay, and 11.9% were other marine-derived clones available in Genbank (Figure S6a).

Ground-truthing the array

To rigorously evaluate our new expanded genome proxy array, we sought to compare pyrosequencing and array data for each of three Monterey Bay samples (0m from 2000 JD298, 2001 JD115 and 2001 JD135). Sequencing produced an average of 400,000 reads per sample, which were trimmed to remove poor quality sequence (~5.5% of reads), then "hybridized" *in silico* using BLAST (Altschul, 1990) to genotypes targeted by the array. BLAST parameters were trained using genomes of *Prochlorococcus* strains whose relative cross-hybridization to the array had been previously investigated (Rich, Konstantinidis and DeLong, 2008), in order to simulate the amount of target divergence tolerated by the array. The sampling depth of the pyrosequencing data was insufficiently deep to meaningfully examine the evenness of BLAST hits to each target (that is, their distribution across the target sequence), whereas such filtering is performed during array data analysis (by requiring >40% of a target probe set to show above-threshold signal to consider that target "present"). Therefore, unfiltered array data were compared to pyrosequencing data for each sample.

The normalized pyrosequencing read recruitment was strongly correlated to the normalized unfiltered mean array intensity (Figure 2; linear regression with R^2 of 0.91 for sample 2000 JD298, 0.88 for 2001 JD115, and 0.85 for 2001 JD135). Such strong correlation between relatively unbiased and comprehensive pyrosequencing, and the high-throughput, inexpensive genome proxy array, supports the array's utility as a tool for profiling studies requiring high sample throughput.

Exploring microbial communities using the genome proxy array

Target derivation versus presence: 57 samples from Monterey Bay (location and relative times and depths of samples indicated in Figure 3b), and 4 samples from Hawaii were hybridized to the array. Targeted coastal genotypes were enriched in Monterey Bay samples, and targeted open ocean genotypes were enriched in Hawaii samples. That is, 74.1% of all target genotype signal in 57 MB samples were from MB-derived clones (Figure S6b), a ~2.5 fold enrichment relative to their representation on the array. Alternately, 59.3% of all target signal in 4 Hawaii samples were from Hawaii-derived targets, 1.39-fold their representation on the array (Figure S6c).

Shallow versus deep genotypes: Hierarchical clustering was used to investigate community depth partitioning. By Pearson correlation-based clustering of the Monterey Bay samples, all 200m (sub-photic zone) samples clustered together to the exclusion of the shallower samples (0m photic zone, and 30m below the mixed layer sample; Figure 4a). Likewise, among four Hawaii samples, hierarchical clustering followed depth (Figure 4b).

Principal component analysis (PCA) of the Monterey Bay hybridization profiles also supported a clear separation of shallow and deep samples (Figure 5), with a slight additional separation of the 0m and 30m samples. The first two principal components account for >90% of the data's variability, and clearly delineate the shallow and deep clusters, recapitulating previously-observed

microbial community stratification along the depth gradient.

The majority of targeted taxa showed differential distributions between shallow and deep samples, in both Monterey Bay (0 and 30m versus 200m; Figure 4a) and Hawaii (25, 75 and 125m versus 500m; Figure 4b) samples. In Hawaii, 500m taxa never occurred in the shallow sample, and vice versa. In the much more extensive Monterey Bay dataset, there were three notable target clusters with particularly strong depth-specific signals (red-dashed boxes in Figure 4a). The first cluster comprised 8 target genotypes that were abundant and consistently present in shallow samples, and spanned a range of phylogenetic clades. This cluster is hereafter referred to as "*shallow-consistent*", and included EB000_31A08, EB000_45B06, alpha_HTCC2255, EB000_39F01, EB000_55B11, EB080_L11F12, EB080_L43F08, and EB080_L27A02. A second cluster of shallow genotypes, "*shallow-frequent*", encompassed 12 frequently-occurring targets: EB000_37F11, EB080_L06A09, EB000_36A07, EB000_46D07, EB000_69G07, EB000_39H12, EB000_49D07, EBAC_27G05, EB000_50A10, HF0010_16H03, *Pelagibacter* HTCC1002 and HTCC1062. The third cluster, "*deep-consistent*", represented 10 taxa with a consistent presence and abundance in the 200m samples: EB000_36F02, DeepAnt_EC39, EB750_10B11, EB750_10A10, HF4000_23L14, EB080_L31E9, EB080_L93H08, EF100_57A08, EB750_01B07, and HF4000_08N17.

Canonical discriminant analysis was used to further examine genotype distributions with depth. Each genotype abundance was correlated to the first two canonical discriminant axes, with the resulting vector length a measure of that genotype's influence on sample variability (Figure 6a). By this analysis, the targets which most drove the separation of the deep from the shallow samples were EB750_01B07, EB750_10B11 EB080_L31E09, and HF4000_08N17, a subset of the deep-specific organisms discerned in the above clustering analysis. For 0m and 30m, the picture was more complex, and included taxa not identified as dominant signals in the clustering analysis. EB080_L43F08, EB000_39F01,

ProMED4, EB080_L27A02, and alpha_HTCC2255 drove the differentiation of 0m from 30m, while EB000_39H12, EBAC_27G05 and EB000_65A11 drove differentiation of 30m from 0m.

Environmental Parameters: We investigated the correlation between clustering patterns observed using the array and environmental parameters, in two ways. First, each sample was assigned to its "oceanographic season", a designation based on average annual upwelling patterns in Monterey Bay (spring/summer, fall, or winter, described in e.g. Pennington *et al.*, 2007) and these designations were compared to the samples' clustering patterns (Figure 4a).

Second, canonical discriminant analysis was used to examine the correlation between individual nutrient (phosphate, nitrate, nitrite and silicate) concentrations and sample variability (Figure 6b). Here, strong correlations were apparent to each nutrient, reflecting large differences in nutrient conditions at the three depths. Phosphate, nitrate and silicate drove the differentiation of the shallow from the deep samples, while nitrite drove the separation of 30m from 0m.

Since possible correlations at each depth were obscured by the strength of the nutrient signals between depths, samples from each depth were also plotted in separate principal component analyses, and the correlations of each nutrient's variability to the first two principal component axes were calculated (Figure 7). (Principal component analysis was used for this instead of canonical discriminant analysis because whereas with c.d.a. the distance between all defined groups is maximized, in p.c.a. the total variability among all samples is maximized, and we chose not to define subgroups within each depth.) Variation in nutrient concentrations among samples accounted for little of the variability among 0m samples (Figure 7a), with a minor correlation of nitrite. At 30m (Figure 7b), however, nutrient variability correlated relatively strongly to the principal

component axes, with a strong signal of phosphate, nitrate and silicate and a slightly weaker and inverse signal for nitrite. Finally, at 200m (Figure 7c), nitrate and nitrite showed no and weak correlations, respectively, while silicate and phosphate gave equally strong but non-overlapping correlations.

Population variations: Population shifts over time were examined in two ways. First, each target's mean intensity was compared to its tukey biweight intensity within each sample. There was a larger drop of TBW relative to mean for sporadically distributed taxa compared to depth-consistent taxa, and also for common deep taxa compared to common shallow taxa (Figure 8). Second, for particular targets of interest, the pattern of signal across the probe set was compared between samples, and the pair-wise Pearson correlation of these patterns was calculated. Clustering analysis of the Pearson correlations between samples was then used to reveal samples with more and less similar probeset patterns for a given genotype. For the SAR86-II target EB000_45B06, this process is shown in Figure 9.

Discussion

Over the ~4-year sampling period at Station M1 in Monterey Bay, a significant portion of the expanded genome proxy array's targets showed signal (95 out of 268 targets, ~35%, were present in one or more samples). The majority of targets detected by array were uncultivated marine lineages, many of which derived from the environment of study (Figure S6). Broadly, there were three major patterns of target occurrence across the 57 samples hybridized. Some taxa were consistently abundant in most or all samples of a given depth, other taxa were frequently present within their primary depth of occurrence, and many taxa had sporadic distributions in one or more depths.

The genome proxy array platform was previously validated using related

89

target strains added into natural marine community samples at a range of concentrations (Rich, Konstantinidis & DeLong 2008). In this study we further validated the results of the expanded platform by comparing its data to community genomic pyrosequencing data, for three surface samples. This represented a full methodological comparison, encompassing the array's potential biases in both the amplification and labeling steps and the hybridization itself; pyrosequenced DNA was not subjected to the same amplification-and-labeling protocol as the aliquots used for array hybridization. Overall there was strong correlation between taxa abundance measured by mean array intensity and by BLAST-based recruitment of pyrosequences to targeted genomes and genome fragments, with linear regression R^2 values of 0.85-0.91 for the three samples.

In addition, the pyrosequence data indicated what percentage of the community could be surveyed by the array, i.e. what percent of the community was represented by the targets on the array. Based on the number of pyrosequence reads recruited to the array target sequences at the relatively high stringency used to mimic the array hybridization, the array captured 1.9%-2.5% of the total reads in these three samples (7636/395767 for 0m_2000_298, 8743/345650 for 0m_2001_115, and 9252/39197 for 0m_2001_135). A recent analysis of a similarly-obtained marine pyrosequence dataset showed only 50% of reads had identity to any Genbank sequences (Frias-Lopez and Shi *et al.*, 2008), using less stringent criteria. Furthermore, the ten targets with the highest number of recruited reads in each sample accounted for from ~66% to 75% of the total reads. In all three cases, 9 of the top 10 targets were environmental genomic clones, with the tenth being a recently-sequenced genome from the NAC11-7 clade of the *Roseobacteria*. Together with the relative decrease in marine genome observations versus presence on the array in both Monterey Bay and Hawaii samples, these suggest that "native", uncultivated DNA sequences are most effective for investigating marine microbial communities, and that by

being designed from such sequences, the array can provide a useful complement to other means of community investigation.

After cross-validating the expanded genome proxy array with pyrosequence data, we investigated the depth-specific distributions of targets from particular phylogenetic groups, across the Monterey Bay samples, focusing on taxa that occurred in multiple samples. One of the most highly represented groups was Roseobacter, which are known to comprise up to 20% of cells in coastal samples (reviewed in Buchan *et al.*, 2005), are ecologically diverse, and include both cultivated and uncultivated lineages. Roseobacteria have been described previously as abundant in Monterey Bay, accounting for 20-40% of total bacterial SSU DNA in the mid-bay region during an upwelling event (Suzuki *et al.*, 2001). In large-insert genomic libraries from this site, the NAC11-7 and CHAB-I-5 clades accounted for ~22% and ~6%, respectively, of the SSU operon-containing clones of both the 0m and 80m libraries, representing ~65% of the total *Roseobacter* signal in each (Suzuki *et al.*, 2004). The array abundance of *Roseobacter* targets agrees with previous estimations of their abundance (Figures 4a and S7a). A significant number (3 of 8) of taxa in the *shallow-consistent* cluster were NAC11-7 clones (EB080_L11F12, EB080_L43F08, EB080_L27A02) as were 2 of 12 *shallow-frequent* targets (EB080_L06A09 and the NAC11-7 genome Rhodobacterales HTCC2255). Overall, NAC11-7 represented 25% of the targeted taxa that commonly occurred (*frequent* or *consistent* clusters) in shallow samples. Lastly, 1 of 10 *deep-consistent* taxa was a CHAB-I-5 clone (EB000_36F02). In addition to their high surface abundances generally, the differential distributions of three of the *Roseobacter* NAC11-7 targets (EB080_L27A02, EB080_L43F08, and HTCC2255) between 0m and 30m samples helped drive the differentiation of these samples (Figure 6a).

Members of the uncultivated gammaproteobacterial SAR86 clade were also abundant in shallow samples. SAR86 has been commonly reported in marine samples (Eilers *et al.*, 2000, Rappe *et al.*, 2000, Suzuki *et al.*, 2001, Venter *et al.*,

2004, Morris *et al.*, 2006), is known to partition with depth (Morris *et al.*, 2006), and can comprise up to 10% of the cells in a community (Mullins *et al.*, 1995, Eliers *et al.*, 2000, Morris *et al.*, 2006). Furthermore, it has been previously described as abundant in Monterey Bay, as 3-6% of total bacterial SSU DNAs in the Bay during an upwelling event (Suzuki *et al.*, 2001), and as 5.6%, 5.5%, and 1.6% of the SSU operon-containing clones in 0m, 80m and 100m large-insert clone libraries from this location (Suzuki *et al.*, 2004). Array-based sample profiling recapitulated this importance (Figures 4a and S7b), as 2 of 8 *shallow-consistent* taxa were SAR86-II clones (EB000_31A08 and EB000_45B06), and a SAR86-III clone (EBAC_27G05) was among the 12 *frequent-shallow* taxa. All three clones possess proteorhodopsin (PR) genes, and PR-containing SAR86 types have been hypothesized to be photoheterotrophs (Beja *et al.*, 2000, Sabehi *et al.*, 2004, Sabehi *et al.*, 2005, Mou *et al.*, 2007, Sabehi *et al.*, 2007). The distribution of the SAR86-III clone also helped drive the differentiation of 30m samples from those at 0m (Figure 6a).

The alphaproteobacterial SAR11 clade is one of the most abundant in the world's oceans (e.g. Morris *et al.*, 2002) and was isolated from coastal waters approximately 700 miles north of the study area (Rappé *et al.*, 2002). Seven of the 10 targeted SAR11 genotypes were present in ≥ 1 Monterey Bay sample, and each showed depth-specific distribution (Figures 4a and S7c). Pelagibacter HTCC1062 and HTCC1002, cultivated strains both in the SAR11 subgroup 1a, were present only in shallow samples and occurred frequently but not consistently. Several other SAR11 genotypes were present only in deep samples, and occurred frequently or sporadically (HF4000_37C10, HF4000(384)_009C18, HF0770_37D02, EBAC750_11E01, and EB750_09G06). This is consistent with the known depth distributions of the two major SAR11 clades (e.g. Stingl *et al.*, 2007). Furthermore, the distribution of HTCC1062 and HTCC1002 showed no correlation to upwelling season, consistent with previous observations that their numbers do not change under phytoplankton bloom

conditions (Morris *et al.*, 2005).

Proteorhodopsin- (PR)- containing targets produced strong array signals throughout the shallow samples. In addition to the SAR86 clones, a number of PR-containing targets without phylogenetic markers were among the *shallow-consistent* (3 clones) and *shallow-frequent* clusters (4 clones). These targets were designated as various *Proteobacteria* based on BLAST-based identities. In total, targets known to carry the proteorhodopsin gene accounted for 50% of the taxa abundant in shallow samples (5 of 12 *shallow-frequent* and 6 of 10 *shallow-consistent* taxa). Two of these PR-containing clones had sufficiently inverted relative abundances at 0m and 30m to contribute to the differentiation of the two depths (Figure 6a; EB000_39F01 in 0m, and EB000_39H12 in 30m). These observations are in agreement with the increasing awareness of high proteorhodopsin gene abundances in photic zones (Béjà *et al.*, 2000, Sabehi *et al.*, 2004, McCarren *et al.*, 2007, Rusch *et al.*, 2007) and of the emerging suggestions of PR-based photoheterotrophs as abundant components of photic communities (Sabehi *et al.*, 2005, Stingl *et al.*, 2007, Gómez-Consarnau *et al.*, 2007, Moran and Miller, 2007, González *et al.*, 2008).

One of the other *shallow-frequent* taxa was a representative of the OM43 clade (target EB000_36A07), which has been observed to respond to diatom blooms (Morris *et al.*, 2006). These blooms occur in MB during the first upwelling season (e.g. Pennington *et al.*, 2007). In our MB samples, a general correlation between this clone's occurrence and upwelling season was not observed. However, during specific post-bloom samples with particularly high array intensities (see below), this OM43 target was among the small number of targets with the most dramatic increases in intensity.

The final bacterial target within the *shallow-frequent* cluster was a SAR116-I clone (EB000_46D07). Of 12 SAR116 targets, two originated in Monterey Bay, and were the only ones detected. The SAR116-II target (EB00_37G09) was

present only twice, in 0m samples, while the SAR116-I clone was present in 21 of 34 shallow samples. In large-insert environmental libraries from this site, SAR116 comprised 11.3%, 1.4%, and 0.8% of the SSU operon-containing clones in 0m, 80m, and 100m libraries, respectively (Suzuki *et al.*, 2004). This *Rhodospirilalles* clade has broad global distribution and frequently high abundances (e.g. Giovannoni and Rappé, 2000, DeLong *et al.*, 2006, Rusch *et al.*, 2007), but has only recently been isolated (Stingl *et al.*, 2007). Due to the phylogenetic diversity of this clade (at least 10% divergent 16S rRNA, Stingl *et al.*, 2007), it may be that the relative specificity of the array platform prohibited it from tracking other native SAR116 strains; that is, that the other SAR116 targets on the array did not share sufficient identity (i.e. ~<80% ANI) with local populations to produce array signal. An alternative explanation is that the previously-constructed 0m large-insert library captured an unusual bloom of SAR116 at this location, and that they are not normally present at ~10% of surface populations. However, a bloom scenario seems unlikely, because the captured SAR116 rRNA genes in the three libraries spanned the breadth of SAR116 diversity. The array results suggest that additional sequencing of previously-captured SAR116 clones from this location may be appropriate, to further and best represent native populations. To identify which SAR116 clone(s) would be optimal, the surface pyrosequence databases can be queried with this clade's rRNA sequences.

Three marine archaeal targets were among the most abundant targeted taxa in the MB samples. Furthermore, of 15 total archaeal genotypes targeted by the array, 7 were present in at least one MB sample. Typically, marine euryarchaea are seen in low numbers in the water column while marine crenarchaea increase with depth and can account for a significant proportion of the total microbial community in deeper waters (Massana *et al.*, 1997, Karner *et al.*, 2001, Pernthaler *et al.*, 2002). In Monterey Bay, pelagic crenarchaeal cells have also been shown to increase with depth and to represent up to 33% of the

200m community, while euryarchaeal cells were more abundant in shallow samples (up to 12% of the community in the summer, but less than 1% throughout the water column in winter, Pernthaler *et al*, 2002, and frequently below FISH-based detection at 200m, Mincer *et al.*, 2007). These trends were generally reflected in the array data. Four euryarchaeal clones were present in the water column. One (EB000_37F11) was in the *frequent-shallow* clade, two of which were in the abundant *deep-consistent* cluster (DeepAnt_EC39 and EF10057A08), and the last of which (DeepAnt_JyKC7) occurred in only a single 200m samples. The overall frequency of these archaeal targets suggests a consistent presence of these taxa at this location, throughout the water column. Two crenarchaeal targets were present in these Monterey Bay samples, and both were restricted to 200m samples. One (ANT74A4) had only sporadic occurrence, while the other occurred quite frequently (4B7, in 13 of 23 200m samples). Finally, a putatively-archaeal target of unknown identity (EB750_01A01) was sporadically present in the deep samples. The presence of euryarchaeal clones in both shallow and deep samples, and the restriction of crenarchaeal clones to the deepest samples, reflected the general trends seen previously.

It is notable that two euryarchaeal clones were among the most abundant taxa at 200m across all sampling dates, given the clade's previously documented maxima near the surface and their low numbers or absence in some studies of deep waters at this location. However, previous FISH-based studies used surface rather than deep euryarchaeal phylotypes to generate probes, and other studies using rRNA clone libraries have noted appreciable euryarchaeal abundances in deep waters (López-García *et al.*, 2001, Massana *et al.*, 1997). This observation highlights the challenge in cross-comparing techniques with different levels of phylogenetic specificity. While previous FISH-based investigations targeted broad phylogenetic groupings, the genome-proxy array targeted specific genotypes. One of the array's two deep-abundant clones

originated from 100m in Monterey Bay, while the other came from 500m in the Antarctic Polar Front (López-García *et al.*, 2004). The array results can only describe the taxa targeted and cannot be generalized to the clades within which those taxa occur, and thus adding additional archaeal targets to the array will be important in expanding the breadth of archaeal investigations to better span known marine diversity.

Previous work at this location also strongly suggested that *Crenarchaea* play a significant role in ammonia oxidation, and mapped their initial appearance in the water column to the nitracline (Mincer *et al.*, 2007). In addition, co-occurrence of *Crenarchaea* and putatively nitrite-oxidizing *Nitrospina* species indicated a possible metabolic link between these two groups at this location (Mincer *et al.*, 2007). qPCR analyses of the relative ratios of crenarchaeal SSU rRNA and *amoA* genes in four depth profiles showed a 1:1 correlation throughout the water column, an increase with depth, and maxima at 200m in three of the four profiles. In addition to the concordance of the array-based abundance of the two crenarchaeal targets, the array included a *Nitrospina* target, and their distributions also agreed with the previous observations. The *Nitrospina* clone (EB080_L20F04) was apparent sporadically in a single 30m sample and in 200m samples, a subset of those in which the two crenarchaeal targets were present (5 of 13 200m samples), and at lower signal intensities. Previous qPCR surveys showed *Nitrospina* SSU rRNA to parallel the distribution of crenarchaeal SSU rRNA but with lower abundances (Mincer *et al.*, 2007). Interestingly, the array also targeted a betaproteobacterial ammonia-oxidizer *Nitrosomonas* clone (EB080_L12H07) captured from 80m in Monterey Bay, and its distribution was similar to that of the *Nitrospina* clone. qPCR surveys of *Nitrosomonas*-like *amoA* sequences at four sampling dates produced very low counts throughout the water column, reinforcing the sporadic nature of this taxon's presence in Monterey Bay.

Returning to the depth-specific clustering of taxa in the array profiles,

additional *deep-consistent* genotypes included four pelagic relatives of deep-sea invertebrate (e.g. vesicomyid clam) symbionts. Two such 16S-containing clones, a 4000m HOT-derived clone (HF4000_23L14), related to ZD0405 (a pelagic 16S-clone related to symbionts), and an 80m MB-derived clone (EB080_L31E09), were consistently abundant at 200m, with the latter being the most abundant targeted genotype at this depth. In addition, two rubisco-containing clones (EB750_10B11, EB750_10A10) without phylogenetic markers, whose BLAST homology indicated possible relatedness to symbionts, were also abundant in the deep. Furthermore, two of these four clones were important in driving the differentiation of 200m samples from shallower ones (Figure 5). It is not uncommon for such pelagic relatives of symbiont species to be found in marine 16S surveys (e.g. Suzuki *et al.*, 2001, Lopez-Garcia *et al.*, 2001, Bano and Hollibaugh, 2002, Zubkov *et al.*, 2002, Klepac-Ceraj, 2004, thesis). Given the availability of large genomic fragments from these symbiont-related organisms, investigation of their potential lifestyle is possible. In this way, there is a feedback between array data and other methods, as distribution information seen with the array can extend metagenomic snapshots, particularly for groups of emerging interest like the symbiont-relatives which have not been studied in depth with more focused methods such as qPCR. In addition, array-based evidence for different populations can motivate the exploration of particular hypotheses within metagenomic data.

Two deltaproteobacterial clones (EB750_01B07, HF4000_08N17 - the latter within the SAR324 clade) were also within the *deep-consistent* cluster, consistent with the previous depth preference described for this group (e.g. Wright *et al.*, 1997). These targets were also highly correlated to the differentiation of 200m from 0m and 30m samples. Finally, the remaining *deep-consistent* target was the gammaproteobacterial clone EB080_L93H08, which clustered together with deep-sea environmental clones from around the world, notably ZD0417 and DHB-2 (Lopez-Garcia *et al.*, 2001), although the natural history of this clade

remains a mystery. The array data demonstrating the consistency of this clade's presence in 200m Monterey waters, combined with its common occurrence in 16S rRNA-gene clone surveys in a variety of locations, suggest it warrants further study. As this clade remains uncultivated, genome fragments provide an important window into its potential lifestyle, and array profiling of its abundance at other sites over time could help define its habitat.

Interestingly, targeted cyanobacteria did not show strong or consistent array signal in Monterey Bay. However, the episodic surface appearance of *Prochlorococcus* MED4 helped differentiate 0m from 30m samples (Figure 5). Also, the use of a 1.6µm pre-filter during sample collection likely excluded larger *Synechococcus* cells.

The sole Hawaii depth profile showed markedly different taxa abundances than Monterey samples, although it retained strong depth-based clustering (Figure 4b). When clustered together with Monterey Bay samples, the Hawaii 500m sample was more like 200m Monterey samples, although it was basal to that cluster, while the shallower three Hawaii samples formed their own cluster separate from all Monterey samples (Figure 4c). No taxa in the 500m HOT179 sample appeared in the shallower three samples, and vice versa. A notable difference in shallow Hawaii taxa compared to Monterey taxa were the cyanobacteria, with a general lack thereof in Monterey Bay, while *Prochlorococcus* strains 9312 and ASC9601 were the most abundant signals at all three shallow depths. The dominance of these clades was consistent with previous metagenomic work at this location (e.g. Coleman *et al.*, 2006, DeLong *et al.*, 2006). The other shallow taxa were also different from those present in Monterey, with the majority never occurring there or only occurring sporadically. Another notable difference between the two locations' profiles was the appearance of more discrete zonation in the Hawaii data; all shallow samples did not appear similar. However, a large caveat to the HOT179 profile must be offered here. The triplicate array hybridizations used to generate these data were

dimmer than Monterey Bay hybridizations, despite using the same amount of starting DNA. Using the same data-filtering parameters optimized for the Monterey Bay profiles, in order to most robustly allow cross-comparison and clustering, very few taxa were "present" in the Hawaii profiles. There are several possible explanations for the poor quality data obtained from this Hawaii profile: (i) estimates of DNA concentrations were inaccurate, and less DNA was hybridized, (ii) the quality of the DNA was poor, with inhibitors present (though this is unlikely as it was spectrophotometrically clean and had been thoroughly extracted and cleaned), (iii) data processing parameters optimized for one location cannot be transferred to another site; this would confound cross-site comparisons, and is not indicated by previous work with the array, or (iv) there was something else substandard in the HOT179 hybridization or scanning process which resulted in less signal. Based on previous hybridizations with assorted Hawaii samples, I believe a combination of (i) and (iv) is most likely. The data obtained, using either Monterey-tailored filtering parameters, unfiltered data, or empirically tuned filtering parameters, are consistent with taxa expectations for the location. Also, previous research (Rich, Konstantinidis and DeLong, 2008) showed transferability of the prototype array to another coastal location in a different ocean basin (Atlantic) using identical processing parameters. Thus, it seems likely that the array will be able to be used across locations without retuning the data processing pipeline.

In addition to examining clade-related depth distributions, Monterey Bay samples were further investigated for variability. Variability among samples and its causes and significance is a major consideration when dealing with natural environmental samples. As indicated by branch length on the sample clustering, there was much more variability among shallow samples than deep ones, as would be expected based on their more variable oceanographic conditions. Profile variability did not correlate overall, however, to Monterey Bay's typical oceanographic seasons (Figure 4a; spring/summer upwelling, fall upwelling, and

winter non-upwelling, as defined in e.g. Pennington and Chavez, 2000, Pennington *et al.*, 2007). There is substantial yearly oceanographic variability at this location in the timing of upwelling events, though, and phytoplankton abundance and growth rates can be "strikingly pulsed" (Pennington and Chavez, 2000). The dynamics of the sampled periods did not fit the time-averaged seasonal delineations, so it may not be surprising that there was little apparent correlation between sample profiles and the site's typical oceanographic seasons. Profiling of additional years might reveal a stronger cumulative signal among seasons. Alternately, a more focused taxa-by-taxa correlation analysis could reveal correlations to oceanographic season that are not evidenced in community-wide profiles but are present in some subset of taxa present.

Sample variability was reflected not only in cluster branch length, however, but also in the relative intensities in each sample's profile, with much greater heterogeneity in intensity among shallow profiles than deep ones. In particular, several shallow sampling dates were notably intense (red starred samples in Figure 4a). The date with the highest intensities is April 25th, 2001, which occurs just after the largest upwelling event in the first 19-mos sampling period (as indicated by nitrate concentrations; sampling date 481 in Figure 10). Other particularly intense samples include Oct3_2000, Oct25_2000, May15_2001, Oct21_2003, and Mar31_2004. These samples were all collected after upwelling events, during upwelling seasons (Figure 10; red arrows and black dashed vertical lines).

Previous studies have shown that different phytoplankton dominate the spring/summer versus fall upwellings (Pennington *et al.*, 2007), which might suggest that different bacteria would also be apparent after spring versus fall upwelling events, even if there were not strong community differences between the annually-averaged seasons overall. However, not only do the intense post-upwelling profiles not all cluster monophyletically, indicating that their profiles are not consistently most similar to one another, but fall and spring upwelling profiles

do not each cluster together either. This suggested that despite phytoplankton differences among upwelling seasons, those taxa targeted by the array do not follow the same trend, at least within the inter-annual variability encompassed by these samples.

Thus, at this study's sampling frequency, there did appear to be a post-upwelling signature in these data, but at the scale of individual events rather than across seasons, and in the form of increased signal from pre-existing, common, abundant taxa rather than unique ones. The strongest signals came from a group of NAC11-7 targets (EB080_L11F12, EB080_L43F08, EB080_ L27A02, and HTCC2255), and two PR-containing alphaproteobacterial clones lacking phylomarkers (EB000_39F01, EB000_55B11). As described above, these six are all within the *shallow-consistent* or *frequent* cluster of targets. The NAC11-7 roseobacterial clade is often associated with bloom and post-bloom conditions (as reviewed in Buchan *et al.*, 2005), ostensibly due to the common roseobacterial ability to degrade dimethylsulfoniopropionate, an osmolyte produced by a variety of phytoplankton. Thus, the prominent role of NAC11-7 targets in the array data from this coastal upwelling site, and their particular intensity after bloom conditions, is consistent with previous observations of this clade.

It may be surprising, however, that PR-containing targets (the two without phylogenetic markers, and the NAC11-7 HTCC2255 genome) would be among those with the strongest post-bloom responses. The diversity of lineages containing proteorhodopsin genes, and their abundance in a variety of photic marine habitats implies a probable diversity in PR lifestyle use. The role of the PR gene in the ubiquitous SAR11 clade has remained unclear but has been hypothesized to allow survival during lean oligotrophic conditions (Giovannoni *et al.*, 2005, Schwalbach *et al.*, 2005). Alternately, the PR-containing *Bacteroidetes* cultivar *Dokdonia* sp. MED134 showed increased growth in light versus dark conditions in a laboratory culture (Gómez-Consarnau *et al.*, 2007). Many

Bacteroidetes, and *Flavobacteria* in particular (of which MED134 is one), are abundant during and after phytoplankton blooms, and it was hypothesized that in end-bloom conditions of decreasing organic matter, PR might allow MED134 to persist as other heterotrophs declined (Gómez-Consarnau *et al.*, 2007). An additional cellular lifestyle that may be linked, in some lineages, to the PR gene, is a cyclic lifestyle alternating between attached and free-living stages. In this case, PR could provide energy to help cross the "deserts" between particles (postulated for the *Flavobacterial* cultivar *Polaribacter* sp. MED152 in González *et al.*, 2008). Thus, the array-based abundance of PR-containing targets during bloom and post-bloom conditions could have several possible explanations. First, it might simply reflect that these taxa were highly competitive heterotrophs under bloom conditions, with PR genes being incidental to the bloom-related phase of their lifestyle. Second, like the hypothesized role in the MED134 cultivar, PR might have allowed these taxa to persist longer than other heterotrophs as the bloom waned. Lastly, the PR might have played a more an active role in bloom utilization, helping provide the energy for organic matter uptake and/or degradation, and allowing these heterotrophs to compete more effectively for bloom carbon. From the current information, we cannot assess the relative likelihood of each scenario. However, additional oceanographic data from these and adjacent sampling dates could help identify bloom stage. Also, three of the intense array profiles have associated pyrosequence data. It could be used to quantify actual numerical dominance of the PR-containing clones more directly rather than inferred from array intensity, and compared to the other heterotrophs present.

In addition to examining sample variability through the lenses of oceanographic season and of upwelling events and associated blooms, we looked more precisely at the environmental variability through actual nutrient concentrations in each sample, and their correlations to the major variability in the data, to both canonical discriminant (c.d.) and principal component (p.c.)

axes. With the variability among the three depths' array profiles maximized in a single CDA, the strong correlations of each axis to nutrient concentrations (Figure 6b) simply recapitulated the oceanographic differences in nutrient concentrations with depth at this location. The higher concentrations of silicate, phosphate and nitrate in the deeper samples (seen in the oceanographic data plotted in Figure 10) were reflected in those nutrients' correlations to the first c.d. In addition, the correlation of nitrite to the second c.d. indicated that 30m was a chemically, not just photically, distinct environment from 0m (also see Figure 10). A water column nitrite maximum is commonly seen below the mixed layer due to active denitrification of organic nitrogen entering from above, and at Station M1, 30m represents the base of the mix layer through much of the year (Figure S8).

Based on the markedly different chemical and photic environments of the 0m and 30m samples, it is surprising that there were not larger differences in the 0m and 30m array profiles. However, mixed layer depth is quite dynamic at this site, as seen both by the calculated MLD across sampling dates (Figure S8) and in temperature-vs.-depth profiles for each sampling date (not shown), which usually show a gradual decrease of temperature with depth rather than a discrete thermocline. Thus, because of water column mixing, these two communities may have been frequently homogenized. In addition, even without mixing, we would have expected the 30m communities to include a subset of 0m communities, particularly for larger-celled taxa, due to particle sinking. Although the 0m and 30m array profiles did not cluster together, some subtle differences were revealed by the correlation of taxa abundances to CDA axes (Figure 6a), which showed a small number of taxa (EBAC_27G05, EB000_65A11, and EB000_39H12) were differentially common and abundant.

Each of the three sampled depths, when investigated separately, showed distinct relationships between nutrient variability and array profile variability in single-depth PCA correlations to nutrients (Figure 7). At 0m, there was no appreciable correlation between nutrient concentration and sample variability

103

(Figure 7a). This is somewhat surprising given the post-upwelling intensity signature in the communities. However, the uptake of upwelled inorganic nutrients is rapid, and subsequent organic forms of these nutrients were not measured. In addition, the strong wind-based homogenization of the mixed layer might obscure relationships between patchy surface nutrients and community profiles. By 30m, however, array profile variability was related to nutrient concentrations. The nutrient signatures of upwelling events (nitrate, phosphate and silicate) were correlated to sample variability, as were the episodic nitrite maxima caused by remineralization, with an opposite vector direction, as expected (Figure 7b). At 200m, the picture was more complex. Although 200m is a more stable and homogenous chemical environment than shallower depths, there remained considerable intra-annual variability in nutrient concentrations (Figure 10), reflecting deeper upwelling, advection, etc. In this case, however, the upwelling-characteristic nutrients appeared decoupled; the correlation vectors for silicate and phosphate were offset but congruent, while nitrate showed no correlation to array profile variability (Figure 7c). In addition, nitrite produced a correlation vector smaller and roughly perpendicular to those of phosphate and silicate. The 200m samples most influenced by the higher silicate and phosphate (2001_Apr_25, 2002_Apr_11, 2004_Jan_21, 2004_Mar_10, 2004_Mar_31, 2004_May_3) are near the spring bloom timing for each of the sampled years (2003 was not sampled in the spring), although for 2002 the oceanographic data do not indicate a preceding upwelling event. These dates include two which also showed highest 0m array profile intensities. Focusing specifically on the 2004 samples, a decoupling of silicate and phosphate was apparent in the oceanographic data (Figure 10) as well. For example, on May 3rd, phosphate concentration was high, silicate was high, and yet there was a dramatic drop in nitrate levels, compared to the surrounding time periods and occurring throughout the water column.

Diatoms dominate the spring upwelling at this location (Pennington *et al.*,

2007). I hypothesize that the temporal pattern in nitrate, phosphate and silicate concentrations at 200m, particularly evident in dramatic upwelling series in spring 2004, and the strong correlation of array profile variability to silicate and phosphate and decoupling from nitrate, represent post-diatom-bloom remineralization signatures. The sequence of events begins as cold nutrient-rich water upwells through the water column; this is seen most clearly in early spring of 2004. As diatoms bloom and begin to settle through the water column, they are remineralized and may, depending on flux rates, produce a short-lived phosphate increase, as in mid-spring 2004. Depending on the volume of settling material, organic matter degradation may strip that water of some nutrients, which may explain the sharp drop in nitrate throughout the water column so soon after its upwelling-associated spike, concurrent with the high levels of phosphate; remineralized nitrogen in the initial form of ammonia is consumed before it can be converted to nitrate, and existing nitrate is also taken up by the actively degrading community. (Low nitrate levels are not explained by rapid nitrification, since the relatively small spike of nitrite occurs later and is of insufficient magnitude). Finally, as the more recalcitrant frustule-associated component of the sinking diatomaceous organic matter becomes a higher percentage of the total available organic matter, silicate concentrations increase as silicate is remineralized. Additional oceanographic data may shed light on the likelihood of this post-bloom remineralization hypothesis as an explanation for the observed 200m correlations.

Variability among samples can be considered not only in the local context, with Monterey Bay as a particularly dynamic environment, but also in the context of the marine environment more broadly. Ocean surveys can be affected by strong spatial and temporal heterogeneity, as strongly evidenced in several studies of chemical, physical and/or biological variability. In one study, the Cytosub, an un-manned autonomous underwater vehicle with an inline flow cytometer, was tethered for 30 days inside a semi-enclosed harbor within the

Bay of Marseilles and collected data on phytoplankton abundances every 30 minutes (Thyssen *et al.*, 2008). After accounting for diel variations and measurement error, 25% of successive samples had >=32% unexplained variability. It was speculated that this rapid variability at a single sampling point might have been due to be genuine biological patchiness, physical forcings (e.g. winds, tides), community dynamics (e.g. grazing, lysis), or behavior (e.g. migration) (Thyssen *et al.*, 2008). Although the sample proximity to shore likely exacerbated variability from episodic terrestrial inputs of nutrients, etc., this study demonstrated that temporal variability in ocean habitats, particularly coastal ones, is poorly understood.

In addition to temporal variability, spatial variability may significantly impact observations. Unlike a terrestrial environment, a single sampling point in a marine habitat may represent very different water masses as currents shift. In addition, spatial patchiness cannot be explained solely by physical forcing (Martin *et al.*, 2005). Another high-resolution flow-cytometry-based study investigated spatial heterogeneity of *Synechococcus* and heterotrophs over a 120-km diameter region of the Celtic Sea (Martin *et al.,* 2005). Repeated triangular transects indicated that the variability between sampled communities 12km apart could equal the variability seen over seasonal cycles in this area. Furthermore, correlations to variability in physical factors (temperature, salinity and density) could account for *at most* 44% of the observed variability. Nor could the fluctuations be due to population doubling, or to mixing from below. The authors suggested that all time-series studies be accompanied by in-depth spatial surveys of the region as well, periodically through the sampling duration, to better constrain the percent of observed variability that could be apportioned to temporal dynamics versus what is just patchiness.

In this vein, Station M1 is in mid-Monterey Bay and is significantly affected by the seasonal Davenport Upwelling Plume which leaves the coast at Santa Cruz and flows southward through the middle of the Bay (Pennington and

Chavez, 2000). Conditions can shift this plume, and biological oceanographic parameters can be dramatically different from its edge to its middle (Pennington and Chavez, 2000). Therefore one might expect additional variability among samples from this site, due to movement of the plume, which is a major driver of site biology.

Lastly, in addition to examining particular taxa and their potential correlations to environmental parameters, the genome proxy array has the ability to indicate the presence of non-target strains and to reveal population shifts over time. This process, demonstrated for the SAR86 target EB000_45B06 in the Monterey Bay data in Figure 9, allows one to tunnel in from the overall array-based target probeset intensity, to the likely genetic relatedness of the hybridizing strain to that targeted, to the similarity in the pattern of hybridization across the target probeset among different samples. For EB000_45B06, the second stage of this analysis – that is, looking at the Tukey biweight signal across the probeset – suggested that hybridized DNAs all had fairly similar identities to the targeted strain. However, the finer-scale level of analysis suggested the presence of four different hybridization patterns in the 39 samples in which this target was present, based on the clustering of pair-wise Pearson correlations of the pattern among samples (Figure 9). The ecological relevance of these potential populations was suggested by the sample origins of each cluster; rather than each cluster being from a mix of depths, as expected if these differing non-target but related DNAs had similar ecology, three of the four clusters had cohesive occurrence patterns. Two clusters arose from 0m samples, each with one aberrant 30m sample, and one cluster arose from 30m samples, with one aberrant 0m sample.

The array-based conclusion of EB000_45B06 population heterogeneity could be cross-checked using the pyrosequence data. However, two of the three pyrosequenced samples fall into the same pattern-cluster, and the third is in an adjacent cluster but has a high correlation to the first two. Ideally, a target

present in all three pyrosequenced samples but with quite distinct patterns in at least two would be cross-validated. BLASTing the target sequence against each of the three databases and plotting the results as recruitment plots, showing relative identity to the target, might not reveal the differences seen with the array, since, for example, in the case of EB000_45B06, the TBW analysis already suggests all four hypothetical populations have similar identity to the target. Therefore, the identity of recruited fragments to one another would ideally be compared as well, and this can only be achieved with overlapping recruits, which requires that the taxa be abundant. To lay the groundwork for this further analysis, the probe-set hybridization pattern correlations were clustered for every target occurring in all three pyrosequenced samples (Figure S9). The differences in these diagrams highlight the differing levels of population homogeneity among lineages at this site. Based on the dual requirements of showing high mean signal intensity in all three samples, and having dissimilar patterns of hybridization, the best candidates for BLAST-based investigation of the pyrosequence data are EB000_55B11 and EB000_39F01, both of which carry the PR genome but lack phylogenetic markers. Future work will examine the population structure of these two clones in the three pyrosequence datasets, if either provides sufficient coverage to do so.

In conclusion, exploration of the array profiles and the underlying causes of their variability allows a more refined understanding of target natural history, and of community dynamics over time, relative to most other methods available. Thus far, we tracked the genotype abundances of 268 marine target taxa through 57 samples collected across four years in Monterey Bay, at three oceanographically-distinct depths. 95 taxa were present in at least one sample, and most taxa showed differential distribution with depth. Highly abundant shallow taxa included representatives of the SAR86, SAR116, SAR11, and Roseobacter clades. Notably, the majority of abundant shallow taxa contained the proteorhodopsin gene. Highly abundant deep taxa included representatives

of marine pelagic euryarchaea, deltaproteobacteria (including the SAR324 clade), and relatives of invertebrate symbionts. All 200m samples clustered together to the exclusion of 0m and 30m samples, although there was no clear clustering of each of the shallower depths. No clustering-based correlation of sample profile to oceanographic season was seen, but overall profile intensity "blooms" were observed in profiles after episodic upwelling events, and possible post-bloom remineralization events were indicated in several 200m samples. In addition, the single depth profile from Hawaii also showed depth-specific taxa distributions, whose composition was markedly different than Monterey Bay samples, although the 500m sample clustered basal to the MB 200m samples.

A unique potential contribution of this array platform is the ability to delineate different populations of closely-related cells, and their dynamics over time. A key next step will be validating this array-based population mapping through the three 0m pyrosequence datasets. In addition, further correlations of environmental data for these samples, and temporal autocorrelation analysis, will help clarify temporal patterns in the array profiles, and define the strength of annual community cyclicity.

Time-series ecology in marine microbial systems is vital to expanding our knowledge of marine microbes from snapshots of taxa, gene contents, and biogeochemical potentials, into a more realistic view of the dynamic nature of these communities, likely variable on the scale of hours and milliliters. Until it is practicable to sequence large numbers of environmental samples for time-series studies, tools that can inexpensively and precisely track native taxa, at levels of phylogenetic discrimination relevant to their ecology, remain an important goal of microbial ecological methodology. Furthermore, sifting of vast metagenomics datasets without a priori hypotheses remains challenging and unwieldy, and complementary tools are required to help direct such investigations. Monterey Bay is one of the best-studied sites in the global ocean, and this genome-proxy array-based investigation of its dynamics brings new nuance to the picture of its

microbial communities. In addition, the array-based evidence for multiple populations, potentially with distinct ecological niches, poses specific questions for exploration in metagenomic datasets from this and other locations.

Cyanobacteria
Cyanothece 1
Prochlorococcus 7
Synechococcus 8

16

Chloroflexi
SAR202 2
Unclassified 3

5

Acidobacteria

3

Gemmatimonadetes

2

EF100-108A04
Cluster

2

SAR406
Cluster

4

Euryarchaeota
Marine Grp I 3
Unclass 3

6

Crenarchaeota

3

Actinobacteria
Janibacter-related 2
Microthrix-related 4

6

Planctomycetes

5

Gammaproteobacteria

Alteromonadales 4	agg47 2
Chromatiales 3	OHB-2 2
Oceanospirillales 6	EB000-65A11 1
Pseudomonadales 4	KI89C 3
Thiotrichales 2	NEP4 2
Vibrionales 8	OM60 4
Rexneles 1	SAR86 10
	SAR156 2

54

Verrucomicrobia

6

Spirochaetes

1

Betaproteobacteria

Burkholderiales 2	Rhodocyclales 1
Methylophilales 2	OM156 1
Nitrosomonadales 1	

7

24

15

Bacteroidetes
Flavibacteria 10
Sphingobacteria 5

61

Delta-
proteobacteria
Desulfobacterales 2
Myxococcales 1
SAR324 16
OM27 2
Unclassified 3

48
No SSU rDNA

Alphaproteobacteria 1
Gammaproteobacteria 13
Deltaproteobacteria 1
Proteobacteria (unresolved) 10
Cyanobacteria 2
Bacteroidetes 1
Planctomycetes 1
Euryarchaeota 4
Unknown 2

Alphaproteobacteria

Parvularcusales 1	*Pelagibacter*-related 2	nuHF1 1
Rhizobiales 4	*Wolbachia*-related 1	Unclassified 2
Rhodobacterales 27	SAR116 8	
Rhodospirales 3	OM75 3	
Sphingomonadales 2	731_112 1	

0.10

Figure 1. Radial tree illustrating the phylogenetic relationships among the 268 targets of the expanded genome proxy array. Numbers indicate the number of targets within each phylogenetic clade. Sequences from clones lacking a small subunit rRNA gene (SSU) phylomarker are represented separately by the hexagon. Tree was created based on alignment of 16S rRNA sequences using ARB.
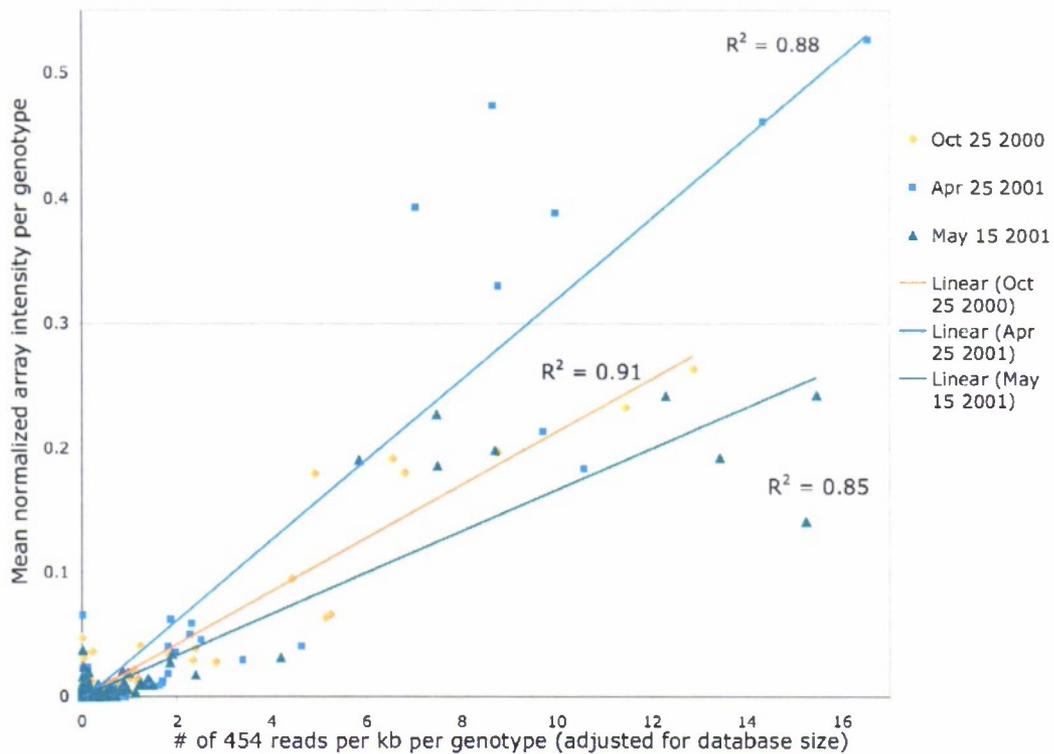
Figure 2. Cross-comparison of array- and pyrosequence-based target abundances for three MB samples. Using BLASTN parameters optimized to mimic array cross-hybridization, all 268 targeted genomes and genome fragments were BLASTed against the pyroseqeunce database for each sample. Pyrosequences were assigned to one or more array targets, proportional to the bitscore of each match. The number of pyrosequences matching each target was normalized to target length and database size, and compared to the unfiltered array signal (see Methods and Results) of the same clone. Correlation lines were not forced through the origin.

a)

b)

Temperature (deg. C)
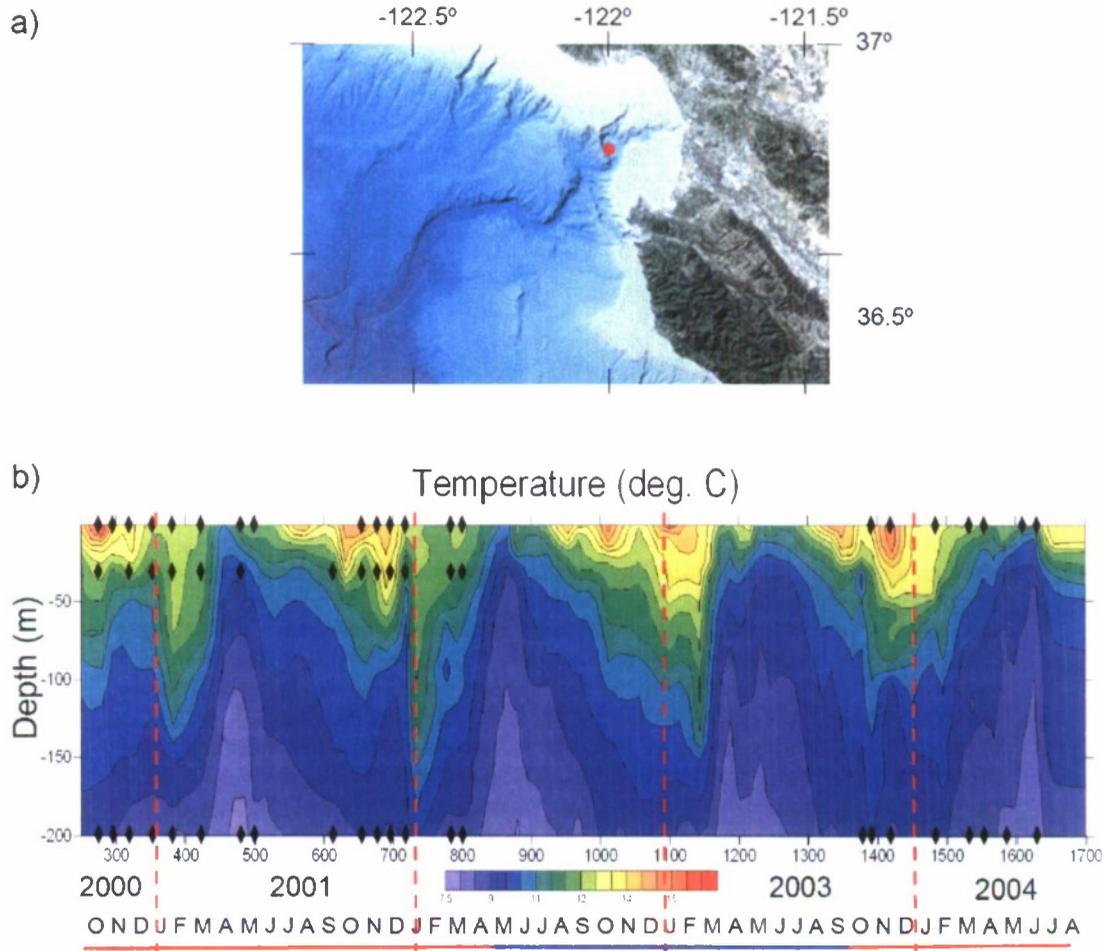
Figure 3. Monterey Bay sample origin. (a) Samples were collected in Monterey Bay, California, at Station M1 (red circle on satellite image from XXXX). (b) Samples from three depths over ~ 4 years, hyrbidized to the array in this study, are shown in relation to the site's temperature vs. depth profile for the same period. Samples (black diamonds) were collected during two consecutive sampling periods (horizontal solid red bars), separated by a sampling hiatus (horizontal blue bar). X-axis numbers represent sampling duration from January 1st, 2000, with years indicated below and delineated by dashed vertical red lines. Months are indicated by their first-letter designations. 42 samples from the first 19-mos period at 0m, 30m, and 200m, and 15 samples from 0m and 200m over the final ~9-mos of sampling, were hybridized to the array.

113

Figure 4. Clustering of hybridizations by sample and by genotype. Hierarchical clustering was performed in GenePattern using Pearson correlation (see Methods) and is shown across the top for samples and along the side for genotypes. Genotypes are color-coded by phylogenetic identity and/or gene content of particular interest (see color legend). Intensity of yellow-to-red color for each genotype and sample date indicates relative mean organismal signal. (a) Monterey Bay samples. Samples are named Depth_Year_CollectionDate, and are color-coded by depth and by oceanographic season (see color legend and text). The break between shallow and deep samples is additionally indicated by the blue vertical dashed line. Clusters of targets referred to in the text, "shallow-consistent", "shallow-frequent", and "deep-consistent", are boxed with dashed red lines. Red asterisks denote samples with particularly intense 0m profiles; the 30m and 200m samples for the same dates, where available, are indicated by blue asterisks. (b) Hawaii samples from cruise HOT179, named by depth of sample. (c) Hawaii samples and MB samples clustered together. Hawaii samples denoted by red dashed box; the three shallow HOT179 samples cluster separately from all MB samples, while the 500m HOT179 clusters basally to the 200m MB samples.

a)

Samples

Genotypes Present

Sample Legend

0m
30m
200m

spring/summer
winter
fall

Target Genotype Legend

Cyanobacteria
PR-containing
bchl-containing
Roseobacteria
SAR11
Archaea
possible S-oxidizing link
16S- or 23S- containing
MB-derived clones
HOT-derived clones
genome

Array Signal

Less   More

cyanobacteria
PR-containing
bchl-containing
SAR11
Archaea
poss. S-oxidizers
16S- or 23S- cont.
MB clones
HOT clones
genomes

b) Samples

Target Genotype Legend

- Cyanobacteria
- PR-containing
- SAR11
- Archaea
- possible S-oxidizing link
- 16S- or 23S- containing
- MB-derived clones
- HOT-derived clones
- genomes

Array Signal

Less   More

Genotypes Present

Name	Description
EB750_10B11	unk. Gammaproteobacteria related to S-oxidizing symbionts
EBAC750_11E01	Alphaproteobacteria ; SAR11 ; subgroup 2
EB750_03B02	Deltaproteobacteria ; SAR406
DeepAnt_EC39	Archaea ; Euryarchaeota G2 ; came from 500m Antarctic Polar Front
EB750_01B07	unk. Deltaproteobacteria
EB750_02H09	Gammaproteobacteria ; SAR156
HF4000_08N17	Deltaproteobacteria ; SAR324 cluster ; ctg_NISA008 clade
HF4000_37C10	Alphaproteobacteria ; SAR11
EB000_36F02	Alphaproteobacteria ; Rhodobacterales ; CHAB1-5
HF4000_384_009C18	Alphaproteobacteria ; Rickettsiales ; SAR11 cluster
EB750_09G06	Alphaproteobacteria ; looks like SAR11
EB75010A10	unk. Gammaproteobacteria
HF0770_09N20	Deltaproteobacteria ; SAR324
HF4000_23L14	Gammaproteobacteria ; Thiotricales ; ZD0405 clade
EB080_69G07	unk. Alphaproteobacteria
HF0070_15B21	Deltaproteobacteria ; SAR324
HOT2C01	unk. Alphaproteobacteria
HF10_45G01	unk. Proteobacteria
HF0130_04F21	Betaproteobacteria ; OM156
Pro_NATL2A	Cyanobacteria ; Prochlorales ; Prochlorococcus LL clade ; high B/A clade I
HF0070_08D07	Gammaproteobacteria ; K189A clade
EBACred_20E09	Gammaproteobacteria ; SAR86-t
HF0010_05E14	Gammaproteobacteria ; SAR92
HF0010_10C01	Alphaproteobacteria ; Rhodobacterales ; SAR102 ; SAR102
HF0070_14E07	Alphaproteobacteria ; OM75
HOT4E07	Gammaproteobacteria ; SAR86-I
Pro_MIT_9515	Cyanobacteria ; Prochlorales ; Prochlorococcus HL clade ; low B/A clade I
HF0070_30B07	Deltaproteobacteria ; SAR324 cluster ; SAR276 clade
HF0130_05G09	Deltaproteobacteria ; SAR324
HF10_12C08	unk. Alphaproteobacteria
HF0010_11K06	Gammaproteobacteria ; SAR86 cluster ; SAR89 clade
ProMED4	Cyanobacteria ; Prochlorales ; Prochlorococcus HL clade ; low B/A clade I
Pro_9312	Cyanobacteria ; Prochlorales ; Prochlorococcus HL clade ; low B/A clade II
Pro_AS9601	Cyanobacteria ; Prochlorales ; Prochlorococcus HL clade ; low B/A clade II

Sample columns: 25m_HOT179, 75m_HOT179, 125m_HOT179, 500m_HOT179

116

c)

Samples

Genotypes Present

Sample Legend

0m
30m
200m

spring/summer
winter
fall

Target Genotype Legend

Cyanobacteria
PR-containing
bchl-containing
Roseobacteria
SAR11
Archaea
possible S-oxicizing link
16S- or 23S- containing
MB-derived clones
HOT-derived clones
genome

Array Signal

Less   More

117

Figure 5. Principal component analysis of Monterey Bay samples by array hybrization data. Deep samples are separated from shallow samples, and the variability of 30m samples is mostly encompassed within the 0m sample varaibility.

Figure 6. Canonical discriminant (c.d.) analysis of Monterey Bay sample (0m ○ , 30m + , and 200m △) array data, with parameter correlations to c.d. axes indicated by vector length and direction. (a) Genotype abundance correlations to c.d. axes; the distribution of particular taxa drive the differentiation of depths. (b) Nutrient correlations to c.d. axes; nutrients are dramatically different between the three depths, and this strong difference is recapitulated in the correlations to c.d. axes.

Figure 7. Principal component (p.c.) analyses of Monterey Bay samples at each depth, with nutrient (nitrate, nitrite, phosphate and silicate) correlations to p.c. axes indicated by vector length and direction. Each sample is desigated by its month and year. (a) 0m samples; the sample varaibility among 0m samples is not strongly correlated to differing nutrient conencentrations. (b) 30m samples; there is a strong correlation to all four nutrients, reflecting the strong upwelling signature at the base of the mixed layer. (c) 200m samples; nitrite, phosphate and silicate each correlate to sample variability, in distinct ways.
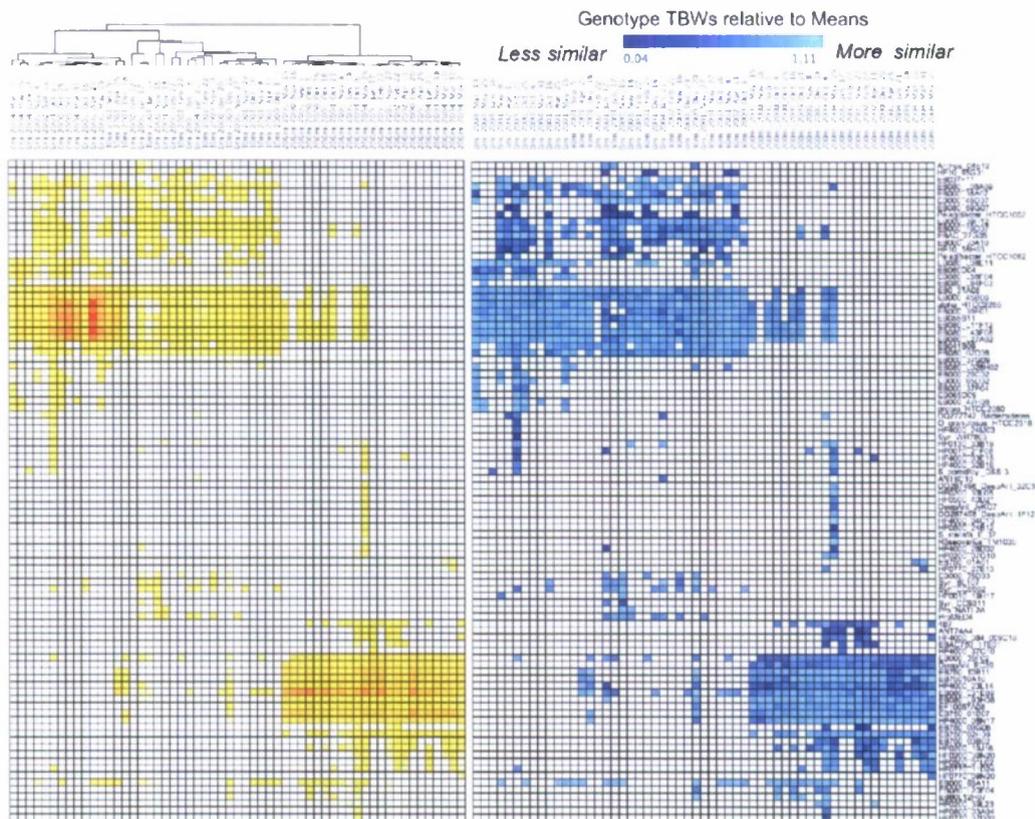
Figure 8. Evaluating the genetic relatedness of community DNA hybridized to the array. On the left are mean organism signals as shown in Figure 4a, repeated here for side-by-side examination. On the right are the relative ratios of the Tukey Biweights (TBW) to the means for each organism (samples in same order as clustering based on mean signals, on left). This ratio is related to the identity of hybridized DNA to the target sequence. Hybridized DNAs with a large relative drop in signal when assessed as TBW rather then as mean (darker blue) have a less even signal across their target probesets, and are thus inferred to be less closely related to the target sequence (i.e., 80-90% ANI), whereas hybridized DNA with higher TBW:Mean ratios (lighter blue) are inferred to be genotypes more closely related to targeted sequences (i.e. >90% ANI), as in Rich, Konstantinidis and DeLong (2008).
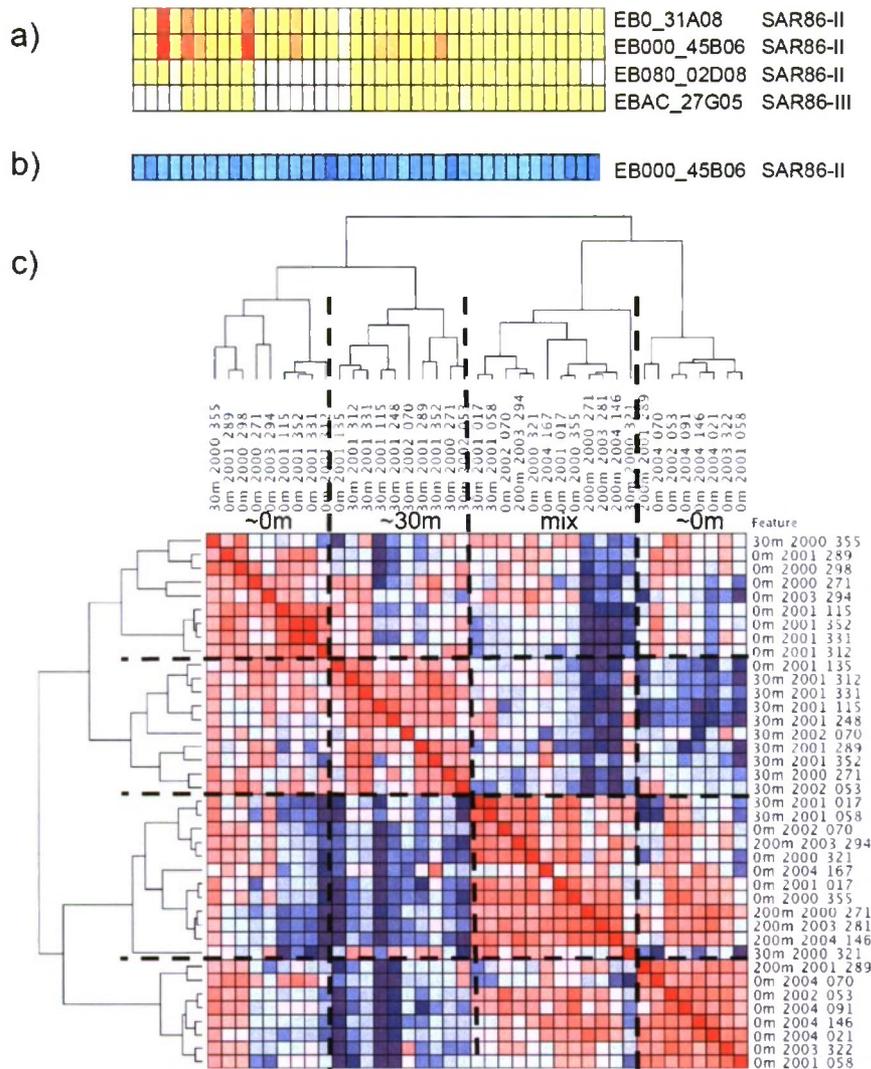
Figure 9. Revealing population heterogeneity by the genome proxy array: tunneling in on strain and population information. (a) Mean target intensity for SAR86 target strains present in Monterey Bay samples (as in Figure 4a). EB000_45B06 is ubiquitous in shallow samples. (b) Tukey biweight intensity for the SAR86-II target EB000_45B06 (as in Figure 8). By this index alone, subpopulations are not strongly evident, (c) Pair-wise Pearson correlations of the signal pattern across the EB000_45B06 probset, between every sample in which it occurred. Samples are clustered based on similarity of probset pattern (assessed by Pearson correlation). Four major clusters of samples are present, delineated by black dashed lines, evident in both the clustering patterns and in the matrix diagonal. Red indicates high Pearson correlation, white is intermediate, blue is low.

Figure 10. Sample oceanographic context. Panels show nitrate, nitrite, silicate and phosphate concentrations through the sampling period. Black diamonds denote samples hybridized to array. Blue arrows at top of each panel indicate samples whose 0m array profiles were particularly intense. Red arrows at bottom of panels indicate 200m samples whose variability was correlated to silicate and phoshphate (Figure 7).

Table 1: Array Targets

(colored just to allow easier viewing)

Accession:	Identity:	Array Probeset Name	Phylogenetic Affiliation
U40208	ORF 4B7	4B7	Archaea ; Crenarchaeota GI
AF393466	ANT_74A4	ANT74A4	Archaea ; Crenarchaeota GI
AF268611	EB000_37F11	EB037F1t	Archaea ; Euryarchaeota G2
not yet in	EF100_57A08	EF10057A08	Archaea ; Euryarchaeota G2
DQ257435	HF10_03D09	HF10_03D09	Archaea ; Euryarchaeota G2
DQ257434	HF70_19B12	HF70_19B12	Archaea ; Euryarchaeota G2
DQ156348	HF70_59C08	HF70_59C08	Archaea ; Euryarchaeota G2
EF089401	HF10_29C11	HF10_29C11	Archaea ; Euryarchaeota G2
AY316120	DeepAnt_EC39	DeepAnt_EC39	Archaea ; Euryarchaeota G2
AY534910	DeepAnt_JyKC7	DeepAnt_JyKC7	Archaea ; Euryarchaeota G2
DQ118403	Alv_FOS1	DQ118403_Alv_FOS1	Archaea ; Euryarchaeota G2
DQ118404	Alv_FOS4	DQ118404_Alv_FOS4	Archaea ; Euryarchaeota G2
DQ078751	Alv_FOS5	DQ078753_Alv_FOS5	Archaea ; Euryarchaeota G2
DQ156349	HF70_39H11	HF70_39H11	Archaea ; Euryarchaeota G2
AY458629	EB750_01A01	EB750_01A01	putative Archaea
	Agreia sp. PHSC20c1	actino_PHSC20C1	Actinobacteria ; Actinobacteria ; Actinomycetales
	Janibacter sp. HTCC2649	Janibacter_HTCC2649	Actinobacteria ; Actinobacteria ; Actinomycetales
In-house	HF200_20K23	HF0200_20K23	Actinobacteria ; Actinobacteria ; Candidatus "Microthrix"
In-house	HF70_17F14	HF0070_17F14	Actinobacteria ; Actinobacteria ; Candidatus "Microthrix"
In-house	HF4000_04C13	HF4000_04C13	Actinobacteria ; Actinobacteria ; Candidatus "Microthrix"
In-house	EB080_L08E11	EB080_L08E11	CFB ; uncultivated Cytophaga
In-house	HF0770_11D24	HF0770_11D24	Chlamydiae/Verrucomicrobia group ; Verrucomicrobia ; Uncultured Verrucomicrobia
In-house	HF70_30L02	HF0070_30L02	Chlamydiae/Verrucomicrobia group ; Verrucomicrobiales
In-house	HF200_39L05	HF0200_39L05	Chlamydiae/Verrucomicrobia group ; Verrucomicrobiales
In-house	HF4000_13K17	HF4000_13K17	Chlamydiae/Verrucomicrobia group ; Verrucomicrobiales
In-house	HF10_05E02	HF0010_05E02	Chlamydiae/Verrucomicrobia group ; Verrucomicrobiales ; MB11C04 clade
In-house	HF200_06I16	HF0200_06I16	Chloroflexi ; Chloroflexi (class) ; Unclassified Chloroflexi
In-house	HF200_09I09	HF0200_09I09	Chloroflexi ; Chloroflexi (class) ; Unclassified Chloroflexi
In-house	HF0500_03M05	HF0500_03M05	Chloroflexi ; Chloroflexi (class) ; Unclassified Chloroflexi
In-house	HF0770_09E03	HF0770_09E03	Chloroflexi ; Chloroflexi (class) ; Unclassified Chloroflexi
In-house	HF4000_28I02	HF4000_28I02	Chloroflexi ; Chloroflexi (class) ; Unclassified Chloroflexi
In-house	HF130_06E03	HF0130_06E03	EF100_108A04 cluster, which was previously in Agg47 by Suzuki et al 2004
In-house	HF0500_16O16	HF0500_16O16	EF100_108A04 cluster, which was previously in Agg47 by Suzuki et al 2004
In-house	HF200_23L05	HF0200_23L05	Fibrobacteres/Acidobacteria group ; Acidobacteria ; Acidobacteria (class) ; Acidobacteriales
In-house	HF4000_26D02	HF4000_26D02	Fibrobacteres/Acidobacteria group ; Acidobacteria ; Acidobacteria (class) ; Acidobacteriales
In-house	HF0770_27F21	HF0770_27F21	Fibrobacteres/Acidobacteria group ; Acidobacteria ; Acidobacteria (class) ; Unclassified Acidobacteriales
DQ295242	AntFos_04D05	AntFos_04D05	Gemmatimonadetes ; Gemmatimonadales ; Gemmatimonadaceae ; Gemmatimonas
In-house	HF130_03D03	HF0130_03D03	Gemmatimonadetes ; Gemmatimonadales
DQ295238	AntFos_04E12	AntFos_04E12	Gram Positive High G + C
	Lentisphaera araneosa HTCC2155	L_araneosa_HTCC2155	Lentisphaerae ; Lentisphaerales
EF591885			Planctomycetes ; Prolixibacter
EF591884			Planctomycetes ; Prolixibacter
In-house	HF0500_18F23		Planctomycetes
EF107103	HF10_45G01	HF10_45G01	Proteobacteria
EF089397	EB000_35D03	EB000_35D03	Proteobacteria
EF107099	EB000_49D07	EB000_49D07	Proteobacteria
EF100190	HF10_19P19	HF10_19P19	Proteobacteria
EF089399	EB000_39H12	EB000_39H12	Proteobacteria
EF100191	HF10_25F10	HF10_25F10	Proteobacteria
AY322455	HF07_02C01	HF07_02C01	Proteobacteria ; Alphaproteobacteria
EF089398	EB000_39F01	EB000_39F01	Proteobacteria ; Alphaproteobacteria

Table 1 continued

EF107105	EB080_69G07	EB080_69G07	Proteobacteria ; Alphaproteobacteria
EF107102	HF10_12C08	HF10_12C08	Proteobacteria ; Alphaproteobacteria
AE009920	EB000_29C02	EB000_29C02	Proteobacteria ; Alphaproteobacteria
in-house	HF130_06E21	HF0130_06E21	Proteobacteria ; Alphaproteobacteria ; D10R clade
AY458084	EB750_09G06	EB750_09G06	Proteobacteria ; Alphaproteobacteria ; putative SAR11
in-house	HF130_31E21	HF0130_31E21	Proteobacteria ; Alphaproteobacteria ; nalE1 clade
in-house	HF70_14E07	HF0070_14E07	Proteobacteria ; Alphaproteobacteria ; OM75
in-house	HF130_20P23	HF0130_20P23	Proteobacteria ; Alphaproteobacteria ; OM75
	Parvularcula bermudensis HTCC2503	Parvularcula_HTCC2503	Proteobacteria ; Alphaproteobacteria ; Parvularcules
	Fulvimarina pelagi HTCC2506	F_pelagi_HTCC2506	Proteobacteria ; Alphaproteobacteria ; Rhizobiales
	Nitrobacter sp. Nb 311A	Nitrobacter_Nb_311A	Proteobacteria ; Alphaproteobacteria ; Rhizobiales
in-house	HF4000_48A13	HF4000_48A13	Proteobacteria ; Alphaproteobacteria ; Rhizobiales ; Parvibaculum
in-house	HF4000_32B18	HF4000_32B18	Proteobacteria ; Alphaproteobacteria ; Rhizobiales ; SIMO CL-S30-58 clade
CP000264	Jannaschia CCS1	Jannaschia_CCS1	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH672414	Loktanella vestfoldensis SKA53	L_vestfoldensis_SKA53	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH672428	Oceanicaulis alexandrii HTCC2633	O_alexandrii_HTCC2633	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH724131	Oceanicola batsensis HTCC2597	O_batsensis_HTCC2597	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH724107	Oceanicola granulosus HTCC2516	O_granulosus_HTCC2516	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH902578	Rhodobacterales HTCC2654 aka Maritimibacter alkaliphilus HTCC2654	Rhodobacterales_HTCC2654	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
N/A	Roseobacter sp. CCS2	Roseobacter_CCS2	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH902583	Roseobacter sp. MED193	Roseobacter_MED193	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
N/A	Roseobacter sp. SK209-2-6	Roseobacter_SK_209_2_6	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH902584	Roseovarius sp. 217	Roseovarius_217	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
DS022279	Roseovarius sp. HTCC2601 aka Pelagibaca bermudensis HTCC2601	Roseovarius_HTCC2601	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
ABCL01000012	Roseovarius sp. TM1035	Roseovarius_TM1035	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CP000031	Silicibacter pomeroyi DSS-3	S_pomeroyi_DSS_3	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
U58356 (16S only)	Sagittula stellata E37	S_stellata_E_37	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH959310	Sulfitobacter sp. EE-36	Sulfitobacter_EE36	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
CH959312	Sulfitobacter sp. NAS-14.1	Sulfitobacter_NAS14_1	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
AY458649	EB080_L28H02	EB080_L28H02	Proteobacteria ; Alphaproteobacteria ; putative Rhodobacterales
in-house			Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
in-house			Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
in-house			Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
in-house			Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
in-house			Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
in-house			Proteobacteria ; Alphaproteobacteria ; Rhodobacterales
DS022282			
AATR01000000	Rhodobacterales HTCC2255	alpha_HTCC2255	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; OM42 clade
N/A	Roseobacter HTCC2150	Rhodobacterales_HTCC2150	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; RCA clade
in-house	HF10_04M21	HF0010_04M21	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Roseobacter-like
in-house	HF70_10D05	HF0070_10D05	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Roseobacter-like
in-house	HF4000_03E16	HF4000_03E16	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Roseobacter-like
in-house	HF0010_10C01	HF0010_10C01	Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; SAR102
in-house	EB000_55D11	EB055D11	Proteobacteria ; Alphaproteobacteria ; Rhodobacter-like
in-house	HF70_31K06	HF0070_31K06	Proteobacteria ; Alphaproteobacteria ; Rhodospirillales
in-house	HF200_01O14	HF0200_01O14	Proteobacteria ; Alphaproteobacteria ; Rhodospirillales
in-house	HF4000_24M03	HF4000_24M03	Proteobacteria ; Alphaproteobacteria ; Rhodospirillales
in-house	HF100_102A06	HF100_102A06	Proteobacteria ; Alphaproteobacteria ; Rhodospirillales ; OM75
in-house	HF0070_02E07	HF0070_02E07	Proteobacteria ; Alphaproteobacteria ; Rickettsiales
AAPV00000000	Pelagibacter ubique HTCC1002	Pelagibacter_HTCC1002	Proteobacteria ; Alphaproteobacteria ; Rickettsiales ; SAR11 clade ; subgroup 1a
CP000084	Pelagibacter ubique HTCC1062	Pelagibacter_HTCC1062	Proteobacteria ; Alphaproteobacteria ; Rickettsiales ; SAR11 clade ; subgroup 1a
in-house	HF4000_[384]00UC18	HF4000_384_00UC18	Proteobacteria ; Alphaproteobacteria ; Rickettsiales ; SAR11 clade
in-house	HF0010_09O16	HF0010_09O16	Proteobacteria ; Alphaproteobacteria ; Rickettsiales ; SAR11 clade
in-house	HF0770_37D02	HF0770_37D02	Proteobacteria ; Alphaproteobacteria ; Rickettsiales ; SAR11 clade ; Pelagibacter
in-house	HF4000_37C10	HF4000_37C10	Proteobacteria ; Alphaproteobacteria ; Rickettsiales ; SAR11 clade
AY458633	EB750_11E01	EBAC750_11E01	Proteobacteria ; Alphaproteobacteria ; Rickettsiales ; SAR11 clade ; SAR11 ; subgroup 2
AY458637	EB750_02H05	EB75002H05	Proteobacteria ; Alphaproteobacteria ; Rickettsiales ; SAR11 clade ; SAR11 ; subgroup 2
in-house	EF100_94U01	EF100_94U01	Proteobacteria ; Alphaproteobacteria ; roots rhodovibrio
in-house	EB080_L84F03	EB080_L84F03	Proteobacteria ; Alphaproteobacteria ; Roseobacter
DQ295239	AntFos_24C04	AntFos_24C04	Proteobacteria ; Alphaproteobacteria ; Roseobacter
AE008921	EB000_60D04	EB060D04	Proteobacteria ; Alphaproteobacteria ; Roseobacter-like (by best BLAST hits)
AY671989	eBACred_25D05	eBACred_25D05	Proteobacteria ; Alphaproteobacteria ; Roseobacter-like bacteria puf/bch
AY244399	eBACred_02C11	EBred_02C11	Proteobacteria ; Alphaproteobacteria ; SAR116 ; putative SAR116-1
in-house	HF10_13E22	HF0010_13E22	Proteobacteria ; Alphaproteobacteria ; SAR116
in-house	HF70_17D04	HF0070_17D04	Proteobacteria ; Alphaproteobacteria ; SAR116
in-house	HF0070_05I22	HF0070_05I22	Proteobacteria ; Alphaproteobacteria ; SAR116
in-house	HF0070_34A12	HF0070_34A12	Proteobacteria ; Alphaproteobacteria ; SAR116
in-house	HF0070_34E11	HF0070_34E11	Proteobacteria ; Alphaproteobacteria ; SAR116
in-house	EBAC_46D07	EB000_46D07	Proteobacteria ; Alphaproteobacteria ; SAR116 ; SAR116-I
in-house	EB000_37G09	EB000_37G09	Proteobacteria ; Alphaproteobacteria ; SAR116 ; SAR116-II
	Erythrobacter litoralis HTCC2594	E_litoralis_HTCC2594	Proteobacteria ; Alphaproteobacteria ; Sphingomonadales
in-house	HF0500_24D12	HF0500_24D12	Proteobacteria ; Alphaproteobacteria ; Sphingomonadales ; Erythrobacteraceae ; Erythrobacter
in-house	HF10_30A23	HF0010_30A23	Proteobacteria ; Alphaproteobacteria ; 131_112 clade
EF089400	EB000_41B09	EB041B09	Proteobacteria ; Betaproteobacteria
CP000316	Polaromonas sp. JS666 - draft	Polaromonas_JS666	Proteobacteria ; Betaproteobacteria ; Burkholderiales ; Polaromonas
in-house	HF4000_05M23	HF4000_05M23	Proteobacteria ; Betaproteobacteria ; Burkholderiales ; Delftia
	Not yet validly described, OM43 clade HTCC2181	Methylophilales_HTCC2181	Proteobacteria ; Betaproteobacteria ; Methylophilales
AY458645	EB080_L12H07	EB80L12H07	Proteobacteria ; Betaproteobacteria ; Nitrosomonas
in-house	HF130_04F21	HF0130_04F21	Proteobacteria ; Betaproteobacteria ; OM156
AY458647	EB000_36A07	EB000_36A07	Proteobacteria ; Betaproteobacteria ; OM43
in-house	HF0010_04H24	HF0010_04H24	Proteobacteria ; Betaproteobacteria ; Rhodocyclales ; Rhodocyclaceae ; Zoogloea
in-house	HF0010_10D05	HF0010_10D05	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0070_07E19	HF0070_07E19	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0070_15B21	HF0070_15B21	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0130_05G09	HF0130_05G09	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0130_20J24	HF0130_20J24	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0200_14D13	HF0200_14D13	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0200_39I23	HF0200_39I23	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0200_39N20	HF0200_39N20	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0500_03A04	HF0500_03A04	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324

Table 1 continued

in-house	HF0770_09N20	HF0770_09N20	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0770_45N15	HF0770_45N15	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0130_19C70	HF0130_19C70	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF0200_19H16	HF0200_19H16	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324
in-house	HF4000_08N17	HF4000_08N17	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324 cluster clg_N1SA008 clade
in-house	HF0070_30B07	HF0070_30B07	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria ; SAR324 cluster SAR276 clade
in-house	HF0010_08B07	HF0010_08B07	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria
AY458630	EB750_01B07	EB750_01B07	Proteobacteria ; delta/epsilon subdivisions ; Deltaproteobacteria
in-house	HF0200_07G10	HF0200_07G10	Proteobacteria ; delta/epsilon subdivisions; Deltaproteobacteria ; Desulfobacterales ; Nitrospina-like
in-house	HF0130_12L15	HF0130_12L15	Proteobacteria ; delta/epsilon subdivisions; Deltaproteobacteria ; Myxococcales ; E48F11CD clade
in-house	HF0010_01J10	HF0010_01J10	Proteobacteria ; delta/epsilon subdivisions; Deltaproteobacteria ; OM27
in-house	HF0070_10I02	HF0070_10I02	Proteobacteria ; delta/epsilon subdivisions; Deltaproteobacteria ; OM27
AY458631	EB750_03B02	EB750_03B02	Proteobacteria ; delta/epsilon subdivisions; Deltaproteobacteria ; SAB406
in-house	HF0500_01L02	HF0500_01L02	Proteobacteria ; delta/epsilon subdivisions; Deltaproteobacteria ; SAB406 cluster
in-house	HF0010_18O13	HF0010_18O13	Proteobacteria ; delta/epsilon subdivisions; Deltaproteobacteria ; SAB406 cluster A31300B clade
in-house	HF4000_22B16	HF4000_22B16	Proteobacteria ; delta/epsilon subdivisions; Deltaproteobacteria ; SAB406 cluster ESP200-K10-15 clade
DQ267495	DeepAnt_1F12	DQ267495_DeepAnt_1F12	Proteobacteria ; Deltaproteobacteria ;
DQ267496	DeepAnt_32C6	DQ267496_DeepAnt_32C6	Proteobacteria ; Deltaproteobacteria
EF106972	EB080_L20F04	EB080_L20F04	Proteobacteria ; Deltaproteobacteria ; Nitrospinaceae
			Proteobacteria ; Gammaproteobacteria
			Proteobacteria ; Gammaproteobacteria
			Proteobacteria ; Gammaproteobacteria
			Proteobacteria ; Gammaproteobacteria
			Proteobacteria ; Gammaproteobacteria
			Proteobacteria ; Gammaproteobacteria
			Proteobacteria ; Gammaproteobacteria
in-house	EF100_93H11	EF100_93H11	Proteobacteria ; Gammaproteobacteria ; AGG47
in-house	HF0200_40H22	HF0200_40H22	Proteobacteria ; Gammaproteobacteria ; AGG47
	Not yet validly described TW-7	Alteromonadales_TW_7	Proteobacteria ; Gammaproteobacteria ; Alteromonadales
CH959302	Alteromonas macleodii Deep ecotype	A_macleodii_Deep	Proteobacteria ; Gammaproteobacteria ; Alteromonadales
AAOH01000000	Pseudoalteromonas tunicata D2	Pseudo_tunicata_D2	Proteobacteria ; Gammaproteobacteria ; Alteromonadales
in-house	HF4000_16C08	HF4000_16C08	Proteobacteria ; Gammaproteobacteria ; Alteromonadales
AE017340	Idiomarina loihiensis L2TR	I_loihiensis_L2TR	Proteobacteria ; Gammaproteobacteria ; Alteromonadales
DQ295237	AntFos_04D03	AntFos_04D03	Proteobacteria ; Gammaproteobacteria ; ArRCTIC96B-16
AY458646	EB080_L31E09	EB080_L31E09	Proteobacteria ; Gammaproteobacteria ; ARCTIC96BD-19 clade
	Nitrococcus mobilis Nh 231	N_mobilis_Nb_2311_1	Proteobacteria ; Gammaproteobacteria ; Chromatiales
in-house	HF0200_41F04	HF0200_41F04	Proteobacteria ; Gammaproteobacteria ; Chromatiales ; Bivalve endosymbiont clade
in-house	HF0770_27E13	HF0770_27E13	Proteobacteria ; Gammaproteobacteria ; DHB-2 Cluster
jgi 3640063	EB000_65A11	EB000_65A11	Proteobacteria ; Gammaproteobacteria ; EB000_65A11 clade
in-house	HF0010_01E20	HF0010_01E20	Proteobacteria ; Gammaproteobacteria ; K189A_clade
in-house	HF0010_16J05	HF0010_16J05	Proteobacteria ; Gammaproteobacteria ; K189A_clade
in-house	HF0070_08D07	HF0070_08D07	Proteobacteria ; Gammaproteobacteria ; K189A_clade
in-house	EB000_37F04	EB000_37F04	Proteobacteria ; Gammaproteobacteria ; KTcl119 clade
in-house	HF4000_36I10	HF4000_36I10	Proteobacteria ; Gammaproteobacteria ; NEP4 cluster (close to OM60 cluster)
	Marinomonas sp. MED121	Marinomonas_MED121	Proteobacteria ; Gammaproteobacteria ; Oceanospirillales
in-house	HF0070_21F08	HF0070_21F08	Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; Alcanivorax-like
in-house	HF4000_13G19	HF4000_13G19	Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; Alcanivorax-like
in-house	HF0130_25G24	HF0130_25G24	Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; OM182 clade
in-house	HF4000_23O15	HF4000_23O15	Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; OM182 clade
AY458643	EB080_L37D05	EB080_L37D05	Proteobacteria ; Gammaproteobacteria ; OM60
in-house	HF0010_05D02	HF0010_05D02	Proteobacteria ; Gammaproteobacteria ; OM60
in-house	HF0130_01E34	HF0130_01E34	Proteobacteria ; Gammaproteobacteria ; OM60
	Not yet validly described KT-71	gamma_KT_71	Proteobacteria ; Gammaproteobacteria ; OM60 clade
	OM60 clade; HTCC2080	proteo_HTCC2080	Proteobacteria ; Gammaproteobacteria ; OM60 ; OM60 clade
in-house	HF0500_12O04	HF0500_12O04	Proteobacteria ; Gammaproteobacteria ; Pseudomonas ; Pseudomonas
AY458632	EB750_02H09	EB750_02H09	Proteobacteria ; Gammaproteobacteria ; SAR156
in-house	HF4000_19M20	HF4000_19M20	Proteobacteria ; Gammaproteobacteria ; SAR156 cluster ; EB750_02H09 clade
in-house	HF0010_11B23	HF0010_11B23	Proteobacteria ; Gammaproteobacteria ; SAR86 cluster ; CHAB-I-7 clade
in-house	HF0130_23J23	HF0130_23J23	Proteobacteria ; Gammaproteobacteria ; SAR86 cluster ; CHAB-I-7 clade
in-house	HF0070_03O15	HF0070_03O15	Proteobacteria ; Gammaproteobacteria ; SAR86 cluster ; CHAB-I-7 clade
in-house	HF0010_11K08	HF0010_11K08	Proteobacteria ; Gammaproteobacteria ; SAR86 cluster ; SAR86 clade
AY533945	eBACred_29E09	EBACred_29E09	Proteobacteria ; Gammaproteobacteria ; SAR86 I
AY458665	eBAC_04E07	EBAC04E07	Proteobacteria ; Gammaproteobacteria ; SAR86 II
AY129016	EB080_31A08	EB080_31A08	Proteobacteria ; Gammaproteobacteria ; SAR86 II
AY372454	EB080_45B06	EB080_45B06	Proteobacteria ; Gammaproteobacteria ; SAR86 II
EF107104	EB080_02E06	EB080_02E06	Proteobacteria ; Gammaproteobacteria ; SAR86 III
in-house	EBAC_27G05	EBAC_27G05	Proteobacteria ; Gammaproteobacteria ; SAR86 III
Broad #L35786	HF0010_05E14	HF0010_05E14	Proteobacteria ; Gammaproteobacteria ; SAR92
in-house	HF4000_23L14	HF4000_23L14	Proteobacteria ; Gammaproteobacteria ; Thiotricales ; ZD0405 clade
AY744196	eBACred_07D11	EBACred_07D11	Proteobacteria ; Gammaproteobacteria ; putative Vibrionales
CP000020	Vibrio fischeri ES114	Vibrio_fischeri_ES114	Proteobacteria ; Gammaproteobacteria ; Vibrionales ; Vibrio
CH724124	Vibrio splendidus 12B01	V_splendidus_12B01	Proteobacteria ; Gammaproteobacteria ; Vibrionales ; Vibrio ; splendidus group
CH902608	Vibrio sp. MED222	Vibrio_MED222	Proteobacteria ; Gammaproteobacteria ; Vibrionales ; Vibrio ; splendidus group
N/A	Vibrio sp. SWAT-3	Vibrio_SWAT_3	Proteobacteria ; Gammaproteobacteria ; Vibrionales ; Vibrio ; splendidus group
in-house	HF0010_22E23	HF0010_22E23	Proteobacteria ; Gammaproteobacteria ; Vibrionales ; Vibrio/Photobacterium-like
CR354531	Photobacterium profundum SS9	Photobacterium_SS9	Proteobacteria ; Gammaproteobacteria ; Vibrionales ; Photobacterium profundum
in-house	EB080_L93H08	EB750_07C09	Proteobacteria ; Gammaproteobacteria ; ZD0408
in-house	EB080_L93H08	EB080_L93H08	Proteobacteria ; Gammaproteobacteria ; ZD0417
DQ068067	MED13K09	DQ068067_MED13K09	Proteobacteria ; putative Gammaproteobacteria
AY458640	EB000_65D02	EB000_65D02	Proteobacteria
AY458643	EB000_47H08	EB000_47H08	Proteobacteria
AY372453	ANT_32C12	ANT32C12	Proteobacteria
AY372452	ANT_8C10	ANTBAC10	Proteobacteria
in-house	HF0500_06B09	HF0500_06B09	Spirochaetes ; Spirochaetales ; Rhizobiales ; Spirochaeta
in-house	HF0010_19B11_A08	L35766_P4N	unknown ; DMSP degradation; phosphonate degradation
DQ068068	RED17I08	DQ068068_RED17I08	unknown ; PR in alphaproteobacterial PR clade
DQ065755	66A11	DQ065755_66A11	unknown ; PR in alphaproteobacterial PR clade
DQ088847	MedeBAC46A06	DQ088847_MedeBAC46A06	unknown ; PR in alphaproteobacterial PR clade
DQ073706	MedeBAC82F10	DQ073706_MedeBAC82F10	unknown ; PR in alphaproteobacterial PR clade
DQ077553	MedeBAC35C06	DQ077553_MedeBAC35C06	unknown ; PR in alphaproteobacterial PR clade
DQ077554	MedeBAC49C08	DQ077554_MedeBAC49C08	unknown ; PR in alphaproteobacterial PR clade

127

Table 2: Summary of Array Targets by Clade

Cluster:	# of Clones Targeted	# of Genomes Targeted
SAR86	11	-
SAR156	2	-
Thiotrichales	2	-
EB000_65A11	-	-
Oceanospirillales	-	1
Alternomondales	1	4
OM60	2	2
NEP17	1	-
OM182	2	-
Alcanivorax	2	-
Psuedomonadales	2	-
Burkholderiales	2	1
Nitrosomonabales	1	-
Methylophilales	1	1
Chromatiales	2	1
DHB-2 unclass Gamma	2	2
Vibrionales	-	5
Agg47	1	-
K189A	3	-
Rhodobacterales	6	18
T31-112 unclass Alpha	1	-
Rhizobiales	2	2
Shingomonadales	1	-
SAR11	6	2
Rhodopirillales	2	-
SAR116	8	-
unclass. Alphas	3	-
OM75	3	-
unclass. Alphas	4	1
SAR324 deltas	17	-
OM27 deltas	2	-
DeepAnt_1F12 deltas	3	-
Shingibacteriales	5	1
Flavobacteriales	2	8
Verrumicrobiales	5	-
Planctomycetales	3	2
SAR406	4	-
EF100-108A04	3	-
Gemmatinomondales	1	-
Acidobacterales	4	2
Desulfobacterales	3	-
SAR202	3	-
Chloroflexi	3	-
Prochlorales	1	16
Microthrix	5	-
Lentisphaerales	-	1
GII Arch	4	-
other Euks	2	-
GI Arch	3	-
no 16S:	50	
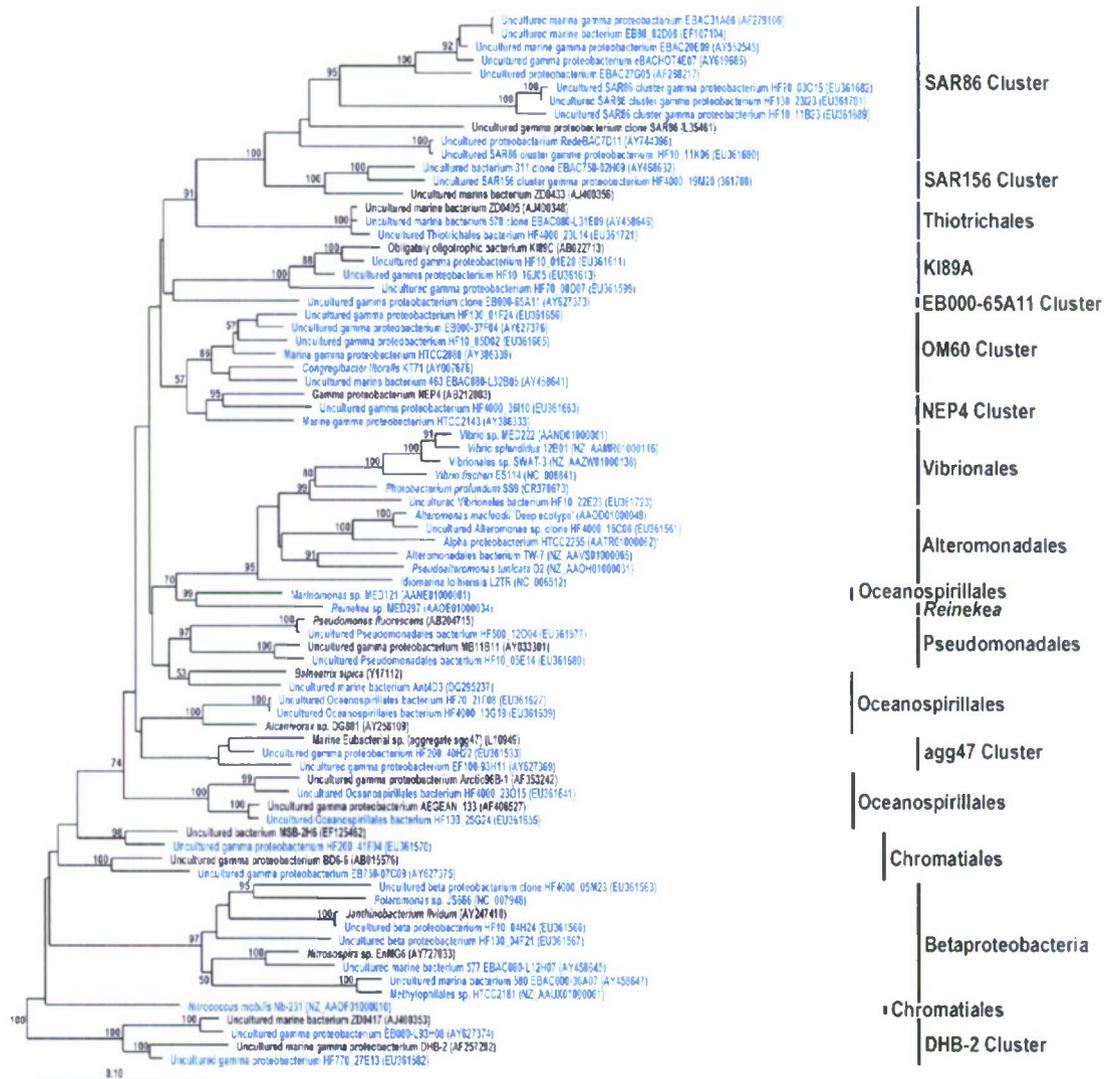
Figures S1-S5. Phylogenetic trees illustrating the relationship of SSU rRNA gene sequences from genomes and uncultivated clones represented on the genome-proxy microarray (blue) and their close relatives (black) as "landmarks". Support for dendrogram topologies is indicated by bootstrap values at nodes determined by the maximum likelihood method (only values >50 are shown). The outgroups used were *Methanomethylovorans victoriae* strain TM (AJ276437) for the bacterial dendrograms, and *Myxococcus xanthus* strain UCDaV1 (AY724797) for the archaeal dendrogram. *The publicly-available SSU rDNA sequence for the *Roseobacter*-like alphaproteobacterial clone HTCC2255 (AATR01000062) is from a Gammaproteobacterium. **S1.** Gamma- and Betaproteobacteria. **S2.** Alphaproteobacteria. **S3.** Deltaproteobacteria and Spirochaetes. **S4.** Other Bacteria. **S5.** Archaea.
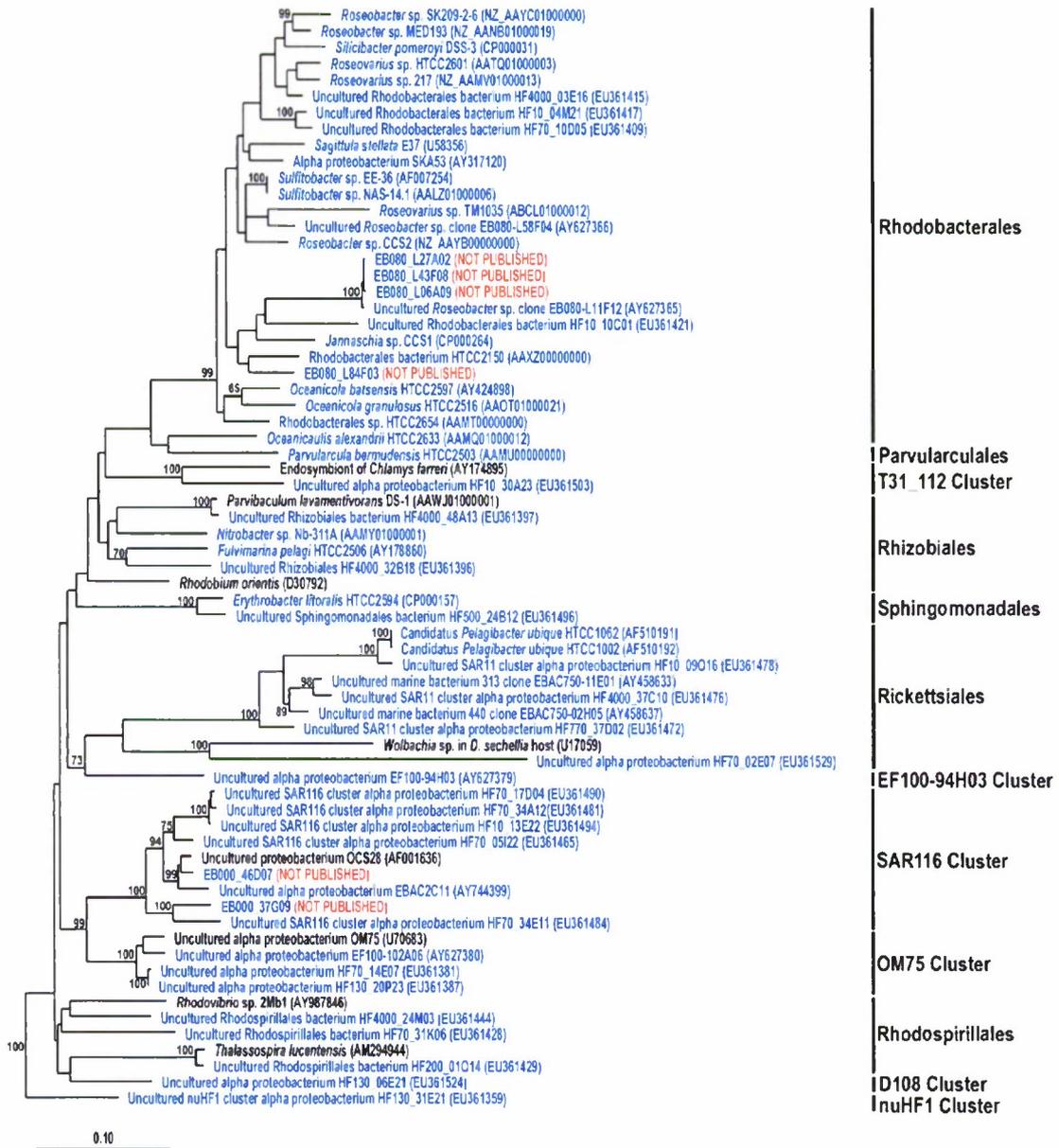
Figure S2. Alphaproteobacterial array targets (blue) and their close "landmark" relatives (black).

130
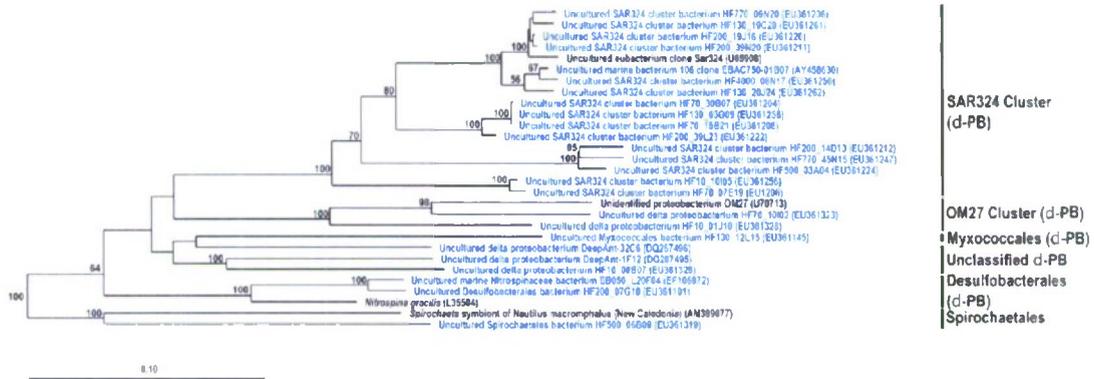
Figure S3. Deltaproteobacterial and Spirochaete array targets (blue) and their close "landmark" relatives (black).
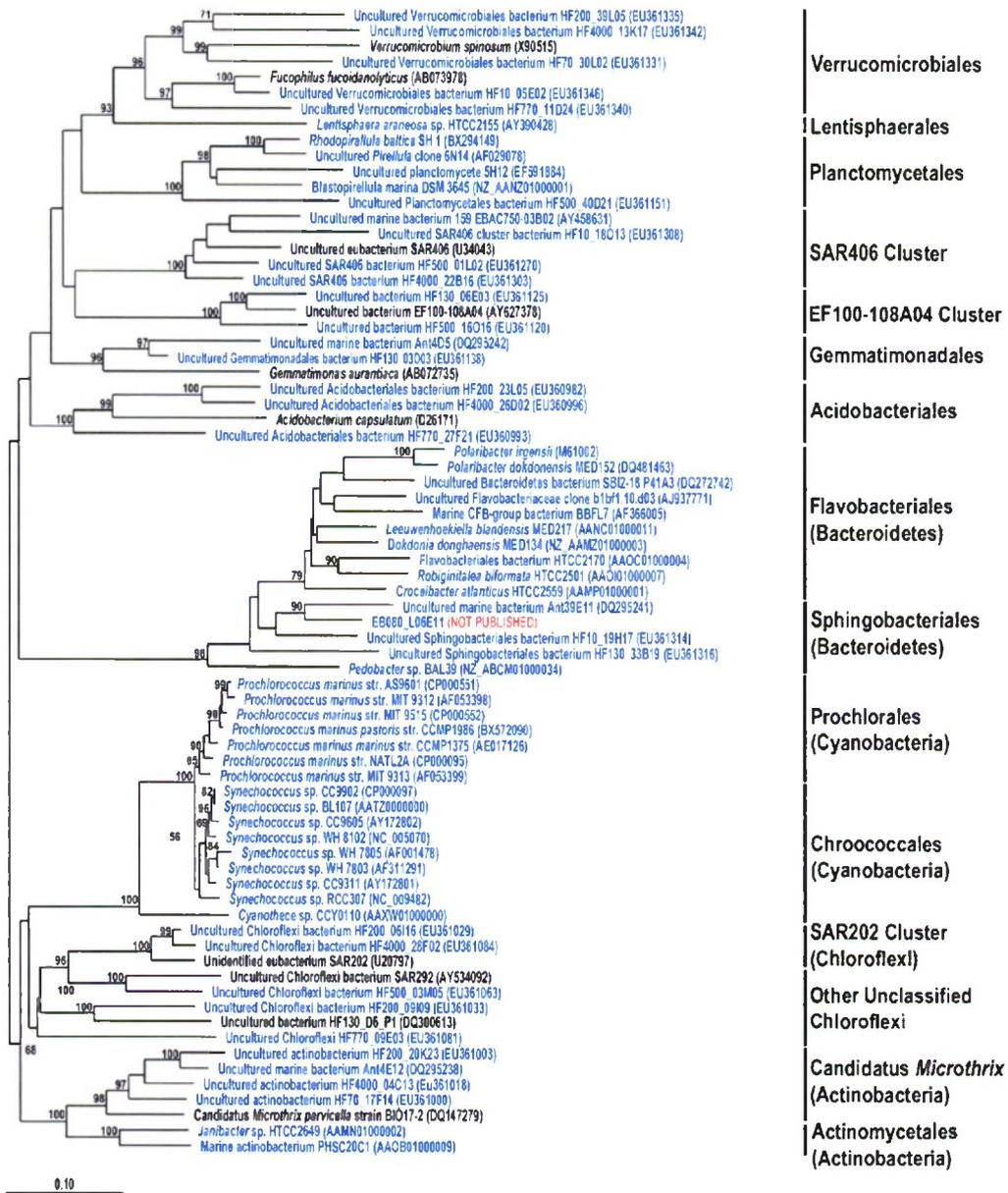
Figure S4. Other bacterial array targets (blue) and their close "landmark" relatives (black).

Figure S5. Archaeal array targets (blue) and their close "landmark" relatives (black).
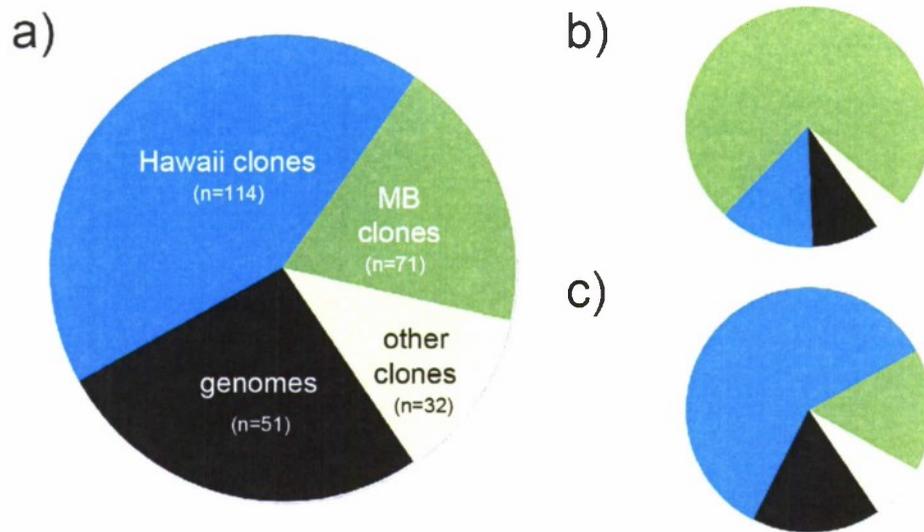
Figure S6. Origin of array targets and their relative array-based occurrences in Monterey Bay and Hawaii samples. (a) Derivation of array targets, either as environmental genome fragments from Hawaii (blue), Monterey (green), other marine sites (beige), or from marine microbial genomes (black). The number of targets in each category is indicated. (b) The proportional abundance of each target type in 57 Monterey Bay samples. (c) The proportional abundance of each target type in 4 Hawaii samples. (b) and (c) are measured as the relative proportion of total array signal across all samples hybridized. Targets derived from a particular environment are proportionally more abundant in that environment.

Figure S7. Array profiles for specific phylogenetic groups of targets. (a) Roseobacter (b) SAR86 (c) SAR11.

a)

Roseovarius_TM1035	Alphaproteobacteria ; Rhodobacterales
S_stellata_E_37	Alphaproteobacteria ; Rhodobacterales
EB000_36F02	Alphaproteobacteria ; Rhodobacterales ; CHAB1-5
S_pomeroyi_DSS-3	Alphaproteobacteria ; Rhodobacterales
O_granulosus_HTCC2516	Alphaproteobacteria ; Rhodobacterales
HF4000_03E16	Alphaproteobacteria ; Rhodobacterales ; Roseobacter-like
O_batensis_HTCC2597	Alphaproteobacteria ; Rhodobacterales
Rhodobacterales_HTCC2654	Alphaproteobacteria ; Rhodobacterales
Roseobacter_CCS2	Alphaproteobacteria ; Rhodobacterales
Roseobacter_MED193	Alphaproteobacteria ; Rhodobacterales
Roseobacter_SK_209_2_6	Alphaproteobacteria ; Rhodobacterales
Roseovarius_217	Alphaproteobacteria ; Rhodobacterales
Roseovarius_HTCC2601	Alphaproteobacteria ; Rhodobacterales
O_alexandrii_HTCC2633	Alphaproteobacteria ; Rhodobacterales
L_vestfoldensis_SKA53	Alphaproteobacteria ; Rhodobacterales
Sulfitobacter_EE36	Alphaproteobacteria ; Rhodobacterales
Sulfitobacter_NAS14_1	Alphaproteobacteria ; Rhodobacterales
EBACred_25D05	Alphaproteobacteria ; possible Roseobacter
Jannaschia_CCS1	Alphaproteobacteria ; Rhodobacterales
Rhodobacterales_HTCC2150	Alphaproteobacteria ; Rhodobacterales ; RCA clade
AntFos_24C04	Alphaproteobacteria ; Roseobacter (best hit to Silicibacter pom
HF0070_10D05	Alphaproteobacteria ; Rhodobacterales ; Roseobacter-like
HF0010_04M21	Alphaproteobacteria ; Rhodobacterales ; Roseobacter-like
EB080_L02BH02	Alphaproteobacteria ; Rhodobacterales ; (best hit to Silicibacter
EB080_L06A09	Alphaproteobacteria ; Rhodobacterales ; NAC11-7
alpha_HTCC2255	Alphaproteobacteria ; Rhodobacterales ; OM42 clade
EB080_L27A02	Alphaproteobacteria ; Rhodobacterales ; NAC11-7
EB080_L11F12	Alphaproteobacteria ; Rhodobacterales ; NAC11-7
EB080_L43F0B	Alphaproteobacteria ; Rhodobacterales ; NAC11-7
EB055B11	Alphaproteobacteria ; Rhodobacter-like
HF0010_10C01	Alphaproteobacteria ; Rhodobacterales ; SAR102 , SAR102
EB080_L58F04	Alphaproteobacteria ; Rhodobacterales ; CHAB1-5
EB060D04	Alphaproteobacteria ; Roseobacter-like
EB080_L84F03	Alphaproteobacteria ; Roseobacter

b)

HF70_03O15	Gammaproteobacteria ; SAR86 cluster ; CHA8-I-7 clade
HF0010_11K06	Gammaproteobacteria ; SAR86 cluster ; SAR89 clade
E8ACred_20E09	Gammaproteobacteria ; SAR86-I
HOT4E07	Gammaproteobacteria ; SAR86-I
HF0010_11823	Gammaproteobacteria ; SAR86 cluster ; CHA8-I-7 clade
HF0130_23I23	Gammaproteobacteria ; SAR86 cluster ; CHA8-I-7 clade
E80_31A08	Gammaproteobacteria ; SAR86-II
E8000_45806	Gammaproteobacteria ; SAR86-II
E8080_02D08	Gammaproteobacteria ; SAR-86-II, syntenic and high ID
E8AC_27G05	Gammaproteobacteria ; SAR86-III

c)

Pelagibacter_HTCC1062	Alphaproteobacteria ; Rickettsiales ; SAR11 clade ; subgroup 1a
Pelagibacter_HTCC1002	Alphaproteobacteria ; Rickettsiales ; SAR11 clade ; subgroup 1a
HF4000_384_009C18	Alphaproteobacteria ; Rickettsiales ; SAR11 cluster
EBAC750_11E01	Alphaproteobacteria ; SAR11 ; subgroup 2
HF4000_37C10	Alphaproteobacteria ; SAR11
EB750_09G06	Alphaproteobacteria ; looks like SAR11
HF0770_37D02	Alphaproteobacteria ; Rickettsiales ; SAR11 cluster ; Pelagibacter

136

Figure S8. Mixed layer depth (MLD) over the sampling period, with hybridized samples indicated. MLD was calculated as the first depth (≥10m) with >0.1 deg C difference from the previous meter (per MBARI BOG group, Reiko Michisaki, pers. comm.). X-axis indicates sampling date in continuous numbered days since Jan. 01, 2000, and y-axis indicates depth. Dashed red line highlights 30m depth. Trendline shows moving averageof MLD with period of 2. The MLD at this location is typically deepest in the winters and shallowest toward the end of the spring/summer upwelling season. 30m samples were both within and below the ML, and the site shows high MLD variability.

a)

b)
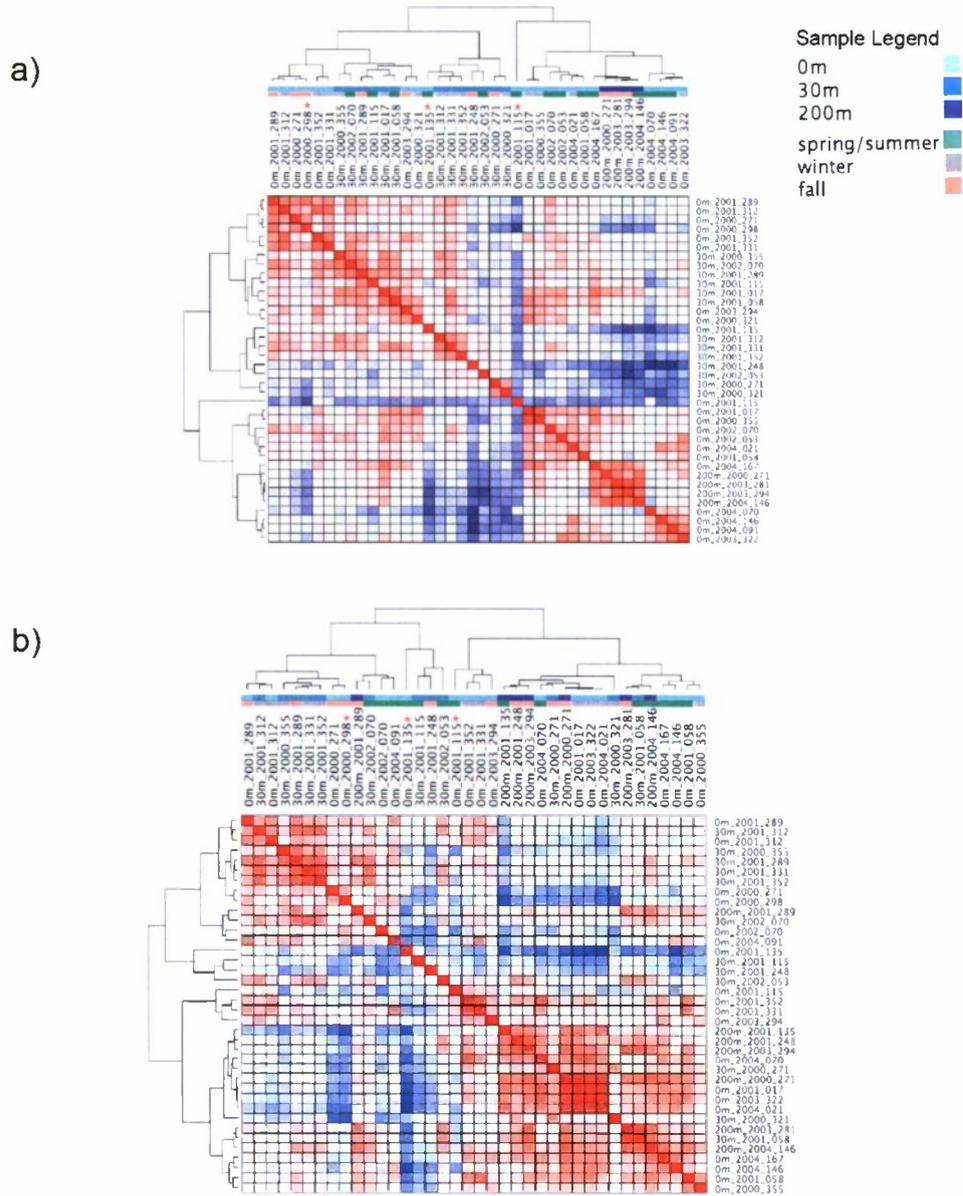
Figure S9. Population heterogeneity in all targets occurring in all three pyroseqeunced samples. Sample colors indicate depth and oceanographic season, as in color sample legend in (a). Red stars next to samples across the top denote the three pyrosequenced samples. (a) EB000_31A08. (b) EB000_39F01. (c) EB000_41B09. (d) EB000_55B11. (e) EB080_02D08. (f) EB080_L08E11. (g) EB080_L11F12. (h). EB080_L27A02. (i). EB080_L43F08.

c)

d)

139
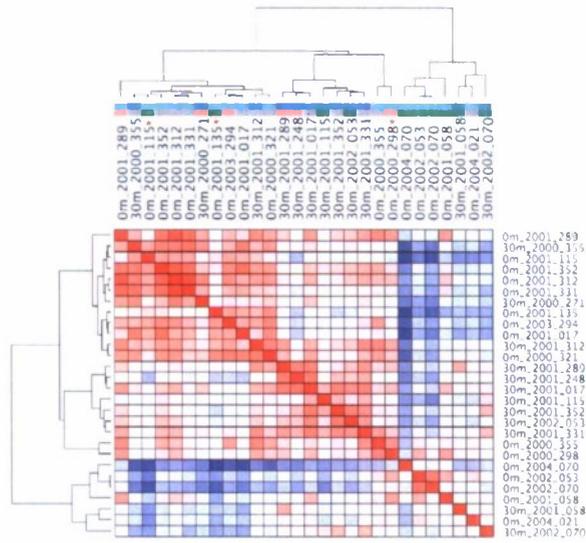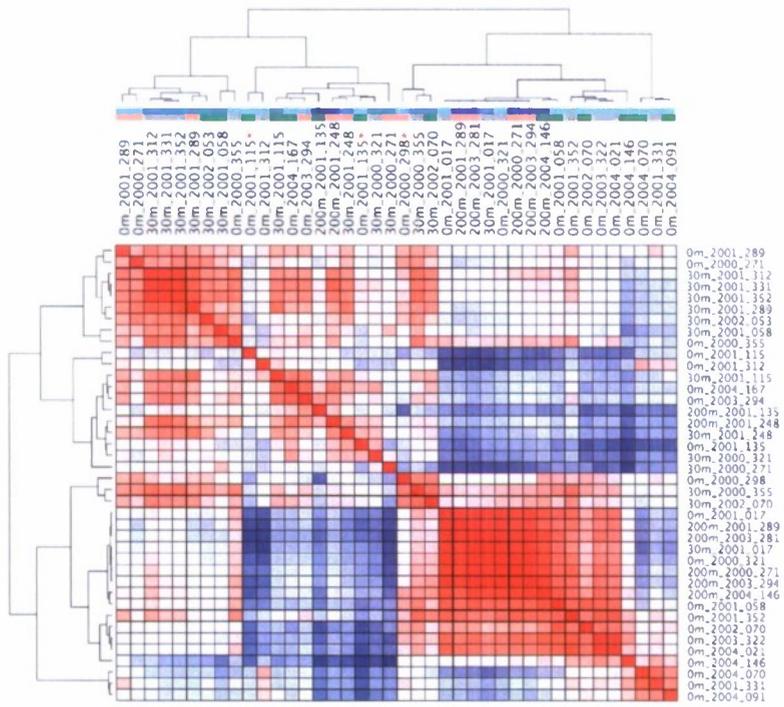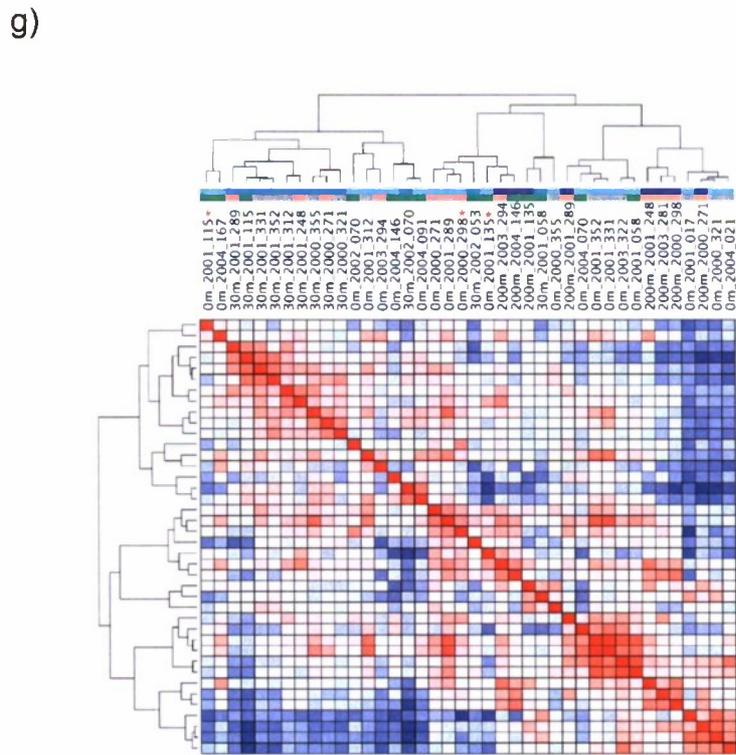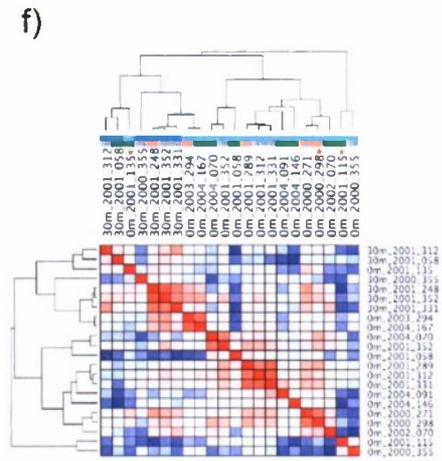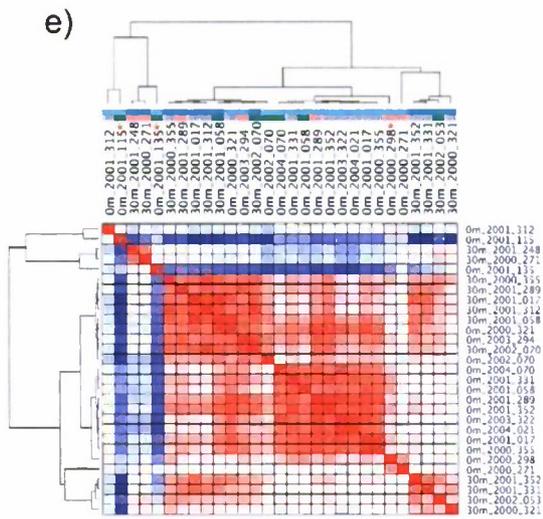
h)
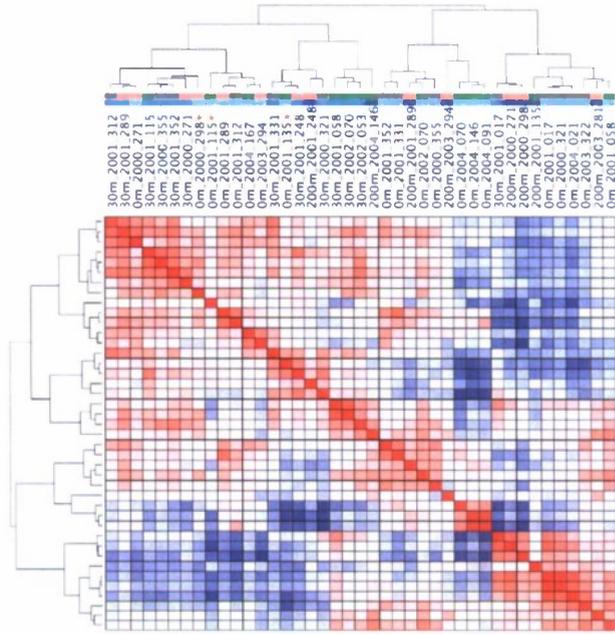
i)
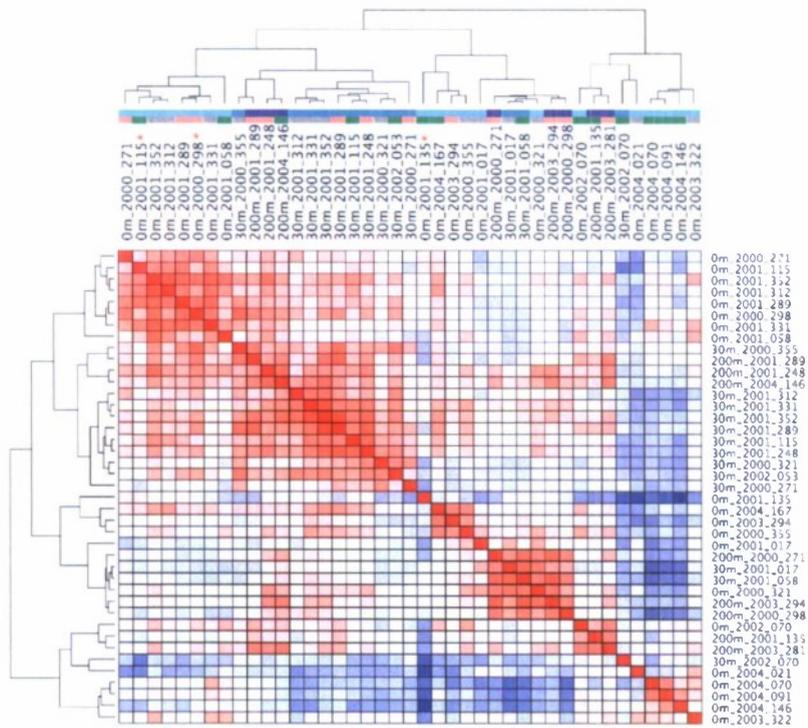
Chapter 4

Conclusions and Future Directions

In this thesis, a new tool for profiling marine microbial communities was developed and validated. It was then applied to study time-series ecology of marine microbial communities in Monterey Bay, CA.

The prototype genome proxy array (Chapter 2) targeted thirteen environmental genome fragments derived from Monterey Bay, the Hawaii Ocean Time-series station ALOHA, and Antarctic coastal waters, as well as three 8-kb regions of the *Prochlorococcus* MED4 genome. Multiple (n=20-60) 70-mer oligonucleotide probes targeted each genome fragment or genome, and were distributed along each ~40-160kbp contiguous genomic region. When hybridized to targets or target relatives, the array correctly identified the presence or absence of each, and could discern related non-target genotypes. It showed minimal cross-hybridization to organisms with ≤~75% ANI to the targets. When target cells and their relatives were spiked into a background of seawater, the array's discriminatory ability did not diminish, and the array-based organism intensity correlated linearly to cell numbers across six orders of magnitude (R^2 of 1.0). A related strain (86% ANI to target) also showed a linear correlation of signal to concentration (R^2 of 0.9999). The limit of detection in a natural background was 0.1% of the community for targeted genotypes, and 1% of the community for their cross-hybridizing relatives. Cross-hybridizing genotypes produced distinct hybridization patterns across each target probe set, allowing related strains to be distinguished.

Having developed and validated the prototype, an expanded version was developed and constructed targeting 268 genotypes, representing all major marine microbial clades, and spanning the relevant intra-clade variability among abundant groups where possible (Chapter 3). This array was used to profile 57 samples collected over four years in Monterey Bay (MB), at three oceanographically distinct depths (the photic zone, just below the mixed layer, and the subphotic zone, 0m, 30m and 200m respectively), as well as a single

depth profile (25m, 75m, 125m and 500m) from Hawaii. Three MB 0m samples were additionally pyrosequenced to allow for cross-comparison with array data. This cross-comparison showed a strong linear correlation between a targeted genotype's array signal and its metagenomic abundance (R^2=0.85-0.91 for the three samples). In addition, the Monterey Bay targets produced *in silico* hybridization to 1.9% - 2.5% of the pyrosequence reads in the three samples.

Based on similarity of array profiles, MB samples clustered into shallow (0m and 30m) and deep (200m) samples. ~35% of targeted genotypes were present in one or more MB sample, and the majority showed depth-specific distributions. Targeted clades expected to be present in this environment (e.g. SAR11, SAR86, etc) showed signal on the array, with depth distributions generally consistent with previous observations of each clade at this site and elsewhere. A relatively small number of taxa accounted for much of the variability between profiles from different depths. Reflecting the highly variable and dynamic mixed layer depth at this site, 30m samples did not cluster separately from 0m samples. In addition, no correlation to oceanographic season was observed among profiles, although bloom and post-bloom signatures were evident as increased profile intensities in 0m samples and in decoupled nutrient correlations to sample variability in 200m samples. Together with oceanographic data, array-based population insights are allowing us to work towards identifying ecotypes in poorly-understood marine clades.

The expanded genome proxy array has begun to realize its potential for high-throughput profiling of marine samples, tracking a larger number of targets simultaneously at higher phylogenetic resolution than has been previously possible with other methods.

In addition to the ongoing analyses of Chapter 3 data mentioned in its concluding paragraphs, there are several other current or future projects involving the genome proxy array that may be worth pursuing. These fall into two

categories: protocol improvements and application to particular research questions in marine microbial community ecology.

There are a number of protocol improvements that might ideally be made with this array platform. In particular, the random-primed A/B/C PCR-based amplification and labeling method is laborious, requiring several very long days for each set of reactions and several physically-challenging steps. We have discussed testing MDA versus A/B/C PCR, and the use of MDA with microarrays has been described in the literature (e.g. Wu et al 2006 for MDA). The relative biases and skewing of MDA remain unclear and several studies indicate that they are not insubstantial. Therefore, before switching array protocols to MDA-based amplification and labeling, a comprehensive set of comparisons would be required. Specifically, a single sample would be treated in three different ways. 1. Direct pyrosequencing. 2. Optimized A/B/C amplification followed by pyrosequencing. 3. Optimized MDA amplification followed by pyrosequencing. If it were important to save financial resources, a single 454 run could be used to address this. Although the depth of coverage in these diverse communities is often not high even for dominant taxa, use of a bloom sample might allow a single 454 run to be split and used for multiple samples.

Another protocol improvement would focus on better consistency of slide PLL coating. Despite improvements, some PLL-coated slide batches exhibit considerable surface irregularities and peeling, which affect the quality and yield of data. Manual removal of poor-quality spots is time-consuming, and for particularly bad arrays additional hybridizations must be performed. A number of coated slides for array printing are commercially available and I tried several earlier in my thesis, but did not have good results. Ideally, the homemade method can be improved, because it is not particularly time-consuming or difficult and it is quite inexpensive (see Appendix 1, Reagent Worksheet, for details). We have addressed issues of water and reagent quality, washing configuration, ambient dust amelioration, drying spin force and duration, and storage. However,

surface performance remains stubbornly variable among batches, in a difficult-to-predict way prior to array use.

Lastly, a major area for continued improvement is in the data processing pipeline. For the prototype array paper, I worked with Kostas Konstantinidis, then a postdoc in the lab, to develop a series of scripts for pre-processing the array data, which is in the form of gpr files from the array scanner. For the Monterey Bay time series paper, I worked with John Eppley, currently part of the lab as a computational expert, to build a next generation script that consolidated tasks into a single workflow. During the switch from prototype to expanded array, Dr. Eppley and I refined several of the preprocessing steps, as described in the methods of Chapter 3. However, there remains room for further optimization. Issues worth particular continued consideration include: i) background subtraction, ii) array-to-array normalization, iii) genome-proxy array-specific filtering thresholds to remove spurious cross-hybridization.

For the Monterey Bay data analysis, I began using an open-source software package initially designed by the Broad Institute for array-based gene expression analysis. This software, GenePattern (mentioned in the methods of Ch. 3), is module-based and designed for tailoring to novel needs. Several of its existing modules are useful for examining pre-processed data from the genome proxy array, and modules can be pipelined together, with an archiving option for documenting and preserving specific combinations of analyses and parameters. Dr. Eppley has written the data preprocessing script to be compatible with the GenePattern architecture, to allow it to easily be added to GenePattern as a module. This will facilitate non-expert use, rapid testing of a range pre-processing parameters, and direct porting of output data into other analysis modules. This will greatly help further optimizing the data processing steps mentioned above. In addition, the two R-based ecological analysis tools used for exploring correlations between array data and environmental parameters will be relatively straightforward to convert into GenePattern modules, since GenePattern is fully

compatible with the R language. These changes will continue to streamline the analysis pipeline for array data.

Among ongoing and future applications of the array, a deeper examination of Hawaiian samples, for comparison to MB samples and to investigate time series dynamics at Station ALOHA, is being performed by a graduate student in the lab, Laure-Anne Ventouras. My Hawaii hybridizations from not only HOT179 but also several other Hawaii profiles, using both the prototype and the expanded array, produced generally good results and the presence of expected taxa. These open ocean samples produced markedly different profiles than the coastal Monterey Bay samples, as expected. As mentioned in Chapter 3, however, some of the hybridizations produced weak signal which, when using the same data processing parameters optimized for MB samples, resulted in very few targets being called "present". Whether this is due to inaccurate DNA quantification, the presence of contaminants, hybridization irregularities, or the need to re-tailor processing parameters for the new site, remains unclear, and needs to be answered to maximize the utility of the array for cross-habitat comparative studies.

In addition to the purely array-based profiling of HOT samples, pyrosequence data are available for several of these samples and a comparison of array and pyrosequence data is ongoing. Working with another graduate student in the lab, Yanmei Shi (a co-author on the Ch. 3 manuscript), we cross-compared pyrosequence and array data for HOT179 75m. The correlations between target abundance by pyrosequence and by array intensity are in the same range as seen for Monterey Bay pyrosequence cross-validations. Using the improved and rapid pyrosequence-to-array pipeline developed for the MB datasets, this 75m sample will be re-examined, together with the 25m, 125m, and 500m HOT179 datasets. In addition, it is likely that deeper investigations of the signals present in the array profiles versus the sequence datasets will be appropriate for these samples, to examine and validate e.g. population variability,

as is underway for the MB datasets. The first stage of this cross-comparison, the correlation of target abundances produced by the two methods, will likely be included in the manuscript of Chapter 3, since the array profiles of these samples are compared therein to the MB samples. The second, more specific cross-comparison of populations may be most usefully performed as part of the Hawaii time-series profiling paper, the work currently being undertaken by Ms. Ventouras.

Comparing the open-ocean Hawaii array profiles with the coastal Monterey Bay profiles poses the question of the habitat specificity of the genome proxy array. Can it only be reliably used at locations from which its targets are derived? Based on existing data, the answer is: No, the array can be an effective cross-location platform. Although the majority of signal at each location came from clones derived from that location, this could be interpreted in two ways. Hawaii versus Monterey samples may represent generically distinct open ocean versus coastal communities, and the clone-derivation-to-signal pattern may simply represent this overall habitat difference. Alternatively, this pattern could be indicative of site rather than habitat specificity; that is, another set of coastal versus open ocean samples would not show appreciable signal, and/or would not show the same inversion of clone-derivation-to-signal in the two habitats. The former explanation seems most likely, for several reasons. First, the prototype array was hybridized to coastal waters from the East coast, a roughly similar habitat but very distant location to the origin of most of the targeted clones. However, a number of these Monterey Bay-derived clones produced strong signal from Woods Hole communities (see e.g. Figure 4a, Rich, Konstantinidis and DeLong, 2008, Chapter 2). Second, a number of clones produce significant array signal in samples from locations, and even habitats, very different to those of their origin. For example, the *deep-consistent* clade described in Chapter 3, whose taxa were consistently present and abundant in MB 200m samples, included two (of 10) targets derived from Hawaii, and one from 500m in the

Antarctic Polar Front. The same is not true for the *shallow-consistent* cluster, which re-emphasizes a point made in the combined clustering of Hawaii and Monterey Bay samples in Chapter 3, Figure 4c, and observed in other marine microbial studies; deep communities from very different environments can be significantly more similar than their surface counterparts. Thus, overall, there is inherently greater habitat specificity involved in probing shallow versus deep communities. However, evidence suggests this is indeed habitat rather than site specificity, which would enable the array to be effectively applied to other coastal and open ocean samples. Further samples from several diverse locations should be hybridized to the array to confirm this hypothesis. Extracted DNAs from Woods Hole, coastal Chile, coastal Oregon, Bermuda, and Antarctica are all available in the lab and so performing a suite of hybridizations would not be a major undertaking.

In addition to examining the genome-proxy array data as overall organism signals and hybridization patterns across probesets, it may also be worthwhile to investigate signals from single probes, representing single genes. For example, cross-hybridization to just one or a few probes in a probeset is considered spurious signal and is not further examined. However, strong signal to a particular gene that is not reflected in the rest of the target organism probes could reveal the importance of particular genes in some samples, with their high conservation and presence in other organisms causing their cross-hybridization. Thus, significant decoupling of the gene and organism signals could be a useful flag for processes of importance. In addition, because probes were not chosen to particular genes but rather selected based on predicted hybridization kinetics, the probes are not limited to genes with already-recognized ecological importance, and include many hypotheticals. Although the sequence divergence of the majority of genes makes strong cross-hybridization unlikely, among the ~5360 genes targeted by the array there are likely to be several interesting stories in the data already obtained from the Monterey Bay samples.

A different and interesting possible application of the genome proxy array is for hybridization to amplified and labeled community cDNA. Frias-Lopez and Shi *et al*. (2008) have optimized a protocol for preferentially amplifying community mRNA from marine samples. Working with Yanmei Shi and Laure-Anne Ventouras, we amplified and reverse transcribed community RNA from a Hawaii sample, HOT179_75m, and Ms. Ventouras and I performed side-by-side replicated DNA and cDNA array hybridizations. Following the logic of community metatranscriptomics in e.g. Frias-Lopez and Shi et al. (2008), in order to interpret transcript abundance it is necessary to also measure gene abundance, in order to normalize expression to gene copy number. A first-pass analysis of the data using the existing array pipeline showed, unsurprisingly, that overall organism-based expression levels were low, when RNA hybridization data were treated identically by the script and averaged across all probes for an organism. However, high expression across all or most of a genome fragment's randomly-targeted genes would not be expected. When genes were examined on an individual basis, a small number of genes, primarily involved in housekeeping and core metabolic functions, had high array-based expression signal. When cross-compared to the overall pyrosequencing-based metatranscriptomics analysis, these genes were not among those with the highest expression levels. However, since the metatranscriptomics analysis was global in scope and the array only targets a small fraction of all community genes, an additional step is required of using just the array targets to recruit metatranscriptomic reads, as was done for the DNA-based array-versus-pyrosequencing comparison described in Chapter 3. That pipeline was not available at the time of the RNA hybridization and the matter has not been revisited, but is worth doing. Due to the small representation of total gene space on the array, this particular array platform would not be a good primary tool for profiling community expression, but might provide useful secondary expression information.

An alternative form of microarray, related to the genome proxy array, is an

environmental clone library array. This library array would be akin to the community genome arrays described in chapters 1 and 2, but instead of each probe spot being a whole genome of a cultivated or isolated microbe, it would be a cloned genome fragment from the environment. This bypasses the need for sequencing clones before they are targeted. Indeed, this method has recently been used as a library-screening tool (e.g. Soule *et al.*, 2006), with thousands of clone library members arrayed and then queried with e.g. labeled PCR amplicons of genes of interest. Rather than using such arrays primarily as a library exploration tool, when the DeLong Lab already has a well-refined library macroarray production and screening pipeline worked out, these arrays would instead be used to query amplified and labeled environmental DNA, just as the genome proxy array is.

This approach has great appeal because it removes a portion of the deterministic nature of array design. For example, in order to be represented on the genome proxy array, all targeted genotypes first had to be sequenced, and both genomes and genome fragments were generally chosen for sequencing because they were already known or suspected to be involved in a process of interest. Instead, if large numbers of clones from genomic libraries were arrayed in a random fashion, then hybridization results would reveal which hitherto unrecognized clones might be highly abundant, or vary along ecological gradients. Those clones would then be targeted for sequencing and further characterization. It is even possible that for particularly abundant taxa, or highly uneven samples, multiple clones deriving from the same taxa or clade could be binned based on similar hybridization intensities, akin to the coverage- and GC-content- based binning of metagenomic data in Tyson *et al.*, 2004.

Lastly, further Monterey Bay investigations should be pursued using the genome proxy array. There is a wealth of samples available that were not included in this first study, for a variety of reasons (e.g. they were collected using different filtration methods, or from different volumes of seawater, etc.). These

samples span several more years and cover two additional stations, in the inner and outer Bay, respectively, and were sequenced with the same frequency as Station M1. In addition, there are several sets of samples from the Monterey Bay Aquarium Research Institute (MBARI) Biological Oceanography Group (BOG) cruises along the old CalCOFI (California Cooperative Oceanic Fisheries Investigations) Line 67 transect which runs from Monterey Bay to 300km offshore. This transect crosses four distinct oceanographic zones, from the seasonal upwelling band along the California coast (up to~20-50km offshore) all the way out to the California Current (170-300km offshore) (Pennington *et al.* 2007). The great majority of all these MB samples have not been extracted.

In the same vein, a fascinating question and suggestion by my co-advisor involved the possibility of hybridizing much older archived DNAs, from samples stored in ethanol or formaldehyde, from the last century. Again, Monterey Bay sits at a nexus of oceanographic and fisheries exploration, with a rich history of sample collection and observation. It exhibits a strong El Niño signature, and has shown warming over the last hundred years. Since the array hybridizes to fragmented DNA, which is amplified randomly prior to and during the labeling process, it is reasonable to think that fairly degraded DNA (e.g. <5000bp, ideally >500bp) could still hybridize reliably and meaningfully to the array. This would have to be tested with sheared DNA. In addition, it would be important to identify preservation biases in the integrity of DNA from different clades; e.g., one can imagine a GC-rich clade perhaps faring better over time. Finally and most importantly, this assumes the existence of such preserved samples, in enough numbers and with sufficient associated metadata to be useful for mapping microbial community change over the time scales involved. Further, it assumes that the guardians of such samples would be willing to give some portion to this endeavor.

In conclusion, there are a number of next-step experiments that could be undertaken to improve the existing protocols and to further define the working

scope of the array. There are also a number of research questions that the array could be a useful tool for answering, given its low cost and relatively high throughput, with good information yield per units of time, money, and DNA, and its unique attributes. It can contribute to our exploration of marine microbial communities at finer scales of both sampling and phylogenetic resolution than are practical using other methods. In addition, the array is uniquely useful for identifying populations of related genotypes, which is currently difficult to do using other methods, even metagenomics, without *a priori* expectations guiding analyses. Therefore, even as the cost of sequencing decreases, the array can be a highly complementary tool in a microbial ecologist's repertoire, by revealing features of the microbial community not readily observed with other tools.

Bibliography

Abulencia, C.B., Wyborski, D.L., Garcia, J.A., Podar, M., Chen, W., Chang, S.H. et al. (2006) Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments. *Applied and Environmental Microbiology* **72**: 3291-3301.

Acinas, S.G., Klepac-Ceraj, V., Hunt, D.E., Pharino, C., Ceraj, I., Distel, D.L., and Polz, M.F. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551-554.

Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F. (2005) PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Applied and Environmental Microbiology* **71**: 8966-8969.

Alonso, C., and Pernthaler, J. (2005) Incorporation of Glucose under Anoxic Conditions by Bacterioplankton from Coastal North Sea Surface Waters. *Applied and Environmental Microbiology* **71**: 1709-1716.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

Amann, R., and Fuchs, B.M. (2008) Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology* **6**: 339-348.

Amy, J.B., Joseph, A.M., Joseph, D.P., Phillipe, C., Fabien, J., and Wade, H.J. (2005) Microbial diversity in a Pacific Ocean transect from the Arctic to Antarctic circles. *Aquatic Microbial Ecology* **41**: 91-102.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C. et al. (2006) The Marine Viromes of Four Oceanic Regions. *PLoS Biology* **4**: e368.

Arrigo, K.R. (2005) Marine microorganisms and global nutrient cycles. *Nature* **437**: 349-355.

Asanuma, H., Rago, T.A., Collins, C.A., Chavez, F.P., and Castro, C.G. (1999) Changes in the hydrography of central California waters associated with the 1997–1998 El Niño. In. Monterey, CA: Naval Postgraduate School, p. 121pp.

Bae, J.-W., Rhee, S.-K., Park, J.R., Chung, W.-H., Nam, Y.-D., Lee, I. et al. (2005) Development and evaluation of genome-probing microarrays for monitoring lactic acid bacteria. *Applied and Environmental Microbiology* **71**: 8825–8835.

Baldwin, A., Moss, J., Pakulski, J., Catala, P., Joux, F., and Jeffrey, W. (2005) Microbial diversity in a Pacific Ocean transect from the Arctic to Antarctic circles. *Aquatic Microbial Ecology* **41**: 91-102.

Bano, N., and Hollibaugh, J.T. (2002) Phylogenetic Composition of Bacterioplankton Assemblages from the Arctic Ocean. *Applied and Environmental Microbiology* **68**: 505-518.

Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P. et al.

(2000) Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1906.

Béjà, O., Koonin, E.V., Aravind, L., Taylor, L.T., Seitz, H., Stein, J.L. et al. (2002) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. *Applied and Environmental Microbiology* **68**: 335-345.

Béjà, O., Suzuki, M.T., Heidelberg, J.F., Nelson, W.C., Preston, C.M., Hamada, T. et al. (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**: 630-633.

Béjà, O., Suzuki, M.T., Koonin, E.V., Aravind, L., Hadd, A., Nguyen, L.P. et al. (2000) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environmental Microbiology* **2**: 516-529.

Bench, S.R., Hanson, T.E., Williamson, K.E., Ghosh, D., Radosovich, M., Wang, K., and Wommack, K.E. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Applied and Environmental Microbiology* **73**: 7629-7641.

Bentley, D.R. (2006) Whole-genome re-sequencing. *Current Opinion in Genetic Development* **16**: 545-552.

Berney, M., Hammes, F., Bosshard, F., Weilenmann, H.-U., and Egli, T. (2007) Assessment and Interpretation of Bacterial Viability by Using the LIVE/DEAD BacLight Kit in Combination with Flow Cytometry. *Applied and Environmental Microbiology* **73**: 3283-3290.

Bernhard, A.E., and Field, K.G. (2000) Identification of Nonpoint Sources of Fecal Pollution in Coastal Waters by Using Host-Specific 16S Ribosomal DNA Genetic Markers from Fecal Anaerobes. *Applied and Environmental Microbiology* **66**: 1587-1594.

Bertilsson, S., Eiler, A., Nordqvist, A., and Jorgensen, N.O.G. (2007) Links between bacterial production, amino-acid utilization and community composition in productive lakes. *ISME Journal* **1**: 532-544.

Biddle, J.F., Lipp, J.S., Lever, M.A., Lloyd, K.G., Sorensen, K.B., Anderson, R. et al. (2006) Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. *Proceedings of the National Academy of Sciences USA* **103**: 3846-3851.

Binga, E., Lasken, R., and Neufeld, J. (2008) Something from (almost) nothing: The impact of multiple displacement amplification on microbial ecology. *ISME Journal* **2**: 233-241.

Bodrossy, L., Stralis-Pavese, N., Murrell, J.C., Radajewski, S., Weilharter, A., and Sessitsch, A. (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environmental Microbiology* **5**: 566-582.

Bohannan, B.J.M., and Hughes, J. (2003) New approaches to analyzing microbial biodiversity data. *Current Opinion in Microbiology* **6**: 282-287.

Bottari, B., Ercolini, D., Gatti, M., and Neviani, E. (2006) Application of FISH

technology for microbiological analysis: current state and prospects. *Applied Microbiology and Biotechnology* **73**: 485-494.

Bozdech, Z., Zhu, J., Joachimiak, M.P., Cohen, F.E., Pulliam, B., and DeRisi, J.L. (2003) Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. *Genome Biology* **4**: R9-R9.15.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J., Salamon, P., and Rohwer, F. (2004) Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* **271**: 565-574.

Brodie, E.L., DeSantis, T.Z., Joyner, D.C., Baek, S.M., Larsen, J.T., Andersen, G.L. et al. (2006) Application of a High-Density Oligonucleotide Microarray Approach To Study Bacterial Population Dynamics during Uranium Reduction and Reoxidation. *Applied and Environmental Microbiology* **72**: 6288-6298.

Brodie, E.L., DeSantis, T.Z., Parker, J.P.M., Zubietta, I.X., Piceno, Y.M., and Andersen, G.L. (2007) Urban aerosols harbor diverse and dynamic bacterial populations. *Proceedings of the National Academy of Sciences USA* **104**: 299-304.

Brown, M.V., Schwalbach, M.S., Hewson, I., Fuhrman, J.A., Brusetti, L., Borin, S. et al. (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series

Usefulness of length heterogeneity-PCR for monitoring lactic acid bacteria succession during maize ensiling. *Environmental Microbiology* **7**: 1466-1479.

Bryant, D.A., Costas, A.M., Maresca, J.A., Chew, A.G., Klatt, C.G., Bateson, M.M. et al. (2007) Candidatus Chloracidobacterium thermophilum: an aerobic phototrophic Acidobacterium. *Science* **317**: 523-526.

Buchan, A., Gonzalez, J.M., and Moran, M.A. (2005) Overview of the marine roseobacter lineage. *Applied and Environmental Microbiology* **71**: 5665-5677.

Campbell, L., Liu, H., Nolla, H.A., and Vaulot, D. (1997) Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991-1994 ENSO event. *Deep Sea Research Part I: Oceanographic Research Papers* **44**: 167-192.

Cebron, A., Bodrossy, L., Chen, Y., Singer, A.C., Thompson, I.P., Prosser, J.I., and Murrell, J.C. (2007) Identity of active methanotrophs in landfill cover soil as revealed by DNA-stable isotope probing. *FEMS Microbiology Ecology* **62**: 12-23.

Chavez, F.P., Pennington, J.T., Castro, C.G., Ryan, J.P., Michisaki, R.P., Schlining, B. et al. (2002) Biological and chemical consequences of the 1997-1998 El NiÒo in central California waters. *Progress In Oceanography* **54**: 205-232.

Chavez, F.P., Ryan, J., Lluch-Cota, S.E., and Niquen C, M. (2003) From

Anchovies to Sardines and Back: Multidecadal Change in the Pacific Ocean. *Science* **299**: 217-221.

Chisholm, S., Olson, R., Zettler, E., Goericke, R., Waterbury, J., and Welschmeyer, N. (1988) A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334**: 340-343.

Cho, J.-C., and Giovannoni, S.J. (2004) Cultivation and Growth Characteristics of a Diverse Group of Oligotrophic Marine Gammaproteobacteria. *Applied and Environmental Microbiology* **70**: 432-440.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., McGarrell, D.M. et al. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Research* **35**: D169-172.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006) Genomic islands and the ecology and evolution of Prochlorococcus. *Science* **311**: 1768-1770.

Connon, S.A., and Giovannoni, S.J. (2002) High-Throughput Methods for Culturing Microorganisms in Very-Low-Nutrient Media Yield Diverse New Marine Isolates. *Applied and Environmental Microbiology* **68**: 3878-3885.

Cottrell, M.T., and Kirchman, D.L. (2003) Contribution of major bacterial groups to bacterial biomass production (thymidine and leucine incorporation) in the Delaware estuary. *Limnology and Oceanography* **48**: 168–178.

Daims, H., Lucker, S., and Wagner, M. (2006) daime, a novel image analysis program for microbial ecology and biofilm research. *Environmental Microbiology* **8**: 200-213.

Dalsgaard, T., Canfield, D.E., Petersen, J., Thamdrup, B., and Acuna-Gonzalez, J. (2003) N2 production by the anammox reaction in the anoxic water column of Golfo Dulce, Costa Rica. *Nature* **422**: 606-608.

de la Torre, J.R., Christianson, L.M., Beja, O., Suzuki, M.T., Karl, D.M., Heidelberg, J., and DeLong, E.F. (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proceedings of the National Academy of Sciences USA* **100**: 12830-12835.

DeLong, E.F. (2005) Microbial community genomics in the ocean. *Nature Reviews Microbiology* **3**: 459-469.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.U. et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496-503.

Denef, V.J., Shah, M.B., Verberkmoes, N.C., Hettich, R.L., and Banfield, J.F. (2007) Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. *Journal of Proteome Research* **6**: 3152-3161.

Derakshani, M., Lukow, T., and Liesack, W. (2001) Novel bacterial lineages at the (sub)division level as detected by signature nucleotide-targeted recovery of 16S rRNA genes from bulk soil and rice roots of flooded rice microcosms. *Applied and Environmental Microbiology* **67**: 623-631.

DeSantis, T., Brodie, E., Moberg, J., Zubieta, I., Piceno, Y., and Andersen, G. (2007) High-Density Universal 16S rRNA Microarray Analysis Reveals Broader Diversity than Typical Clone Library When Sampling the Environment. *Microbial Ecology* **53**: 371-383.

DeSantis, T.Z., Stone, C.E., Murray, S.R., Moberg, J.P., and Andersen, G.L. (2005) Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiology Letters* **245**: 271–278.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M. et al. (2008) Functional metagenomic profiling of nine biomes. *Nature* **452**: 629-632.

Dinsdale, E.A., Pantos, O., Smriga, S., Edwards, R.A., Angly, F., Wegley, L. et al. (2008) Microbial Ecology of Four Coral Atolls in the Northern Line Islands. *PLoS ONE* **3**: e1584.

Dubelaar, G.B.J., Casotti, Rafaella , Tarran, Glen A., Biegala, Isabelle C. (2007) Phytoplankton and their analysis by flow cytometry. In *Flow Cytometry with Plant Cells: Analysis of Genes, Chromosomes and Genomes.* Doležel, J., Greilhuber, J., and Suda, J. (eds). Weinheim: Wiley-VCH, pp. 287-322.

DuRand, M.D., Olson, R.J., and Chisholm, S.W. (2001) Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **48**: 1983-2003.

Eek, K.M., Sessions, A.L., and Lies, D.P. (2007) Carbon-isotopic analysis of microbial cells sorted by flow cytometry. *Geobiology* **5**: 85-95.

Eilers, H., Pernthaler, J., Glockner, F.O., and Amann, R. (2000) Culturability and In Situ Abundance of Pelagic Bacteria from the North Sea. *Applied and Environmental Microbiology* **66**: 3044-3051.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* **95**: 14863-14868.

El Fantroussi, S., Urakawa, H., Bernhard, A.E., Kelly, J.J., Noble, P.A., Smidt, H. et al. (2003) Direct Profiling of Environmental Microbial Populations by Thermal Dissociation Analysis of Native rRNAs Hybridized to Oligonucleotide Microarrays. *Applied and Environmental Microbiology* **69**: 2377-2382.

Eppley, J.M., Tyson, G.W., Getz, W.M., and Banfield, J.F. (2007) Strainer: software for analysis of population variation in community genomic datasets. *BMC Bioinformatics* **8**: 398.

Eyers, L., Smoot, J., Bugli, C., Urakawa, H., Murray, Z., Siripong, S. et al. (2006) Discrimination of shifts in a soil microbial community associated with TNT-contamination using a functional ANOVA of 16S rRNA hybridized to oligonucleotide microarrays. *Environmental Science and Technology* **40**: 5867-5873.

Falkowski, P.G., and de Vargas, C. (2004) Shotgun sequencing in the sea: A blast from the past? *Science* **304**: 58-60.

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R. et al. (2007) Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil. *Applied and Environmental Microbiology* **73**: 7059-7066.

Fisher, M.M., and Triplett, E.W. (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Applied and Environmental Microbiology* **65**: 4630-4636.

Follows, M.J., Dutkiewicz, S., Grant, S., and Chisholm, S.W. (2007) Emergent Biogeography of Microbial Communities in a Model Ocean. *Science* **315**: 1843-1846.

Frank, D.N., and Pace, N.R. (2008) Gastrointestinal microbiology enters the metagenomics era. *Current Opinion in Gastroenterology* **24**: 4-10.

Frias-Lopez, J., Shi, Y., Tyson, G.W., Coleman, M.L., Schuster, S.C., Chisholm, S.W., and Delong, E.F. (2008) Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences USA* **105**: 3805-3810.

Frigaard, N.U., Martinez, A., Mincer, T.J., and DeLong, E.F. (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**: 847-850.

Fuhrman, J.A., Hewson, I., Schwalbach, M.S., Steele, J.A., Brown, M.V., and Naeem, S. (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences USA* **103**: 13104-13109.

Galbraith, E.A., D.A. Antonipouluos, B.A. White (2004) Suppressive subtractive hybridization as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model. *Environmental Microbiology* **6**: 928-937.

Galperin, M.Y. (2004) Metagenomics: from acid mine to shining sea. *Environmental Microbiology* **6**: 543-545.

Gao, H., Yang, Z.K., Gentry, T.J., Wu, L., Schadt, C.W., and Zhou, J. (2007) Microarray-Based Analysis of Microbial Community RNAs by Whole-Community RNA Amplification. *Applied and Environmental Microbiology* **73**: 563-571.

Garcia Martin, H., Ivanova, N., Kunin, V., Warnecke, F., Barry, K.W., McHardy, A.C. et al. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nature Biotechnology* **24**: 1263-1269.

Garrido, P., Gonzalez-Toril, E., Garcia-Moyano, A., Moreno-Paz, M., Amils, R., and Parro, V. (2008) An oligonucleotide prokaryotic acidophile microarray: its validation and its use to monitor seasonal variations in extreme acidic environments with total environmental RNA. *Environmental Microbiology*

10: 836-850.

Gasol, J.M., Zweifel, U.L., Peters, F., Fuhrman, J.A., and Hagstrom, A. (1999) Significance of Size and Nucleic Acid Content Heterogeneity as Measured by Flow Cytometry in Natural Planktonic Bacteria. *Applied and Environmental Microbiology* **65**: 4475-4483.

Gebert, J., Stralis-Pavese, N., Alawi, M., and Bodrossy, L. (2008) Analysis of methanotrophic communities in landfill biofilters using diagnostic microarray. *Environmental Microbiology* **10**: 1175-1188.

Gentry, T.J., Wickham, G.S., Schadt, C.W., He, Z., and Zhou, J. (2006) Microarray applications in microbial ecology research. *Microbial Ecology* **52**: 159-175.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S. et al. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* **312**: 1355-1359.

Giovannoni, S., and Stingl, U. (2007) The importance of culturing bacterioplankton in the 'omics' age. *Nature Reviews Microbiology* **5**: 820-826.

Giovannoni, S.J., Bibbs, L., Cho, J.C., Stapels, M.D., Desiderio, R., Vergin, K.L. et al. (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**: 82-85.

Giovannoni, S.J., Hayakawa, D.H., Tripp, H.J., Stingl, U., Givan, S.A., Cho, J.C. et al. (2008) The small genome of an abundant coastal ocean methylotroph. *Environmental Microbiology* **10**: 1771-1782.

Giovannoni, S.J., and Rappé, M.S. (2000) Evolution, diversity and molecular ecology of marine prokaryotes. In *Microbial Ecology of the Oceans*. Kirchman, D.L. (ed): Wiley and Sons, pp. 47-84.

Goffredi, S.K., Wilpiszeski, R., Lee, R., and Orphan, V.J. (2008) Temporal evolution of methane cycling and phylogenetic diversity of archaea in sediments from a deep-sea whale-fall in Monterey Canyon, California. *ISME Journal* **2**: 204-220.

Goldberg, S.M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R. et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences USA* **103**: 11240-11245.

Gomez-Consarnau, L., Gonzalez, J.M., Coll-Llado, M., Gourdon, P., Pascher, T., Neutze, R. et al. (2007) Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* **445**: 210-213.

Gonzalez, J.M., Fernandez-Gomez, B., Fernandez-Guerra, A., Gomez-Consarnau, L., Sanchez, O., Coll-Llado, M. et al. (2008) Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria). *Proceedings of the National Academy of Sciences USA* **105**: 8724-8729.

Greene, E.A., and G. Voordouw (2003) Analysis of environmental microbial

communities by reverse sample genome probing. *Journal of Microbiological Methods* **53**: 211-219.

Gruden, C., Skerlos, S., and Adriaens, P. (2004) Flow cytometry for microbial sensing in environmental sustainability applications: current status and future prospects. *FEMS Microbiology Ecology* **49**: 37-49.

Grzymski, J.J., Carter, B.J., DeLong, E.F., Feldman, R.A., Ghadiri, A., and Murray, A.E. (2006) Comparative genomics of DNA fragments from six Antarctic marine planktonic bacteria. *Applied and Environmental Microbiology* **72**: 1532-1541.

Hahn, M.W., and Pöckl, M. (2005) Ecotypes of planktonic actinobacteria with identical 16S rRNA genes adapted to thermal niches in temperate, subtropical, and tropical freshwater habitats. *Applied and Environmental Microbiology* **71**: 766-773.

Hallam, S.J., Konstantinidis, K.T., Putnam, N., Schleper, C., Watanabe, Y.-i., Sugahara, J. et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. *Proceedings of the National Academy of Sciences USA* **103**: 18296-18301.

Hallam, S.J., Putnam, N., Preston, C.M., Detter, J.C., Rokhsar, D., Richardson, P.M., and DeLong, E.F. (2004) Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science* **305**: 1457-1462.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *MMBR* **68**: 669-685.

Hansel, C.M., and Francis, C.A. (2006) Coupled Photochemical and Enzymatic Mn(II) Oxidation Pathways of a Planktonic Roseobacter-Like Bacterium. *Applied and Environmental Microbiology* **72**: 3543-3549.

Hardenbol, P., Baner, J., Jain, M., Nilsson, M., Namsaraev, E.A., Karlin-Neumann, G.A. et al. (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology* **21**: 673-678.

Hartmann, M., and Widmer, F. (2008) Reliability for detecting composition and changes of microbial communities by T-RFLP genetic profiling. *FEMS Microbiology Ecology* **63**: 249-260.

He, Z., Gentry, T.J., Schadt, C.W., Wu, L., Liebich, J., Chong, S.C. et al. (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME Journal* **1**: 67-77.

He, Z., Wu, L., Li, X., Fields, M.W., and Zhou, J. (2005) Empirical establishment of oligonucleotide probe design criteria. *Applied and Environmental Microbiology* **71**: 3753–3760.

Hewson, I., Steele, J., Capone, D., and Fuhrman, J. (2006) Remarkable heterogeneity in meso- and bathypelagic bacterioplankton assemblage composition. *Limnology and Oceanography* **51**: 1274-1283.

Hewson, I., Steele, J.A., Capone, D.G., and Fuhrman, J.A. (2006) Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. *Marine Ecology Progress Series* **311**: 67-77.

Hinojosa, M.B., Carreira, J.A., Garcia-Ruiz, R., and Dick, R.P. (2005) Microbial

Response to Heavy Metal-Polluted Soils: Community Analysis from Phospholipid-Linked Fatty Acids and Ester-Linked Fatty Acids Extracts. *Journal of Environmental Quality* **34**: 1789-1800.

Horz, H.-P., Rich, V., Avrahami, S., and Bohannan, B.J.M. (2005) Methane-Oxidizing Bacteria in a California Upland Grassland Soil: Diversity and Response to Simulated Global Change. *Applied and Environmental Microbiology* **71**: 2642-2652.

Howard, E.C., Henriksen, J.R., Buchan, A., Reisch, C.R., Burgmann, H., Welsh, R. et al. (2006) Bacterial Taxa That Limit Sulfur Flux from the Ocean. *Science* **314**: 649-652.

Huber, H., Hohn, M.J., Rachel, R., Fuchs, T., Wimmer, V.C., and Stetter, K.O. (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**: 63-67.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., and Sogin, M.L. (2007) Microbial population structures in the deep marine biosphere. *Science* **318**: 97-100.

Hughes, J.B., Hellmann, J.J., Ricketts, T.H., and Bohannan, B.J.M. (2001) Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. *Applied and Environmental Microbiology* **67**: 4399-4406.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**: R143.

Hutchison, C.A., 3rd, Smith, H.O., Pfannkoch, C., and Venter, J.C. (2005) Cell-free cloning using phi29 DNA polymerase. *Proceedings of the National Academy of Sciences USA* **102**: 17332-17336.

Huyghe, A., Francois, P., Charbonnier, Y., Tangomo-Bento, M., Bonetti, E.-J., Paster, B.J. et al. (2008) Novel Microarray Design Strategy To Study Complex Bacterial Communities. *Applied and Environmental Microbiology* **74**: 1876-1885.

Inagaki, F., Nunoura, T., Nakagawa, S., Teske, A., Lever, M., Lauer, A. et al. (2006) Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. *Proceedings of the National Academy of Sciences USA* **103**: 2815-2820.

Ingalls, A.E., Shah, S.R., Hansman, R.L., Aluwihare, L.I., Santos, G.M., Druffel, E.R.M., and Pearson, A. (2006) Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proceedings of the National Academy of Sciences USA* **103**: 6442-6447.

Janssen, P.H. (2006) Identifying the Dominant Soil Bacterial Taxa in Libraries of 16S rRNA and 16S rRNA Genes. *Applied and Environmental Microbiology* **72**: 1719-1728.

Jaspers, E., and Overmann, J. (2004) Ecological significance of microdiversity: Identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Applied and Environmental Microbiology* **70**: 4831-4839.

Jimenez Esquilin, A.E., Stromberger, M.E., and Shepperd, W.D. (2008) Soil Scarification and Wildfire Interactions and Effects on Microbial Communities and Carbon. *Journal of the American Society for Soil Sciences* **72**: 111-118.

Johnson*, Z.I., Zinser*, E.R., Coe, A., McNulty, N.P., Woodward, E.M.S., and Chisholm, S.W. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-Scale environmental gradients. *Science* **311**: 1737-1740.

Kalyuzhnaya, M.G., Zabinsky, R., Bowerman, S., Baker, D.R., Lidstrom, M.E., and Chistoserdova, L. (2006) Fluorescence In Situ Hybridization-Flow Cytometry-Cell Sorting-Based Method for Separation and Enrichment of Type I and Type II Methanotroph Populations. *Applied and Environmental Microbiology* **72**: 4293-4301.

Karl, D.M. (2007) Microbial oceanography: paradigms, processes and promise. *Nat Rev Micro* **5**: 759-769.

Karl, D.M., and Lukas, R. (1996) The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep Sea Research Part II: Topical Studies in Oceanography* **43**: 129-156.

Karner, M.B., DeLong, E.F., and Karl, D.M. (2001) Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**: 507-510.

Kemp, P.F., Lee, S., and Laroche, J. (1993) Estimating the Growth Rate of Slowly Growing Marine Bacteria from RNA Content. *Applied and Environmental Microbiology* **59**: 2594-2601.

Kennedy, J., Marchesi, J., and Dobson, A. (2007) Metagenomic approaches to exploit the biotechnological potential of the microbial consortia of marine sponges. *Applied Microbiology and Biotechnology* **75**: 11-20.

Kent, A.D., Smith, D.J., Benson, B.J., and Triplett, E.W. (2003) Web-based phylogenetic assignment tool for analysis of terminal restriction fragment length polymorphism profiles of microbial communities. *Applied and Environmental Microbiology* **69**: 6768-6776.

Kent, A.D., and Triplett, E.W. (2002) Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annual Review of Microbiology* **56**: 211-236.

Kent, A.D., Yannarell, A.C., Rusak, J.A., Triplett, E.W., and McMahon, K.D. (2007) Synchrony in aquatic microbial community dynamics. *ISME Journal* **1**: 38-47.

Khrapko, K., Hanekamp, J.S., Thilly, W.G., Belenkii, A., Foret, F., and Karger, B.L. (1994) Constant denaturant capillary electrophoresis (CDCE): a high resolution approach to mutational analysis. *Nucleic Acids Research* **22**: 364-369.

Kita-Tsukamoto, K., Wada, M., Yao, K., Kamiya, A., Yoshizawa, S., Uchiyama, N., and Kogure, K. (2006) Rapid identification of marine bioluminescent bacteria by amplified 16S ribosomal RNA gene restriction analysis. *FEMS Microbiology Letters* **256**: 298-303.

Klepac-Ceraj, V. (2004) Thesis: Diversity and phylogenetic structure of two complex marine microbial communities. In *Dept. of Civil and Environmental Engineering.* Cambridge: Massachusetts Institute of Technology, p. 106.

Knapp, C., Findlay, D., Kidd, K., and Graham, D. (2007) A comparative assessment of molecular biological and direct microscopic techniques for assessing aquatic systems. *Environmental Monitoring and Assessment.*

Kohler, T., Stingl, U., Meuser, K., and Brune, A. (2008) Novel lineages of Planctomycetes densely colonize the alkaline gut of soil-feeding termites (Cubitermes spp.). *Environmental Microbiology* **10**: 1260-1270.

Koizumi, Y., Kelly, J., Nakagawa, T., Urakawa, H., El-Fantroussi, S., Al-Muzaini, S. et al. (2002) Parallel characterization of anaerobic toluene- and ethylbenzene-degrading microbial consortia by PCR-denaturing gradient gel electrophoresis, RNA-DNA membrane hybridization, and DNA microarray technology. *Applied and Environmental Microbiology* **68**: 3215-3225.

Kolber, Z.S., Van Dover, C.L., Niederman, R.A., and Falkowski, P.G. (2000) Bacterial photosynthesis in surface waters of the open ocean. *Nature* **407**: 177-179.

Konstantinidis, K.T., and Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences USA* **102**: 2567-2572.

Korlach, J., Marks, P.J., Cicero, R.L., Gray, J.J., Murphy, D.L., Roitman, D.B. et al. (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of the National Academy of Sciences USA* **105**: 1176-1181.

Kostić, T., Weilharter, A., Rubino, S., Delogu, G., Uzzau, S., Rudi, K. et al. (2007) A microbial diagnostic microarray technique for the sensitive detection and identiWcation of pathogenic bacteria in a background of nonpathogens *Analytical Biochemistry* **360**: 244–254

Krause, L., Diaz, N.N., Goesmann, A., Kelley, S., Nattkemper, T.W., Rohwer, F. et al. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research* **36**: 2230-2239.

Kreil, D.P., Russell, R.R., Russell, S., and Alan Kimmel, a.B.O. (2006) Microarray Oligonucleotide Probes. In *Methods in Enzymology*: Academic Press, pp. 73-98.

Kuenen, J.G. (2008) Anammox bacteria: from discovery to application. *Nature Reviews Microbiology* **6**: 320-326.

Kunin, V., He, S., Warnecke, F., Peterson, S.B., Garcia Martin, H., Haynes, M. et al. (2008) A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Research* **18**: 293-297.

Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A. et al. (2007) Comparative metagenomics revealed commonly enriched gene

sets in human gut microbiomes. *DNA Research* **14**: 169-181.

Kuypers, M.M., Lavik, G., Woebken, D., Schmid, M., Fuchs, B.M., Amann, R. et al. (2005) Massive nitrogen loss from the Benguela upwelling system through anaerobic ammonium oxidation. *Proceedings of the National Academy of Sciences USA* **102**: 6478-6483.

Kuypers, M.M.M., Sliekers, A.O., Lavik, G., Schmid, M., Jorgensen, B.B., Kuenen, J.G. et al. (2003) Anaerobic ammonium oxidation by anammox bacteria in the Black Sea. *Nature* **422**: 608-611.

Lacerda, C.M., Choe, L.H., and Reardon, K.F. (2007) Metaproteomic analysis of a bacterial community response to cadmium exposure. *Journal of Proteome Research* **6**: 1145-1152.

Lasken, R.S. (2007) Single-cell genomic sequencing using Multiple Displacement Amplification. *Current Opinion in Microbiology* **10**: 510-516.

Lechene, C.P., Luyten, Y., McMahon, G., and Distel, D.L. (2007) Quantitative Imaging of Nitrogen Fixation by Individual Bacteria Within Animal Cells. *Science* **317**: 1563-1566.

Lee, N., Nielsen, P.H., Andreasen, K.H., Juretschko, S., Nielsen, J.L., Schleifer, K.-H., and Wagner, M. (1999) Combination of Fluorescent In Situ Hybridization and Microautoradiography---a New Tool for Structure-Function Analyses in Microbial Ecology. *Applied and Environmental Microbiology* **65**: 1289-1297.

Legault, B.A., Lopez-Lopez, A., Alba-Casado, J.C., Doolittle, W.F., Bolhuis, H., Rodriguez-Valera, F., and Papke, R.T. (2006) Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* **7**: 171.

Lehner, A., Loy, A., Behr, T., Gaenge, H., Ludwig, W., Wagner, M., and Schleifer, K.-H. (2005) Oligonucleotide microarray for identification of *Enterococcus* species. *FEMS Microbiology Letters* **246**: 133–142.

Lenaerts, J., Lappin-Scott, H.M., and Porter, J. (2007) Improved Fluorescent In Situ Hybridization Method for Detection of Bacteria from Activated Sludge and River Water by Using DNA Molecular Beacons and Flow Cytometry. *Applied and Environmental Microbiology* **73**: 2020-2023.

Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**: 837-848.

Liebich, J., Schadt, C.W., Chong, S.C., He, Z., Rhee, S.-K., and Zhou, J. (2006) Improvement of Oligonucleotide Probe Design Criteria for Functional Gene Microarrays in Environmental Applications. *Applied and Environmental Microbiology* **72**: 1688-1691.

Liesack, W., and Stackebrandt, E. (1989) Evidence for unlinked rrn operons in the Planctomycete Pirellula marina. *Journal of Bacteriology* **171**: 5025-5030.

Lim, E., Tomita, A., Thilly, W., and Polz, M. (2001) Combination of competitive

quantitative PCR and constant-denaturing capillary electrophoresis for high-resolution detection and enumeration of microbial cells. *Applied and Environmental Microbiology* **67**: 3897-3903.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D., and Knight, R. (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research* **35**: e120-.

Lopez-Garcia, P., Brochier, C., Moreira, D., and Rodriguez-Valera, F. (2004) Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environmental Microbiology* **6**: 19-34.

López-García, P., López-López, A., Moreira, D., and Rodríguez-Valera, F. (2001) Diversity of free-living prokaryotes from a deep-sea site at the Antarctic Polar Front. *FEMS Microbiology Ecology* **36**: 193-202.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J. et al. (2002) Oligonucleotide Microarray for 16S rRNA Gene-Based Detection of All Recognized Lineages of Sulfate-Reducing Prokaryotes in the Environment. *Applied and Environmental Microbiology* **68**: 5064-5081.

Loy, A., Maixner, F., Wagner, M., and Horn, M. (2007) probeBase--an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Research* **35**: D800-804.

Loy, A., Schulz, C., Lucker, S., Schopfer-Wendels, A., Stoecker, K., Baranyi, C. et al. (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the Betaproteobacterial order "*Rhodocyclales*". *Applied Environmental Microbiology* **71**: 1373-1386.

Lucchini, S., Thompson, A., and Hinton, J.C. (2001) Microarrays for microbiologists. *Microbiology* **147**: 1403-1414.

Magurran, A.E. (1988) *Ecological Diversity and Its Measurement*. Princeton, N.J., USA: Princeton University Press.

Marcelino, L.A., Backman, V., Donaldson, A., Steadman, C., Thompson, J.R., Pacocha Preheim, S. et al. (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. *Proceedings of the National Academy of Sciences USA* **103**: 13629 –13634.

Marcelino, L.A., Backman, V., Donaldson, A., Steadman, C., Thompson, J.R., Preheim, S.P. et al. (2006) Accurately quantifying low-abundant targets amid similar sequences by revealing hidden correlations in oligonucleotide microarray data. *Proceedings of the National Academy of Sciences USA* **103**: 13629-13634.

Marcy, Y., Ishoey, T., Lasken, R.S., Stockwell, T.B., Walenz, B.P., Halpern, A.L. et al. (2007) Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet* **3**: 1702-1708.

Marcy, Y., Ouverney, C., Bik, E.M., Losekann, T., Ivanova, N., Martin, H.G. et al. (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth.

Proceedings of the National Academy of Sciences USA **104**: 11889-11894.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

Marhaver, K.L., Edwards, R.A., and Rohwer, F. (2008) Viral communities associated with healthy and bleaching corals. *Environ Microbiol.*

Marsh, T., Saxman, P., Cole, J., and Tiedje, J. (2000) Terminal restriction fragment length polymorphism analysis program, a web-based research tool for microbial community analysis. *Applied and Environmental Microbiology* **66**: 3616-3620.

Martin, A.P., Zubkov, M.V., Burkill, P.H., and Holland, R.J. (2005) Extreme spatial variability in marine picoplankton and its consequences for interpreting Eulerian time-series. *Biology Letters* **1**: 366-369.

Martín-Cuadrado, A.-B., López-García, P., Alba, J.-C., Moreira, D., Monticelli, L., Strittmatter, A. et al. (2007) Metagenomics of the Deep Mediterranean, a Warm Bathypelagic Habitat. *PLoS ONE* **2**: e914.

Martinez, A., Bradley, A.S., Waldbauer, J.R., Summons, R.E., and DeLong, E.F. (2007) Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. *Proceedings of the National Academy of Sciences USA* **104**: 5590-5595.

Maruyama, F., Kenzaka, T., Yamaguchi, N., Tani, K., and Nasu, M. (2005) Visualization and Enumeration of Bacteria Carrying a Specific Gene Sequence by In Situ Rolling Circle Amplification. *Applied and Environmental Microbiology* **71**: 7933-7940.

Mary, I., Cummings, D.G., Biegala, I.C., Burkill, P.H., Archer, S.D., Zubkov, M.V. et al. (2006) Seasonal dynamics of bacterioplankton community structure at a coastal station in the western English Channel

daime, a novel image analysis program for microbial ecology and biofilm research. *Aquatic Microbial Ecology* **42**: 119-126.

Marzorati, M., Wittebolle, L., Boon, N., Daffonchio, D., and Verstraete, W. (2008) How to get more out of molecular fingerprints: practical tools for microbial ecology. *Environmental Microbiology* **10**: 1571-81.

Massana, R., Murray, A.E., Preston, C.M., and DeLong, E.F. (1997) Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Applied and Environmental Microbiology* **63**: 50-56.

McCarren, J., and DeLong, E.F. (2007) Proteorhodopsin photosystem gene clusters exhibit co-evolutonary trends and shared ancestry among diverse marine microbial phyla. *Environmental Microbiology* **9**: 846-858.

McDonald, I.R., Bodrossy, L., Chen, Y., and Murrell, J.C. (2008) Molecular Ecology Techniques for the Study of Aerobic Methanotrophs. *Applied and Environmental Microbiology* **74**: 1305-1315.

McGillicuddy, D.J., Jr., Anderson, L.A., Bates, N.R., Bibby, T., Buesseler, K.O.,

Carlson, C.A. et al. (2007) Eddy/Wind Interactions Stimulate Extraordinary Mid-Ocean Plankton Blooms. *Science* **316**: 1021-1026.

Menke, M.A.O.H., Liesack, W., and Stackebrandt, E. (1991) Ribotyping of 16S and 23S rRNA genes and organization of rrn operons in members of the bacterial genera Gemmata, Planctomyces, Thermotoga, Thermus, and Verrucomicrobium. *Archives of Microbiology* **155**: 263-271.

Michel, C., Pelletier, C., Boussaha, M., Douet, D.G., Lautraite, A., and Tailliez, P. (2007) Diversity of lactic acid bacteria associated with fish and the fish farm environment, established by amplified rRNA gene restriction analysis. *Applied and Environmental Microbiology* **73**: 2947-2955.

Mills, D.K., Entry, J.A., Gillevet, P.M., and Mathee, K. (2007) Assessing Microbial Community Diversity Using Amplicon Length Heterogeneity Polymerase Chain Reaction. *Soil Sci Soc Am J* **71**: 572-578.

Mincer, T.J., Church, M.J., Taylor, L.T., Preston, C., Karl, D.M., and DeLong, E.F. (2007) Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre. *Environmental Microbiology* **9**: 1162-1175.

Moeseneder, M.M., Arrieta, J.M., Muyzer, G., Winter, C., and Herndl, G.J. (1999) Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. *Applied and Environmental Microbiology* **65**: 3518-3525.

Moisander, P.H., Morrison, A.E., Ward, B.B., Jenkins, B.D., and Zehr, J.P. (2007) Spatial-temporal variability in diazotroph assemblages in Chesapeake Bay using an oligonucleotide nifH microarray. *Environmental Microbiology* **9**: 1823-1835.

Moore, L.R., Rocap, G., and Chisholm, S.W. (1998) Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. *Nature* **393**: 464-467.

Moore-Kucera, J., and Dick, R. (2008) PLFA Profiling of Microbial Community Structure and Seasonal Shifts in Soils of a Douglas-fir Chronosequence. *Microbial Ecology* **55**: 500-511.

Moran, M.A., and Miller, W.L. (2007) Resourceful heterotrophs make the most of light in the coastal ocean. *Nature Reviews Microbiology* **5**: 792.

Moran, N.A., Degnan, P.H., Santos, S.R., Dunbar, H.E., and Ochman, H. (2005) The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proceedings of the National Academy of Sciences USA* **102**: 16919-16926.

Morris, R., Vergin, K., Cho, J.-C., Rappe, M., Carlson, C., and Giovannoni, S. (2005) Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. *Limnology and Oceanography* **50**: 1687-1696.

Morris, R.M., Longnecker, K., and Giovannoni, S.J. (2006) Pirellula and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. *Environmental Microbiology* **8**: 1361-1370.

Morris, R.M., Rappe, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., and Giovannoni, S.J. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806-810.

Mou, X., Hodson, R.E., and Moran, M.A. (2007) Bacterioplankton assemblages transforming dissolved organic compounds in coastal seawater. *Environmental Microbiology* **9**: 2025-2037.

Mou, X., Sun, S., Edwards, R.A., Hodson, R.E., and Moran, M.A. (2008) Bacterial carbon processing by generalist species in the coastal ocean. *Nature.* **451**: 708-711.

Moyer, C.L., Tiedje, J.M., Dobbs, F.C., and Karl, D.M. (1996) A computer-simulated restriction fragment length polymorphism analysis of bacterial small-subunit rRNA genes: efficacy of selected tetrameric restriction enzymes for studies of microbial diversity in nature. *Applied and Environmental Microbiology* **62**: 2501-2507.

Muhling, M., Woolven-Allen, J., Murrell, J.C., and Joint, I. (2008) Improved group-specific PCR primers for denaturing gradient gel electrophoresis analysis of the genetic diversity of complex microbial communities. *ISME Journal* **2**: 379-392.

Mullins, T.D., Britschgi, T.B., Krest, R.L., and Giovannoni, S.J. (1995) Genetic comparisons reveal the same unknown bacterial lineages in Atlantic and Pacific bacterioplankton communities. *Limnology and Oceanography* **40**: 148-158.

Murphy, L.S., and Haugen, E.M. (1985) The distribution and abundance of phototrophic ultraplankton in the North Atlantic. *Limnology and Oceanography* **30**: 47-58.

Murray, A.E., Blakis, A., Massana, R., Strawzewski, S., Passow, U., Alldredge, A., and DeLong, E.F. (1999) A time series assessment of planktonic archaeal variability in the Santa Barbara Channel. *Aquatic Microbial Ecology* **20**: 129-145.

Muyzer, G., and Smalla, K. (1998) Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie Van Leeuwenhoek* **73**: 127-141.

Neufeld, J.D., Chen, Y., Dumont, M.G., and Murrell, J.C. (2008) Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environmental Microbiology* **10**: 1526-1535.

Neufeld, J.D., Mohn, W.W., and de Lorenzo, V. (2006) Composition of microbial communities in hexachlorocyclohexane (HCH) contaminated soils from Spain revealed with a habitat-specific microarray. *Environmental Microbiology* **8**: 126-140.

Nocker, A., Burr, M., and Camper, A. (2007) Genotypic Microbial Community Profiling: A Critical Technical Review. *Microbial Ecology* **54**: 276-289.

Noguchi, H., Park, J., and Takagi, T. (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research*

34: 5623-5630.

O'Mullan, G.D., and Ward, B.B. (2005) Relationship of Temporal and Spatial Variabilities of Ammonia-Oxidizing Bacteria to Nitrification Rates in Monterey Bay, California. *Applied and Environmental Microbiology* **71**: 697-705.

Olson, R.J., Chisholm, S.W., Zettler, E.R., and Armbrust, E.V. (1988) Analysis of Synechococcus pigment types in the sea using single and dual beam flow cytometry. *Deep Sea Res.* **35**: 425-440.

Olson, R.J., and Sosik, H.M. (2007) A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods* **5**: 195-203.

Olson, R.J., Vaulot, D., and Chisholm, S.W. (1985) Marine phytoplankton distributions measured using shipboard flow cytometry. *Deep Sea Res.* **32**: 1273-1280.

Ong, L.J., Glazer, A.N., and Waterbury, J.B. (1984) An Unusual Phycoerythrin from a Marine Cyanobacterium. *Science* **224**: 80-83.

Orphan, V.J., House, C.H., Hinrichs, K.-U., McKeegan, K.D., and DeLong, E.F. (2001) Methane-Consuming Archaea Revealed by Directly Coupled Isotopic and Phylogenetic Analysis. *Science* **293**: 484-487.

Osborne, C.A., Rees, G.N., Bernstein, Y., and Janssen, P.H. (2006) New threshold and confidence estimates for terminal restriction fragment length polymorphism analysis of complex bacterial communities. *Applied and Environmental Microbiology* **72**: 1270-1278.

Ottesen, E.A., Hong, J.W., Quake, S.R., and Leadbetter, J.R. (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* **314**: 1464-1467.

Pace, N.R., Stahl, D.A., Olsen, G.J., and Lane, D.J. (1985) Analyzing natural microbial populations by rRNA sequences. *American Society for Microbiology News* **51**: 4-12.

Palmer, C., Bik, E.M., DiGiulio, D.B., Relman, D.A., and Brown, P.O. (2007) Development of the Human Infant Intestinal Microbiota. *PLoS Biology* **5**: e177.

Palmer, C., Bik, E.M., Eisen, M.B., Eckburg, P.B., Sana, T.R., Wolber, P.K. et al. (2006) Rapid quantitative profiling of complex microbial populations. *Nucleic Acids Research* **34**: e5-.

Park, S.J., Kang, C.H., Chae, J.C., and Rhee, S.K. (2008) Metagenome microarray for screening of fosmid clones containing specific genes. *FEMS Microbiol Lett* **284**: 28-34.

Pennington, J., Timothy, Chavez, Francisco P. (2000) Seasonal fluctuations of temperature, salinity, nitrate, chlorophyll and primary production at station H3/M1 over 1989-1996 in Monterey Bay, California. *Deep Sea Research Part II: Topical Studies in Oceanography* **47**: 947-973.

Pennington, J.T., Mahoney, K.L., Kuwahara, V.S., Kolber, D.D., Calienes, R., and Chavez, F.P. (2006) Primary production in the eastern tropical Pacific:

A review. *Progress In Oceanography* **69**: 285-317.

Pennington, J.T., Michisaki, R., Johnston, D., and Chavez, F.P. (2007) Ocean observing in the Monterey Bay National Marine Sanctuary: CalCOFI and the MBARI time series In. Monterey: The Sanctuary Integrated Monitoring Network (SIMoN), Monterey Bay Sanctuary Foundation, and Monterey Bay National Marine Sanctuary p. 24.

Pernthaler, A., and Amann, R. (2004) Simultaneous Fluorescence In Situ Hybridization of mRNA and rRNA in Environmental Bacteria. *Applied and Environmental Microbiology* **70**: 5426-5433.

Pernthaler, A., Dekas, A.E., Brown, C.T., Goffredi, S.K., Embaye, T., and Orphan, V.J. (2008) Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. *Proceedings of the National Academy of Sciences USA* **105**: 7052-7057.

Pernthaler, A., Pernthaler, J., and Amann, R. (2002) Fluorescence In Situ Hybridization and Catalyzed Reporter Deposition for the Identification of Marine Bacteria. *Applied and Environmental Microbiology* **68**: 3094-3101.

Pernthaler, A., Preston, C.M., Pernthaler, J., DeLong, E.F., and Amann, R. (2002) Comparison of Fluorescently Labeled Oligonucleotide and Polynucleotide Probes for the Detection of Pelagic Marine Bacteria and Archaea. *Applied and Environmental Microbiology* **68**: 661-667.

Podar, M., Abulencia, C.B., Walcher, M., Hutchison, D., Zengler, K., Garcia, J.A. et al. (2007) Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Applied and Environmental Microbiology* **73**: 3205-3214.

Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R.D., Buigues, B. et al. (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311**: 392-394.

Polz, M.F., and Cavanaugh, C.M. (1998) Bias in Template-to-Product Ratios in Multitemplate PCR. *Applied and Environmental Microbiology* **64**: 3724-3730.

Polz, M.F., Harbison, C., and Cavanaugh, C.M. (1999) Diversity and heterogeneity of epibiotic bacterial communities on the marine nematode Eubostrichus dianae. *Applied and Environmental Microbiology* **65**: 4271-4275.

Poretsky, R.S., Bano, N., Buchan, A., LeCleir, G., Kleikemper, J., Pickering, M. et al. (2005) Analysis of microbial gene transcripts in environmental samples. *Applied and Environmental Microbiology* **71**: 4121-4126.

Raes, J., Foerstner, K.U., and Bork, P. (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Current Opinion in Microbiology* **10**: 490-498.

Raes, J., Korbel, J.O., Lercher, M.J., von Mering, C., and Bork, P. (2007) Prediction of effective genome size in metagenomic samples. *Genome Biology* **8**: R10.

Ram, R.J., Verberkmoes, N.C., Thelen, M.P., Tyson, G.W., Baker, B.J., Blake,

R.C., 2nd et al. (2005) Community proteomics of a natural microbial biofilm. *Science* **308**: 1915-1920.

Rappe, M.S., Connon, S.A., Vergin, K.L., and Giovannoni, S.J. (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**: 630-633.

Rappe, M.S., Vergin, K., and Giovannoni, S.J. (2000) Phylogenetic comparisons of a coastal bacterioplankton community with its counterparts in open ocean and freshwater systems. *FEMS Microbiology Ecology* **33**: 219-232.

Rashby, S.E., Sessions, A.L., Summons, R.E., and Newman, D.K. (2007) Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proceedings of the National Academy of Sciences USA* **104**: 15099-15104.

Reich M, L.T., Gould J, Lerner J, Tamayo P, Mesirov JP (2006) GenePattern 2.0. *Nature Genetics* **38**: 500-501.

Reisenfeld, C.S., P.D. Schloss and J. Handelsman (2004) Metagenomics: Genome Analysis of Microbial Communities. *Annual Review of Genetics* **38**.

Rhee, S.-K., Liu, X., Wu, L., Chong, S.C., Wan, X., and Zhou, J. (2004) Detection of genes involved in biodegradation and biotransformation in microbial communities by using 50-Mer oligonucleotide microarrays. *Applied Environmental Microbiology* **70**: 4303-4317.

Rice, D.W., Seltenrich, C.P., Spies, R.B., and Keller, M.L. (1993) Seasonal and annual distribution of organic contaminants in marine sediments from Elkhorn slough, moss landing harbor and nearshore Monterey Bay, California. *Environmental Pollution* **82**: 79-91.

Ritchie, N.J., Schutter, M.E., Dick, R.P., and Myrold, D.D. (2000) Use of length heterogeneity PCR and fatty acid methyl ester profiles to characterize microbial communities in soil. *Applied and Environmental Microbiology* **66**: 1668-1675.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A. et al. (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042-1047.

Rodríguez-Valera, F. (2002) Approaches to prokaryotic biodiversity: a population genetics perspective. *Environmental Microbiology* **4**: 628-633.

Rodríguez-Valera, F. (2004) Environmental genomics, the big picture? *FEMS Microbiology Letters* **231**: 153-158.

Roesch, L.F., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K., Kent, A.D. et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME Journal* **1**: 283-290.

Rondon, M.R., August, P.R., Bettermann, A.D., Brady, S.F., Grossman, T.H., Liles, M.R. et al. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and Environmental Microbiology* **66**: 2541-2547.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S.,

Yooseph, S. et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology* **5**: e77.

Sabehi, G., Beja, O., Suzuki, M.T., Preston, C.M., and DeLong, E.F. (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. *Environmental Microbiology* **6**: 903-910.

Sabehi, G., Kirkup, B.C., Rozenberg, M., Stambler, N., Polz, M.F., and Beja, O. (2007) Adaptation and spectral tuning in divergent marine proteorhodopsins from the eastern Mediterranean and the Sargasso Seas. *ISME Journal* **1**: 48-55.

Sabehi, G., Loy, A., Jung, K.-H., Partha, R., Spudich, J.L., Isaacson, T. et al. (2005) New Insights into Metabolic Properties of Marine Bacteria Encoding Proteorhodopsins. *PLoS Biology* **3**: e273.

Sanguin, H., Herrera, A., Oger-Desfeux, C., Dechesne, A., Simonet, P., Navarro, E. et al. (2006) Development and validation of a prototype 16S rRNA-based taxonomic microarray for *Alphaproteobacteria*. *Environmental Microbiology* **8**: 289–307.

Santiago-Vazquez, L.Z., Bruck, T.B., Bruck, W.M., Duque-Alarcon, A.P., McCarthy, P.J., and Kerr, R.G. (2007) The diversity of the bacterial communities associated with the azooxanthellate hexacoral Cirrhipathes lutkeni. *ISME Journal* **1**: 654-659.

Scanlan, D.J., and West, N.J. (2002) Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbial Ecology* **40**: 1-12.

Schmid, M.C., Maas, B., Dapena, A., van de Pas-Schoonen, K., van de Vossenberg, J., Kartal, B. et al. (2005) Biomarkers for In Situ Detection of Anaerobic Ammonium-Oxidizing (Anammox) Bacteria. *Applied and Environmental Microbiology* **71**: 1677-1684.

Schonhuber, W., Fuchs, B., Juretschko, S., and Amann, R. (1997) Improved sensitivity of whole-cell hybridization by the combination of horseradish peroxidase-labeled oligonucleotides and tyramide signal amplification. *Applied and Environmental Microbiology* **63**: 3268-3273.

Schramm, A., Fuchs, B.M., Nielsen, J.L., Tonolla, M., and Stahl, D.A. (2002) Fluorescence in situ hybridization of 16S rRNA gene clones (Clone-FISH) for probe validation and screening of clone libraries. *Environmental Microbiology* **4**: 713-720.

Schutter, M.E., and Dick, R.P. (2000) Comparison of Fatty Acid Methyl Ester (FAME) Methods for Characterizing Microbial Communities. *Soil Sci Soc Am J* **64**: 1659-1668.

Schwalbach, M.S., Brown, M.V., and Fuhrman, J.A. (2005) Impact of light on marine bacterioplankton community structure. *Aquatic Microbial Ecology* **39**: 235-245.

Schweiger, F., and Tebbe, C. (1998) A new approach to utilize PCR-single-strand-conformation polymorphism for 16S rRNA gene-based microbial

community analysis. *Applied and Environmental Microbiology* **64**: 4870-4876.

Sebat, J.L., Colwell, F.S., and Crawford, R.L. (2003) Metagenomic profiling: Microarray analysis of an environmental genomic library. *Applied Environmental Microbiology* **69**: 4927-4934.

Sekar, R., Fuchs, B.M., Amann, R., and Pernthaler, J. (2004) Flow Sorting of Marine Bacterioplankton after Fluorescence In Situ Hybridization. *Applied and Environmental Microbiology* **70**: 6210-6219.

Seymour, J.R., Seuront, L., and Mitchell, J.G. (2005) Microscale and small-scale temporal dynamics of a coastal planktonic microbial community. *Marine Ecology Progress Series* **300**: 21-37.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M. et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**: 1728-1732.

Small, J., Call, D.R., Brockman, F.J., Straub, T.M., and Chandler, D.P. (2001) Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. *Applied Environmental Microbiology* **67**: 4708-4716.

Smolina, I., Lee, C., and Frank-Kamenetskii, M. (2007) Detection of Low-Copy-Number Genomic DNA Sequences in Individual Bacterial Cells by Using Peptide Nucleic Acid-Assisted Rolling-Circle Amplification and Fluorescence In Situ Hybridization. *Applied and Environmental Microbiology* **73**: 2324-2328.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R. et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences USA* **103**: 12115-12120.

Sosik, H.M., and Olson, R.J. (2007) Automated taxonomic classification of phytoplankton sampled with imaging in-flow cytometry. *Limnology and Oceanography: Methods* **5**: 204-216.

Soule, M., Kuhn, E., Loge, F., Gay, J., and Call, D.R. (2006) Using DNA microarrays to identify library-independent markers for bacterial source tracking. *Applied and Environmental Microbiology* **72**: 1843-1851.

Staley, J.T., and Konopka, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology* **39**: 321-346.

Stapels, M.D., Cho, J.C., Giovannoni, S.J., and Barofsky, D.F. (2004) Proteomic analysis of novel marine bacteria using MALDI and ESI mass spectrometry. *Journal of Biomolecular Technology* **15**: 191-198.

Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H., and DeLong, E.F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology* **178**: 591-599.

Steinberg, D.K., Carlson, C.A., Bates, N.R., Johnson, R.J., Michaels, A.F., and Knap, A.H. (2001) Overview of the US JGOFS Bermuda Atlantic Time-

series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Research Part II: Topical Studies in Oceanography* **48**: 1405-1447.

Stepanauskas, R., and Sieracki, M.E. (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proceedings of the National Academy of Sciences USA* **104**: 9052-9057.

Stingl, U., Cho, J.C., Foo, W., Vergin, K., Lanoil, B., and Giovannoni, S. (2008) Dilution-to-Extinction Culturing of Psychrotolerant Planktonic Bacteria from Permanently Ice-covered Lakes in the McMurdo Dry Valleys, Antarctica. *Microbial Ecology* **55**: 395-405.

Stingl, U., Tripp, H.J., and Giovannoni, S.J. (2007) 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 Journal* **1**: 361-371.

Stralis-Pavese, N., Sessitsch, A., Weilharter, A., Reichenauer, T., Riesing, J., Csontos, J. et al. (2004) Optimization of diagnostic microarray for application in analysing landfill methanotroph communities under different plant covers. *Environmental Microbiology* **6**: 347-363.

Streit, W.R.a.R.A.S. (2004) Metagenomics - the key to the uncultured microbes. *Current Opinion in Microbiology* **7**: 492-498.

Strous, M., Pelletier, E., Mangenot, S., Rattei, T., Lehner, A., Taylor, M.W. et al. (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**: 790-794.

Summons, R.E., Jahnke, L.L., and Simoneit, B.R. (1996) Lipid biomarkers for bacterial ecosystems: studies of cultured organisms, hydrothermal environments and ancient sediments. *Ciba Found Symp* **202**: 174-193; discussion 193-174.

Sun, L., Qiu, F., Zhang, X., Dai, X., Dong, X., and Song, W. (2008) Endophytic Bacterial Diversity in Rice (Oryza sativa L.) Roots Estimated by 16S rDNA Sequence Analysis. *Microbial Ecology* **55**: 415-424.

Suzuki, M., and Giovannoni, S. (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology* **62**: 625-630.

Suzuki, M., Rappe, M.S., and Giovannoni, S.J. (1998) Kinetic bias in estimates of coastal picoplankton community structure obtained by measurements of small-subunit rRNA gene PCR amplicon length heterogeneity. *Applied and Environmental Microbiology* **64**: 4522-4529.

Suzuki, M.T., Beja, O., Taylor, L.T., and Delong, E.F. (2001) Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environmental Microbiology* **3**: 323-331.

Suzuki, M.T., Preston, C.M., Beja, O., de la Torre, J.R., Steward, G.F., and DeLong, E.F. (2004) Phylogenetic Screening of Ribosomal RNA Gene-Containing Clones in Bacterial Artificial Chromosome (BAC) Libraries from Different Depths in Monterey Bay. *Microbial Ecology* **48**: 473-488.

Suzuki, M.T., Taylor, L.T., and DeLong, E.F. (2000) Quantitative analysis of small-subunit rRNA genes in mixed microbial populations via 5'-nuclease assays. *Applied and Environmental Microbiology* **66**: 4605-4614.

Tamas, I., Klasson, L., Canback, B., Naslund, A.K., Eriksson, A.S., Wernegreen, J.J. et al. (2002) 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**: 2376-2379.

Taroncher-Oldenburg, G., Griner, E., Francis, C., and Ward, B. (2003) Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Applied and Environmental Microbiology* **69**: 1159-1171.

Teske, A., and Sorensen, K.B. (2007) Uncultured archaea in deep marine subsurface sediments: have we caught them all? *ISME Journal* **2**: 3-18.

Thiel, V., Toporski, J., Schumann, G., Sjovall, P., and Lausmaa, J. (2007) Analysis of archaeal core ether lipids using Time of Flight-Secondary Ion Mass Spectrometry (ToF-SIMS): Exploring a new prospect for the study of biomarkers in geobiology. *Geobiology* **5**: 75-83.

Thompson, J.R., Marcelino, L.A., and Polz, M.F. (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic Acids Research* **30**: 2083-2088.

Thompson, J.R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D.E., Benoit, J. et al. (2005) Genotypic Diversity Within a Natural Coastal Bacterioplankton Population. *Science* **307**: 1311-1313.

Thompson, J.R., Randa, M.A., Marcelino, L.A., Tomita-Mitchell, A., Lim, E., and Polz, M.F. (2004) Diversity and dynamics of a north Atlantic coastal Vibrio community. *Applied and Environmental Microbiology* **70**: 4103-4110.

Thyssen, M., Tarran, G.A., Zubkov, M.V., Holland, R.J., Gregori, G., Burkill, P.H., and Denis, M. (2008) The emergence of automated high-frequency flow cytometry: revealing temporal and spatial phytoplankton variability. *Journal of Plankton Research* **30**: 333-343.

Tiedje, J.M., Asuming-Brempong, S., N.sslein, K., Marsh, T.L., and Flynn, S.J. (1999) Opening the black box of soil microbial diversity. *Applied Soil Ecology* **13**: 109-122.

Tiquia, S.M., Wu, L., Chong, S.C., Passovets, S., Xu, D., Xu, Y., and Zhou, J. (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. *Biotechniques* **36**: 664--672.

Trebesius, K., Amann, R., Ludwig, W., Muhlegger, K., and Schleifer, K.-H. (1994) Identification of Whole Fixed Bacterial Cells with Nonradioactive 23S rRNA-Targeted Polynucleotide Probes. *Applied and Environmental Microbiology* **60**: 3228-3235.

Treusch, A.H., Kletzin, A., Raddatz, G., Ochsenreiter, T., Quaiser, A., Meurer, G. et al. (2004) Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. *Environmental Microbiology* **6**: 970-980.

Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang,

H.W. et al. (2005) Comparative metagenomics of microbial communities. *Science* **308**: 554-557.

Tringe, S.G., Zhang, T., Liu, X., Yu, Y., Lee, W.H., Yap, J. et al. (2008) The Airborne Metagenome in an Indoor Urban Environment. *PLoS ONE* **3**: e1862.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**: 1027-1031.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.

Urisman, A., Fischer, K.F., Chiu, C.Y., Kistler, A.L., Beck, S., Wang, D., and DeRisi, J.L. (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biology* **6**: R78.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Vestal, J.R., and White, D.C. (1989) Lipid Analysis in Microbial Ecology. *BioScience* **39**: 535-541.

Villanueva, L., Navarrete, A., Urmeneta, J., White, D.C., and Guerrero, R. (2004) Combined Phospholipid Biomarker-16S rRNA Gene Denaturing Gradient Gel Electrophoresis Analysis of Bacterial Diversity and Physiological Status in an Intertidal Microbial Mat. *Applied and Environmental Microbiology* **70**: 6920-6926.

von Mering, C., Hugenholtz, P., Raes, J., Tringe, S.G., Doerks, T., Jensen, L.J. et al. (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126-1130.

Wagner, M., Horn, M., and Daims, H. (2003) Fluorescence in situ hybridisation for the identification and characterisation of prokaryotes. *Current Opinion in Microbiology* **6**: 302-309.

Wagner, M., Loy, A., Klein, M., Lee, N., Ramsing, N.B., Stahl, D.A., and Friedrich, M.W. (2005) Functional marker genes for identification of sulfate-reducing prokaryotes. *Methods Enzymol* **397**: 469-489.

Wagner, M., Nielsen, P.H., Loy, A., Nielsen, J.L., and Daims, H. (2006) Linking microbial community structure with function: fluorescence in situ hybridization-microautoradiography and isotope arrays. *Current Opinion in Biotechnology* **17**: 83-91.

Wagner, M., Smidt, H., Loy, A., and Zhou, J. (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microbial Ecology* **53**: 498-506.

Wang, D., L Coscoy, M. Zylberberg, P.C. Avila, H.A. Boushey, D. Ganem, and J.L. DeRisi (2002) Microarray-based detection and genotyping of viral

pathogens. *PNAS* **99**: 15687-15692.

Ward, B.B. (1982) Oceanic distribution of ammonium-oxidizing bacteria determined by immunofluorescent assay. *Journal of Marine Research* **40**: 1155-1172.

Ward, B.B. (2005) Temporal variability in nitrification rates and related biogeochemical factors in Monterey Bay, California, USA. *Marine Ecology Progress Series* **292**: 97-109.

Ward, B.B., Eveillard, D., Kirshtein, J.D., Nelson, J.D., Voytek, M.A., and Jackson, G.A. (2007) Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environmental Microbiology* **9**: 2522-2538.

Ward, D.M., Cohan, F.M., Bhaya, D., Heidelberg, J.F., Kuhl, M., and Grossman, A. (2008) Genomics, environmental genomics and the issue of microbial species. *Heredity* **100**: 207-219.

Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T. et al. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**: 560-565.

Waterbury, J.B., Watson, S.W., Guillard, R.R.L., and Brand, L.E. (1979) Widespread occurrence of a unicellular marine planktonic cyanobacterium. *Nature* **277**: 293-294.

Wegley, L., Edwards, R., Rodriguez-Brito, B., Liu, H., and Rohwer, F. (2007) Metagenomic analysis of the microbial community associated with the coral Porites astreoides. *Environmental Microbiology* **9**: 2707-2719.

Whitaker, R.J., and Banfield, J.F. (2006) Population genomics in natural microbial communities. *Trends in Ecology and Evolution* **21**: 508-516.

Wilhelm, L.J., Tripp, H.J., Givan, S.A., Smith, D.P., and Giovannoni, S.J. (2007) Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biology Direct* **2**: 27.

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I. et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* **3**: e1456.

Wilson, K.H., Wilson, W.J., Radosevich, J.L., DeSantis, T.Z., Viswanathan, V.S., Kuczmarski, T.A., and Andersen, G.L. (2002) High-Density Microarray of Small-Subunit Ribosomal DNA Probes. *Applied and Environmental Microbiology* **68**: 2535-2541.

Woebken, D., Fuchs, B.M., Kuypers, M.M.M., and Amann, R. (2007) Potential Interactions of Particle-Associated Anammox Bacteria with Bacterial and Archaeal Partners in the Namibian Upwelling System. *Applied and Environmental Microbiology* **73**: 4648-4657.

Wommack, K.E., Bhavsar, J., and Ravel, J. (2008) Metagenomics: Read length matters. *Applied and Environmental Microbiology* **Epub ahead of print**.

Wright, S., and Jeffrey, S. (2006) Pigment Markers for Phytoplankton Production. In *Marine Organic Matter: Biomarkers, Isotopes and DNA*, pp. 71-104.

Wright, T.D., Vergin, K.L., Boyd, P.W., and Giovannoni, S.J. (1997) A novel delta-subdivision proteobacterial lineage from the lower ocean surface layer. *Applied and Environmental Microbiology* **63**: 1441-1448.

Wu, D., Daugherty, S.C., Van Aken, S.E., Pai, G.H., Watkins, K.L., Khouri, H. et al. (2006) Metabolic complementarity and genomics of the dual bacterial symbiosis of sharpshooters. *PLoS Biology* **4**: e188.

Wu, L., Liu, X., Schadt, C.W., and Zhou, J. (2006) Microarray-based analysis of subnanogram quantities of microbial community DNAs by using whole-community genome amplification. *Applied and Environmental Microbiology* **72**: 4931-4941.

Wu, L., Thompson, D., Li, G., Hurt, R., Tiedje, J., and Zhou, J. (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Applied and Environmental Microbiology* **67**: 5780-5790.

Wu, L., Thompson, D.K., Liu, X., Fields, M.W., Bagwell, C.E., Tiedje, J.M., and Zhou, J. (2004) Development and Evaluation of Microarray-Based Whole-Genome Hybridization for Detection of Microorganisms within the Context of Environmental Applications. *Environmental Science and Technology* **38**: 6775-6782.

Yang, S.-Y., Hsiung, S.-K., Hung, Y.-C., Chang, C.-M., Liao, T.-L., and Lee, G.-B. (2006) A cell counting/sorting system incorporated with a microfabricated flow cytometer chip. *Measurement Science and Technology*: 2001.

Yeates, C., and Blackall, L.L. (2006) Construction and analysis of a metagenomic library from an enhanced biological phosphorus removal biomass. *Water Science and Technology* **54**: 277-284.

Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K. et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biology* **5**: e16.

Yutin, N., Suzuki, M.T., Teeling, H., Weber, M., Venter, J.C., Rusch, D.B., and Beja, O. (2007) Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environmental Microbiology* **9**: 1464-1475.

Zeidner, G., Preston, C.M., Delong, E.F., Massana, R., Post, A.F., Scanlan, D.J., and Béjà, O. (2003) Molecular diversity among marine picophytoplankton as revealed by psbA analyses. *Environmental Microbiology* **5**: 212-216.

Zeng, Y., Liu, W., Li, H., Yu, Y., and Chen, B. (2007) Effect of restriction endonucleases on assessment of biodiversity of cultivable polar marine planktonic bacteria by amplified ribosomal DNA restriction analysis. *Extremophiles* **11**: 685-692.

Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W., and Church, G.M. (2006) Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology* **24**: 680-686.

Zhang, L., Hurek, T., and Reinhold-Hurek, B. (2007) A nifH -based Oligonucleotide Microarray for Functional Diagnostics of Nitrogen-fixing Microorganisms. *Microbial Ecology* **53**: 456-470.

Zhang, T., and Fang, H.H. (2006) Applications of real-time polymerase chain reaction for quantification of microorganisms in environmental samples. *Applied Microbiology and Biotechnology* **70**: 281-289.

Zhou, J. (2003) Microarrays for bacterial detection and microbial community analysis. *Current Opinion in Microbiology* **6**: 288-294.

Zubkov, M.V., Fuchs, B.M., Archer, S.D., Kiene, R.P., Amann, R., and Burkill, P.H. (2002) Rapid turnover of dissolved DMS and DMSP by defined bacterioplankton communities in the stratified euphotic zone of the North Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **49**: 3017-3038.

Zwirglmaier, K., Ludwig, W., and Schleifer, K.H. (2004) Recognition of individual genes in a single bacterial cell by fluorescence in situ hybridization - RING-FISH. *Molecular Microbiology* **51**: 89-96.

Appendix 1

Protocols & Source Sheets Developed during this thesis for the Genome Proxy
Array.

184

Preparing poly-lysine slides for printing microarrays:

<u>Goal:</u> To coat glass microscope slides with poly-lysine ("PLL"), so they are "sticky". Then when the arraying robot prints DNA spots onto them (making a microarray) the DNA sticks.

<u>Materials:</u>

Gold Seal Micro Slides, Cat. No. 3010, 3" x 1", 1mm thick, Fisher # 12-518-100A
NaOH pellets, e.g. Sigma # S8045
95% EtOH
lots of Milli-Q H2O
1X or 10X PBS
Poly-L-lysine solution, 0.1% (w/v), Sigma # P8920
slide boxes, Fisherbrand, foam-lined not cork-lined, Fisher #03-448-4

1 secondary containment tray
3L glass beakers (2+)
1 1L plastic beaker
plastic boxes for PLL-coating of slides (pipette-tip boxes work fine)
metal slide racks
plastic washing container
plastic wrap or tinfoil
rubber band

<u>Notes:</u>

*rinse all containers with RO before using, to remove dust, etc.. Keep these containers separate (at Gin's bench).

* throughout this process, make sure the slides in the racks stay well separated, by running a gloved finger across the top edge of the slides, etc – during rinsing, etc. – at *all* steps

* to keep dust from sticking to the slides, once the protocol is begun try to keep the slides submerged in solution at all times, &/or covered.

* do NOT use powdered gloves during this protocol.

* slides **MUST be stored at least two weeks before spotting DNA** (DeRisi lab says 2 weeks, Schoolnik lab says 3 weeks, Somero lab says 1 week). **Don't use slides that are > 4 mos. old for printing,** sometimes the poly-lysine degrades (DeRisi says 4 mos., Schoolnik labs says 3 mos.).

<u>1. Wash slides:</u> Although the slides come "clean" in a box, there is still a fair bit of dust and dirt on them and they need to be extremely clean before they are coated, because any dirt will cause irregularities in the coating which will be weak spots, and may peel off.

<u>make up wash solution:</u>
     need about 600mls of solution to cover a metal slide rack in a 3L beaker
     in a <u>glass</u> beaker (2L or bigger), mix:

<u>For 60 slides: 1200mls</u>	<u>For 90 slides: 1800mls</u>	<u>For 120 slides: 2400mls</u>
120 g NaOH pellets	180 g NaOH pellets	240 g NaOH pellets
720 mls 95% EtOH	1080 mls 95% EtOH	1440 mls 95% EtOh
480 mls Milli-Q H2O	720 mls Milli-Q H2O	960 mls Milli-Q H2O

(final conc. for this solution is 57% EtOH, 10% w/v NaOH)

mix with stir-bar on stir-plate (no heat) until fully dissolved (takes ~15min)
Thoroughly rinse 1 glass 3L beaker for each 30 slides.

Place slides into metal racks (our current racks take 30 slides each), use air canister to spray off slides.

Put racks into beakers, and beakers into one large plastic tub (for secondary containment on shaking table – this solution is highly basic and we don't want it to spill). Put secondary containment tray on shaker, into hood.

pour 600mls wash solution into each 3L beaker. Cover the beakers, with plastic wrap or foil, to prevent dust from getting in.

shake gently on table to wash, about 2hrs

2. Rinse slides: The wash solution must be fully rinsed from the slides or it will interfere with the coating process.

wash slides 5x *vigorously* with **clean Milli-Q water** – e.g., put two racks at a time in long narrow plastic tubs, place a rubber band around the slides - **as close to the two ends as possible**!!! - to hold them in the rack securely, run Milli-Q water over racks and then swoosh racks up and down *and* back and forth in water vigorously, repeatedly, for maybe 30 seconds. Dump water and repeat process 4 more times.

let the slides sit in clean water while you prepare the poly-lysine solution

3. Coat slides:
only use PLASTIC with poly-lysine!
So, to coat the slides use the empty plastic pipette-tip boxes labeled "for poly-lysine". To make up poly-lysine solution, use a plastic beaker or dish.

For coating 2 racks of slides at once:

For 750mls poly-lysine solution:
(This solution's final concentrations are ~0.0169% w/v PLL, and 0.0984X PBS.)
550mls Milli-Q H2O
73.8mls 1X PBS (kept in fridge once opened, post-autoclaving)
126.72mls poly-lysine solution, 0.1% w/v in H2O
mix ingredients in order listed, in plastic, with stir-bar on stir-plate

dump excess water from slides

place each of two slide racks into its own poly-lysine pipette-tip box

immediately pour 375mls poly-lysine solution over each rack

By Virginia Rich, graduate student in the DeLong Lab, vrich@mit.edu
Last modified 2.12.08
VR's protocol, adapted from Somero lab protocol, as adapted from DeRisi lab protocol

close each box

put the two boxes into a plastic tub for secondary containment

let slides shake on shaker table in poly-lysine solution for 30 minutes

(Note: each box of poly-lysine solution can be re-used the same day, to coat one more rack of slides. Also, if a lot of slides will be prepared on several consecutive days, the solution can be filtered after the first day and stored in plastic in the fridge. When ready to use again, add an additional 3-5 mls of poly-lysine. This can be repeated a maximum of six times.)

4. Rinse slides: All the excess poly-lysine solution should be removed, so that the coating is as even as possible.

lift the metal slide racks out of the poly-lysine solution, and place the two racks into the long narrow plastic tub for washing. Wash slides 5x vigorously with clean Milli-Q water – following the same protocol as first rinse step.

(Meanwhile, if needed, start next racks of slides coating.)

Keep the rinsed slides in water until they are spun – they can wait a *few* minutes if someone else is using the centrifuge, for example.

Drain the excess water from racks, and put them into the centrifuge, on top of several folded up large kimwipes.

Position the racks with the slides in a consistent orientation so you know which way is "facing" the direction of motion. (Imagine riding on a very fast merry-go-round coating in maple syrup. As you move, the syrup would dry off the front side of you first, and also flow around your edges a little to build up along the sides of your back. We see this with the slides – on their back faces, the edges show signs of PLL wrapping around during the spin... No problem on the backside, but not so great for the side we want to print on.)

Make sure the racks are balanced in weight (same number of slides) *and position*. Check the separation of the slides in the racks (finger run across top edge).

spin ~150xg (in Ed's plate-spinner centrifuge, a Sorvall Legend RT, with the plate-spinner rotor in it) for 5-10 min., *at room temperature*, until dry. Before putting the slides in the centrifuge, spray out the rotor, plate holders, etc., with an air canister, to prevent dust.

remove the slides from the racks and place into a slide storage box, (having sprayed the box dust-free with an air-can), all slides in the same orientation (e.g. all with the slide faces that were at the "front" – in the direction of spinning motion – pointing the same direction in the box. Label the outside of the box with tape saying "PLL-coated slides" and **the date**, and your name, and the directionality of the slides in the box (based on direction of spin, and see below).

By Virginia Rich, graduate student in the DeLong Lab, vrich@mit.edu
Last modified 2/12/08
VR's protocol, adapted from Somero lab protocol, as adapted from DeRisi lab protocol

Place storage box into dessicator cabinet.

5. To QC the slides: It's a good idea to see how the slides look, to make sure nothing obvious has gone wrong. This can be done the same day or in the next several days.

  - breathe on them, look for surface irregularities.

  - check their background by scanning on the array scanner
    - edge effects not that uncommon, shouldn't interfere with where array goes

    - slide sidedness where one face of slide is more speckly than the other also not uncommon. If this occurs, **orient slides** methodically in the case so that the better side always faces the same direction, and note this on the outside of the storage box – this will be the face used for spotting the DNA onto.

    - it's probably good to save a few of the slide scans, and print out on a single page, to keep a record of that batch's general properties post-coating. Just doing Preview Scans is fine. Save the file, and also export it using both our standard brightness/contrast and the auto-brightness/contrast.

Our standard settings for scanning PLL slides:
    635nm laser (red): 100%, PMT Gain: 600
    532nm laser (green): 100%, PMT Gain: 600
        brightness at 95, contrast at 93
but, also look at and save a few images using the auto-brightness/contrast button

our file naming convention:
    Year_Month_Day_description_slidenumber_front_or_back_autosettings
e.g. 2005_01_19_newPLLslides_1f   or   2004_12_20_newPLLslides_3b_auto

Protocol Summary:

1. Wash slides
make up wash solution:
    in a glass beaker, mix:

For 60 slides: 1200mls	For 90 slides: 1800mls	For 120 slides: 2400mls
120 g NaOH pellets	180 g NaOH pellets	240 g NaOH pellets
720 mls 95% EtOH	1080 mls 95% EtOH	1440 mls 95% EtOh
480 mls Milli-Q H2O	720 mls Milli-Q H2O	960 mls Milli-Q H2O

pour 600mls wash solution into each *clean* 3L beaker, containing rack of slides.
Cover the beakers.
shake gently on table to wash, about **2hrs**

2. Rinse slides: wash slides 5x *vigorously* with **clean Milli-Q water**
 let the slides sit in clean water while you prepare the poly-lysine solution

3. Coat slides:
only use PLASTIC with poly-lysine!
        for 750mls poly-lysine solution:
                550mls Milli-Q H2O
                73.8mls 1X PBS (kept in fridge once opened, post-autoclaving)
                126.72mls poly-lysine solution, 0.1% w/v in H2O
        mix ingredients in order listed, in plastic, with stir-bar on stir-plate

dump excess water from slides
place each of two slide racks into its own poly-lysine pipette-tip box
immediately pour 375mls poly-lysine solution over each rack
close boxes & put into a plastic tub for secondary containment
let slides shake on shaker table in poly-lysine solution for **30 minutes**

any additional racks of slides can sit in Mill-Q water while waiting their turn for the poly-lysine

4. Rinse slides:
        Wash slides 5x vigorously with clean Milli-Q water
                (Meanwhile, if needed, start next racks of slides coating.)
        Drain the excess water from racks, and put them into the centrifuge, on top of several folded up large kimwipes. Balance them, and note direction of motion of slides (which faces point "forward").
        spin **150g** for **5 min.**, or until dry.
        remove the slides from the racks and place into a slide storage box, all in the same orientation. Label the outside of the box. Place storage box into dessicator. cabinet.

5. QC the slides

Form for Preparing Poly-Lysine coated slides, for microarrays:

Date:_____   Preparer:_____

Number of slides being prepped:_____

slide lot #:_____   NaOH lot #:_____   EtOH lot#:_____

NaOH slide wash:

time in washing:_____   time out:_____   total time:_____

PLL slide coating:

1st batch, time in coating:_____   time out:_____   total time:_____

2nd batch, time in coating:_____   time out:_____   total time:_____

Notes about any changes, observations, accidents, etc:

Q/C of slides:
1. Visual inspection when breathed upon:

2. General appearance using scanner:

3. A few representative scans, using both standard settings (brightness 95, contrast 93) and auto-settings:

Generating oligonucleotide probes for the marine microbial microarray:
all instructions are for a Mac set-up

Step 1: Generate a fasta file with all genes from each genome/genome fragment:

There are a variety of ways to do this. I now use a perl script. However, an easy way to do one or a few sequences is via the program Artemis.

- launch **Artemis**

> within a Terminal window, cd to the Directory containing Artemis and then type "art" and return – if this doesn't work, type "csh", return, then try again. Artemis should launch. If you search for Artemis in your computer, you may find that you have an icon-launchable version of it.

> 1. Options -> click off eukaryotic mode
>
> Open -> File -> file of choice, gb file (yes, ignore the error)

> 2. Select -> CDS features
>
> Write -> bases of selection -> fasta format
>
> use a ".fna" file extension, with appropriate prefix

Naming conventions: Location with some indication of clone type & depth if possible, and coordinates, eg. HF10_04G06 is Hawaii fosmid 10m, library plate 4 coordinates G06. For files, the convention is that ".fna" is used for fasta nucleic acids, and ".faa" is used for fasta amino acids

> 3. Select -> All
>
> Write -> bases of selection -> fasta format

> 4. Click in the lower gene list window and hit
>
> <control> and mouse click, then -> Save List as

This list is useful for annotating your oligos and for quickly checking the gene content of a fosmid/BAC later without having to launch Artemis. Some versions of Artemis won't let you do this; if not, no worries.

- then open the ".fna" file in **BBEdit** (or a similar text-editing program, one which won't add a bunch of stuff like Word does).

clean up file names as needed so that they lack spaces or funny characters.

This will be a different process from file-to-file since for many in-house sequences we're still working with unpublished versions that may not be perfectly named, etc. Generally, I prefer to keep each gene name as the CloneIdentifier_GeneIdentifier – often times this will be the sequence location of the gene because the gene names haven't yet been added, or it may just be CDS_001, _002, etc., depending on what information the Artemis-parsed CDS list contains. An ideal naming would be, eg., AntFos04D03_0to633, meaning AntFos library clone 04D03, the CDS from 0 t0 633.

- copy the file into the ArrayOligoSelector folder

Step 2: Use Array Oligo Selector to generate the potential probes:

ArrayOligoSelector is available, along with all documentation, at
http://arrayoligosel.sourceforge.net/
If working on a Mac you will need to download the version of formatdb and blastall from the
NCBI website whose date corresponds to the same release date (or as close as possible) of the
AOS version you're using, because the bundles AOS download comes with a Linux-compatible
version of formatdb and blastall. If you're doing more complicated things with AOS than what is
described below, there will also be other things you would need to download separately to
allow complete functionality of the program, but for the scripts we run, this is sufficient. I have
compared the results of AOS set up this way on a Mac Powerbook G4 to those from AOS set up
on a Linux machine and they are identical.

- in the Terminal window, change into the ArrayOligoSelector directory
        % cd [drag and drop ArrayOligoSelector folder for pathname]

- script 1 generates a list of all possible oligos from the input sequences, in a
sliding window manner. The output file is "output 0", you can view it as text.

        % pick70_script1
if this doesn't work, try typing ./pick70_script1
if you just type this you'll get a USAGE error telling you exactly what parameters you need to
input:
        *inputseq: gene/NUCLEOTIDE sequences submit for design in FASTA format

        *genome: genome GENE/NUCLEOTIDE sequences in FASTA format

        *oligo_size: in basepair

        *MaskByLowercase: You can exclude sub-sequences from the compuation using lower case.
        Those sub-regions will be flagged in the outputs. To use lowercase for this purpose, type
        "yes"; otherwise, type "no".

In the case of this array, we're choosing 70-mers, and are not doing any masking of sequence.

So, what we'd really like to do is:

        % pick70_script1 <input>.fna <input>.fna 70 no
For historical reasons, we use the same CDS output fna file as both CDS file and
genome file, against which ArrayOligoSelector checks for uniqueness.
We discussed a dizzying array of possibilities for what to use as the genome file: concatenating
all the fosmid and BAC sequences, using all the prokaryotic sequences in the nr database –
these days one could imagine using all available environmental sequences as a "genome"...
BUT, for our purposes, we did the simplest variant possible – using the CDs file as CDS and
genome. For different organisms there is different coverage of the nearby related "sequence
space", and this coverage changes all the time. One could try to make an array with much more
specific probes, even to the point of doing alignments, etc, as other groups have done for other
arrays, but that's not the purpose of this array. The goal was to see whether a "blind" design
approach would allow discrimination among related genotypes, and with the prototype array I
demonstrated that it did. If one were designing a different array for different purposes, or a
different system, one might want to use a different design strategy.

Script 1 will show a warning error because it can tell we're not running Linux
and they want to make sure we've got the correct versions of blastall and such

installed, and python – even if everything is good, it still gives you this warning
– so type "yes" to proceed when queried.

– script 2 chooses among the many possible oligos for each gene to give you
the ones closest to your desired parameters.

% pick70_script2

again, if this doesn't work, try typing ./pick70_script2
again, if you just type this you'll get a USAGE error telling you exactly what parameters you
need to input:

*GC: GC percentage (eg: 35.5, positive float or integer number)

*Oligo_len: length of Oligo in bp(positive integer)

*Number_Oligo: how many oligos do you want to design (positive integer)

*OPTIONAL binding energy cutoff: 0 is the default

*OPTIONAL masking parameters: if used, all the optional masking parameters are required

*Mask_Length: maximum length of subsequence allowed containing the
Mask_Symbols eg: 20

*Mask_Symbol (ATGCN): masking bases eg:AT or N

*Mask_Tolerance (0 -1) : percent of other bases allowed eg:0.1

So, we'd like to do 40% GC content (which was the average GC content of the few tens of fully-
sequenced clones present in the lab database at the time I started this), and 70-mers, and 1
probe per gene, with no binding energy cutoff and no masking:

% pick70_script2 40 70 1

– copy the oligodup and oligofasta output files from the ArrayOligoSelector
folder into a new location (remember, ArrayOligoSelector has to rewrite those
intermediate files each time, so you have to save them before you can run it
again), and rename the files based on the clone/organism name.

Step 3: Choose which output oligos to use as probes:
*Again, there are different ways to process the AOS outfiles... I now use a
perlscript to do this, which will get posted on the website too, but this is a
simple, alternate way to do the same thing manually.*

– open either output files in BBEdit, select all, and <apple> <F> to find and
replace – click on the lower left box for a Multifile search to include the other
file, and use grep:
	find: \r
	replace: (just a blank space)
	then

find: >
replace: \r>
save files

- open both files in Excel with a space as column delimiter.
merge the files into one

sort by %GC (column D or E, depending)

if there are <20 oligos with 40%GC, then take those just higher and lower until you have 20. Highlight these 20 oligos – these are your probes.

if there are >20 oligos with 40%GC, then sort *among those* oligos by ΔG of hybridization (column G usually), and take the 20 oligos with the lowest (=most negative) ΔG values, within those that have 40%GC. Highlight these 20, these are your probes.

ΔG has been shown to correlate inversely with hybridization signal for microarray probes, which makes good sense – so if you've got a surfeit of potential probes with the "right" %GC, ΔG makes a good criterion for selecting among them!

Copy and paste your chosen oligos into your master oligo file, and proceed as you see fit.

An **important thing to note** here is that "blind" probe design means that the process outlined above does *not* targeting particular genes of interest.

Resuspending oligo plates for arraying

Goal: I use 70-mer oligos for printing our microarrays. I get them from a company (Illumina) and have them make aliquot plates for us of 400 pmoles of each oligo in each well. I have Illumina ship the plates dried down. (Side note: I also have them ship the remainders, since their aliquotting robots apparently have some lame limitations and so can't dispense the last few aliquots for us – but we can do it ourselves from the remainder plates...) So, the dried-down aliquot plates need to be resuspended before they can be used for printing.

1. Volumes:
First, determine what the resuspension volume will be. I typically resuspend a new oligo-aliquot-plate at 10ul per well, which gives 40pmol/ul of each oligo.

After a plate is used for printing, I dry it down in a speedvac vacuum centrifuge with a plate-holding rotor. The oligos store better dried-down, and then I don't need to worry about evap over time, etc.

Then in subsequent resuspensions, I add the volume that should cause the remaining oligo to be at ~40pmol/ul again. I do this calculation by assuming that for each inking, the wells lose 0.5ul of fluid (the pin takes ~0.3ul, but we assume 0.5ul total loss to account for evaporation – based on personal communication with Kevin Visconti, Schoolnik Lab, Stanford. However, it's **always** good to check a few wells at random after a printrun to see what is really left, since your loss will depend upon the ambient humidity, and how long your plates were in the stacker – e.g. if you put them all in at once they'd all be exposed (with lids but no Al-foil seals) a lot longer than if you put them in one or two at a time). Thus, if I had a new plate with 10ul/well and used it for one bed of slides that required two inkings, I would assume the new resuspension volume the next time I used the plate would be 9.0ul.

For the total volume of printing buffer required, I multiply the volume per well, by the plate size, by the number of plates. So, 9 ul/well * 384 wells/plate * 15 plates = 51840 ul = 51.84mls, just over a Falcon tube's worth.

2. Printing buffer:
For the first resuspension, I use 3X SSC. For subsequent resuspensions (after the plate has been used and dried down), I use 0.3X SSC, to account for the small amount of salt lost during evaporation (this is what I learned from the Schoolnik Lab). I have a 20X SSC stock (Ambion).

In addition, in recent printruns I have been adding the co-spot oligo to the print buffer. I make a 100pmol/ul co-spot oligo stock solution, in 3X or 0.3X SSC. Then I make 1pmol/ul working solution, so I dilute 1:100.

Primary Resuspension Fluid (first time a plate is used):

final concentration	recipe per 40ml
3X SSC	6mls 20X SSC
1pmol/ul co-spot oligo	400ul of 1nmol/ul co-spot oligo
water	33.6 mls Ambion water

Secondary Resuspension Fluid (each subsequent time a plate is used):

final concentration	recipe per 40ml
0.3X SSC	0.6mls 20X SSC
water	39.4 mls Ambion water

195

3. Using the *aliQuot* robot by Genetix

1. Sonicate the disassembled manifold:
       in 3% aQu clean for >15 minutes
       in MilliQ water for >10 minutes
       in 95% EtOH for >15 minutes

2. Replace manifold (see diagram in manual if necessary).

3. Run 50mls of 3% bleach (1 Chlorox : 1 MilliQ; Chlorox is 6%) through the robot, letting sit in the tubing for ~15", per the online recommendations for removing DNA.

4. Run 3 x 50mls of MilliQ through the system.

5. Rinse with 1 x 50mls of 80% EtOH.

6. See page 13 of the aliQuot instruction manual, for the section "Running a Filling Routine".

      Adjust the manifold's start position, dispense height, and tilt, in order to minimize splashes. Start position is adjusted in the software, the height and tilt are adjusted mechanically.

      Use a dummy plate of the same make as you'll be using to test the settings out to see if there's splashage.

      PCR mode dispenses volumes as multiple aliquots of smaller volumes, to decrease splashing. Accessible in software.

7. Before using real plates, test aliquotting accuracy in both of two ways:
      i. weigh a plate before and after aliquotting into it.
      ii. use a pipetteman to test the volume several wells of different rows, since each row is filled by a different pin.

8. Note: The bottle fill type is a 50ml Falcon-type tube, BUT Falcons, Fisher brand, and BD brand do not work. A Greiner tube is in there currently, and we have a limited supply of other Greiners. They don't all work smoothly either. In fact, if you undo the existing tube it can be very difficult to get back on. So, I clean the tube thoroughly and then use it for dispensing if I can. Also, I keep it screwed in to the black connector piece, and unscrew that piece instead from the arm, whenever possible.

The dead volume (volume taken up from the Falcon tube before it gets to the manifold) of the aliQuot is ~3mls, as stated in its instruction book.

Side note: I tested evaporation from a 384-well plate, unsealed but with plastic lid (I think), and it ended up averaging ~ 0.05ul per well per hour... with the caveat that perimeter wells evap. faster than internal ones, and that the evap process should be non-linear. This was calculated from an 18-hour benchtop exposure with 10ul per well of 3X SSC, and September humidity. (So the total loss over 18 hours was 0.9ul).

196

Instructions for Doing a Printrun of 70-mer Oligos in 3X SSC on our Genetix QArray2 Arrayer

Quick checklist for starting a run:
- ☐ fill and start external humidifier
- ☐ turn big red power knob on right side of machine
- ☐ press the red reset button on the front
- ☐ turn on the computer
- ☐ start the software
- ☐ check the water level in the nebulizer (see below), and turn on
- ☐ wipe down inside with 70% EtOH, and spray with airgun
- ☐ fill the water and ethanol bottles, and pull them forwards against the front bar
- ☐ you MAY need to refill these during the run, check every few hours!
- ☐ make sure the waste bottles are empty
- ☐ you MAY need to empty these during the run, check every few hours!
- ☐ check for pinched tubing
- ☐ sonicate & dry pins (see below)
- ☐ load pins
- ☐ load one test slide, and fill remainder of column
- ☐ vacuum on
- ☐ print test (1 slide, 2 fields)
- ☐ check test slide, if OK, load blotting slide and print slides
- ☐ spin down print plates and load into right-hand stacker (recall A1 goes front-right!)
- ☐ load correct protocol in program, confirm parameters are correct, check data tracking, and start

Quick checklist for ending a run:
- ☐ remove plates, re-seal and freeze or start drying in speed-vac
- ☐ turn on light, and use head icon to remove the pins, and start them washing
- ☐ turn vacuum back on, affix labels to slides, then turn off vacuum, remove slides and place on clean surface to cut label strips
- ☐ turn off the software
- ☐ turn off the computer
- ☐ turn off the machine
- ☐ turn off the external humidifier
- ☐ pull the arraying head into the middle of the bed, close door.

The Physical Set-UP:
The pins:
Officially, the 150um tips produce a ~190um spot, use when spotting densities reach 20,000-30,000
The 75um tips produce a 90um spot

Both diameters are "regular volume" *unless* otherwise specified – both come in low volume options. <mark>Regular volume is approx 250nl per inking</mark>, low volume is 100-150nl.

In the pin boxes, Cheryl doesn't keep the rubber stoppers on the bottom of the pins, she just pushes them in gently with the top parts flush with the foam so that the pin tips are well clear of the foam. She also numbers the pins in the foam and keeps them in the same order, for tracking purposes

<u>To wash the pins prior to loading in the machine:</u> Place them in the specially-designed tip-washing manifold/stand. Place the stand into a beaker with an approx. 2% solution of aQu clean, or dilute detergent. Place beaker in sonicating water bath. Sonicate 10-15 minutes. She says she RE-USES the 2% aQu solution several times. She also bought just the smallest sonicating water bath that VWR sells and uses that.
Then rinse in Milli-Q H2O, sonicating, for 10-15 minutes.
Then either air-dry, or dunk in EtOH.
*If the pin is persistently clogged, she heats up the 2% soap in the microwave and lets the pins sit sonicating in the warm soapy water.

You NEVER should need to clean the Head in the microarrayer, which actually holds the pins. It's made inside of ball bearings so don't *ever* take it apart! Maybe, if the pins aren't sliding in smoothly, use compressed air, gently.

<u>To load pins:</u> Clicking the head icon in the software will bring the head to the front left of the bed for loading. The top of the pins are not radially symmetric, they are rounded and then have one straight edge, this edge should sit flush with the metal bars on the top of the head.

<u>The plates:</u>
Genetix X7022, which has covers and is V-bottom. She recommends no lower than 4ul in them, but has gone as low as 2ul.

<u>Loading plates:</u> Recall they go in REVERSE ORDER (from one prespective), with the first plate being on the BOTTOM of the "in" stack. Also, they go BACKWARDS, with the A1 well going in the bottom front right!! Plates go into the right-hand stacker. Twist the knob to right to lock stacks into place.
The stackers are interchangeable.

Evaporation from stackers. Usually not a problem, but one customer put a plastic bag over the stackers and shoved a humidifier tube up inside, but that was for a 70-plate 30,000-spot run.

<u>The machine:</u>

The tip washing manifold in the arrayer is a flow-through design, first H2O, then EtOH. Filled from containers underneath, and each into their own waste container. We filled with Milli-Q water and **80% EtOH**. The smallest container underneath is the equalization bottle – you should *never* see liquid in there, if you do call Tech Support.

Fill the humidifier, from the capped opening on the front right corner of the platter, inside the arrayer hood. Fill with DI-H2O or MilliQ. It takes about two liters to fill up, and use a funnel to fill it, and fill it until the level reaches the *bottom* of the plastic aperture through which you're filling. There's a digital display on the front right of the machine that controls the humidity, and holds it at +/- 2.0%. For a run in the winter set to 55-65%. The ambient can get down to 8% in the winter! If the humidity is low, the spots can bleb together, and if it's really low, the spotting solution can dry in the pins and then those pins won't print those spots. Use 1-2 external humidifiers in the room as well, with door closed.
**Talk to facilities and get the fan turned down in that room during the winter.

On the START page of the software, clicking the ☐ "Reset Outputs After Run" will stop the humidifier after the run is completed, which is good so that it doesn't run dry – especially a problem in the winter – **if it runs dry the motor on the pump can burn out**.

Loading slides:
You have to fill up an entire column of the bed with slides to get a vacuum seal for that column – you don't have to print on them all though, you can just use junk slides to finish filling a column in need be. You can control air flow to each column separately.
Always front-right justify the slides (except for the blotting slide(s) see below), e.g. using forceps, and then turn on the vacuum (icon at top bar of program).
The knobs at the front of each column control the vacuum for that column.
When the vacuum didn't come on, Cheryl tried bleeding the line at the thingie on the right side of the machine, lower half, where there's a pressure gauge – she bled the line here (by pushing something in?) until the compressor came on. That didn't fix it, but seemed to be her first trouble-shooting step. Dave had to come and fix it later.

Cleaning:
before each run she wipes down the inside – walls, everything – with EtOH and kimwipes. She also blows compressed air *gently* down the grooves on the slide-bed, pushing dust etc. towards the back, and then wipes across the back. If wiping the slide ruts themselves, then wipe right-to-left, since the left-hand side of the slides is less important since that (based on our current config) is where the labels will go.

The Software Set-up:
you can store protocols.

WELCOME tab (meaningless)
DESCRIPTION tab: lets you gave a name and write lab notes for the printrun
HEAD CONFIGURATION tab
      you can use a total of 48 pins
      they have many configs programmed in the pull-down that let you work
from either 96-well or 384-well source plates
      you can set up novel configs as defined in a config file
      we choose e.g. "16 pin Microarraying head" (default is 384-well)
      you don't need to tell it what diameter pins, it doesn't care
      it shows you the *correct* way to load the pins in the head, based on the
config you've selected. Follow its instructions!

SOURCE tab:
Plate Holder: Stacker Source Plate Holder   use stacker ☑
Plate type: 7022
Total Plates: 1, etc. (If >70 it pauses to let you refill)
Source order by:
      ○ columns (means the head dips into the source plate proceeding from
A1→ P1, so A1, E1...)
      ● rows (means the head dips into the source plate proceeding from A1→
A24, so A1, A5, A9)

SLIDE DESIGN tab:
slide: 3" x 1" (16pins/ 4fields)
      you match this to the actual number of pins being used, and the "field" is
the # of times that that printhead could physically fit onto that size slide; e.g.,
using the 48-pin head config., it could only fit on the slide once.
      with a 16-pin config you can actually print at a higher density
field layout: can organize replicate fields, etc.

** double-click on any of the spots to open a new screen, allowing you to edit
more parameters of slide design:

      spot view:
            ○ layout   estimated spot size: 200 for 150um tips,
90 for 75um tips
            ● actual (click actual)   (she nudges this a little higher so
there's good spacing)

      Pattern dimensions tab:
            calculate with calculator,

384 pins * 6 replicates / 16 pins = 144 spots, = a 12x12 matrix given the entered matrix, it automatically gives you the max. pitch for the row and columns; you can decrease the pitch to bring the spots closer if desired.
e.g. if the pitch =300, and the spot size = 200, then there's 100um between spots, which is fine

Fill tools tab:
Replicate type: replicate count = 6
    ○ Adjacent
    ○ Sector   where to start (diagrams):
    ● Cyclic   (choose upper left prob)
    ○ Random

Statistically, random really is the best

If using marker spots from a separate plate, click on markers, then hit remove all, then highlight the desired spots and type M, these then become red, designated as marker spots, and then when you fill in the spots it will assign the pattern around these marker spots, leaving room for them separately.

SLIDE LAYOUT tab:
    # slides: _____
    # blots:_____ (uses a sample slide location, and can blot multiply on a single "blotter" which in this case is just a discard slide. You can set the blotting pitch farther apart and overprint the same blotting spots
        □ change blotters   blotters to use: □ (this lets you change the number of blotters, it will pause the run for you to change blotters)
        Blot overprint method: □ same sample
                           □ no overprint

slide order: (diagrammatic)
can drag slide layout around is you want to . The layout can become important if you want to do humidity testing, arrange the spots to be printed around the edges to check the corners of beds.

PRINT tab:
Max stamps per inking: (approx. 200-300 in SSC)
stamps per spot: = 1.
    You can change this if you're doing e.g. protein microarraying. – can overstamp, or stamp in an offset circle centered around primary spot, etc.
Stamp time: 0
    She does 0 but you can increase it 20-30 (all in milliseconds) if you're seeing non-uniform features, or have a more viscous fluid.

Inking time: = 2000 (this is appropriate for regular volume pins, you can decrease the number for lower volume pins.

Print depth adjustment: she adjusts this so the pins *just barely* hit the slide - *very* tough to see, but makes a slight tapping noise during a run. The manual says to do this through Datum Points, and then use the print depth adjustment to just vary if there's a known alternate slide type you use, etc.

TOUCH OFF tab:
won't need this for SSC. It's used for more viscous buffers, to wick off the excess solution from the *outside* of the pins, by touching the surface of the liquid in the plate.

STERILIZE tab:
Water – 4 each at 1000ms wash time, 0 dry time, 500 wait time
EtOH – 1 – at 2500ms wash time, 7000 dry time, 500 wait time
You can bring the EtOH up to 3500ms and also up the water times if you're starting to worry about cross-contam, but these are the params. she likes

(there's two diff washing philosophies, one as above, the other with longer washes, eg 1 water wash at 5000ms, same EtOH.)

DATA TRACKING tab:
First, on the desktop, change the comma-delimited .csv file from Excel to .txt, then open the separate Data Tracking Program. The username: dtuser, password: dtuserpw

admin username: sa, pword: genetixsapw

Tools → Import Process File, Files of type Qsoft... etc.
Close Data Tracking Program, and back in Qsoft under Data Tracking tab:
File name: (for gal file)   File format: gal4.1 works
Open Groups → OligoPrototypeArray (whatever group you want, you've loaded)
highlight the plates wanted, and click Add.
the TOP plate in the software = the first plate out of stacker = the BOTTOM plate in stacker

Always backup the Database before reading new (e.g. rubbish) data in!

BARCODING tab: N/A

START tab:
○ Normal
○ Test plate (inks ONLY from the first 16-pin quadrant, just to print a few slides)
○ Print Test (inks and then prints onto just the front left slide, without re-inking. Let's you see if all the pins are clean, see the numbers of times you can print from a given inking, and the # of times you might need to blot, etc.)

○ Data tracking export only (just generates the gal file for a given data tracking import)

Other Important thing on the Software:

If you ever want to just get the gal file made for a given run, or made alone, you find it through: My Computer → C → Program Files → Genetix → QArray → logs → gal

Configuration Icon on Toolbar → Defined Objects → 3x1 (16pins/4 fields) (double click)

scroll down:

offset field 1X	this controls how far from the skinny "top" edge the field starts. She changed it from 1000 to 1500, for the purposes of the print test.
offset field 1y	this controls how far from the long "side" edge the field starts. She changed it from 3900 to 4100, eventually to 4300, to get the field more centered on the slide.
(offset field x, y	just control the distance between fields)

Inking depth: This is how deep into a plate you go. She has it set for their plates, and set for 4-10ul volume. If we go above 10, to 15 or so, it's worth changing the defined object inking depth so we waste less oligo on the outside of the pin. Currently it's set for about 200um above the bottom of the plate. You can change the inking depth for a new defined object, and save, or you can change it within a given run only.

Print depth: This needs to be done for particular slides. Go to **Configure Datum Points** in the toolbar, and set the heights for the slides in **each column** empirically. To adjust the print depth, negative numbers = up higher.

She saved a protocol called "Print Test" which prints two fields, and prints one slide only. She changed the field layout for it.

Hammer and Screwdriver icon at top = Diagnostic button, lets you move things around, etc. You can just drag the head where you want it, make a wash happen, etc.

Slide layout for first run was:
        sample slides: 14
        # blots: 5
        blot pitch: 300
        # overprints: 1   BUT check off No overprint, below.

Water = 2000 wash time, 4 replicates, 0 dry time, wait 500
EtOH = 4000 wash time, once, 10000 dry time, wait 500

save routine icon at top, looks like save as (double disk)

To back-up datum points and protocols:
Robot config → Database → Back-up   (on C drive, → Qsoft back-up)

Maintenance and Troubleshooting:

If the arrayer is not going to be used for >= week or so, she says to empty the bottles and leave everything dry, so that nothing grows. She also says you may want/need to wash the tubing once in a while: says some customers have run a mild fungicide through and then flushed with water – this is under I/O diagnostic section, Microarraying Wash, cick on Port 1 (=wash), Switch, and Vacuum, and then the wash runs continuously through both lines until you stop it.

To test if the pins are drying post-wash, and if they're washing enough or if there's carryover, cut a small piece of e.g. nylon and tape it on a slide. Then print (print test) on it and see if it's wicking off moisture, and if it's colored (use a dilute blue food coloring in your spotting buffer to see if there's carryover). Use an empty source plate if you're just trying to see if the dry time is sufficient – it needs a plate in there to go through the printing motions.

Blotting:
5-10 blots OK for 3xSSC
she set for 5, all on 1 slide (program figures how many slides automatically) the blot slide doesn't stick to field size but starts at far left hand side – for this reason, this slide should NOT be bottom-right justified but be a little more centered!

Troubleshooting: back pins not printing, tho freshly cleaned. Tried washing tray again, still no. Switched pins front-to-back, and back half still not printing – so pins are OK. Checked balance of head and bed with a mini-balance. Turns out head itself is not level.

Cheryl's free advice on plate-sealers (any gunk left on top may prevent top from coming of plate cleanly, and screw up the machine):
- small hand-held heat sealers units (for sealing foil) from Marsh (now bought by Fisher)
- or HyperTask, small local company, have foil plate-sealers she's liked and not had probs with.

U.K. Tech Support (US toll free #):
1-877-436-3275
favorite dude: Tim Roberts, ext. 4796, **expert** on QArray2
tim.roberts@genetix.com

U.S. Tech Support:
1-877-436-3849
spoke previously to Joe Jordan there.

Another local tech/rep person is Ken Adams, 617-549-6050

Array Post-processing Protocol

Adapted from DeRisi, Schoolnik, and Somero Lab protocols (developed by Andy Gracey), DeRisi microarray course notes, and personal experience.

Goal: After the arrays have been printed with DNA spots, this protocol is used to bind the DNA more tightly to the slide, remove excess DNA that hasn't been bound, and block any free lysines on the poly-lysine slide coating (those free lysines are "sticky" and if not blocked they could non-specifically bind labeled probe during an array hybridization.) Since slides age more quickly after they've been post-processed, it makes sense to just post-process a small batch at a time, as they are needed.

Required Equipment *most of it in my bench cupboards*

slide rack(s)	piece of cardboard)
a 1L and a 50ml Erlenmeyer flask	centrifuge with adapter for slide rack
2 3L beakers	diamond scribe etcher (in my bench
50ml Erlenmeyer	drawer)
rotator-table	powder-free gloves
humid chamber (Sigma H6644)	UV-crosslinker (Stratagene above Steve's
heating block	bench)
dust-free board for cross-linking (e.g.	

Required Chemicals

20X SSC, 10% SDS

1M Sodium Borate, filtered (make from Boric acid, adjust pH with NaOH): Boric Acid's FW = 61.83

1-Methyl-2-pyrrolidinone (anhydrous, 99.5%, FW=99.13) – Sigma M6762-1L - **do not use if it appears yellowish**

Succinic Anhydride, 6g (99+%, FW=100.07, a moisture sensitive irritant) – Aldrich 239690-50G - **do not use if it has been exposed to moisture. Keep in a dessicator, sealed with parafilm.**

95% Ethanol **(do not make from 100%, it is made differently) – do not use if it is cloudy or has particulates**

Before you start: *prep work can take up to ~ an hour*

** always rinse all glassware, etc. with Milli-Q water before use to remove dust**

** always use powder-free gloves when working with arrays!**

1) take Methyl-pyrrolidinone out of fridge and put in hood to come to room temperature.

2) make up *fresh* Na-Borate *(do this first so it has time to cool before pHing)*

you'll need 25.71mls, so make e.g. 40mls and then discard excess:

in a 50ml Erlenmeyer, with a small stir bar:

35mls Milli-Q H2O

2.473g Boric Acid

mix on high, with heat, until dissolved

cool to room temp

adjust to pH = 8.0 with 10N NaOH (takes >20 drops, so start there. Do NOT work back with HCl if you go past. Start fresh with new solution.)

check volume – add Milli-Q H2O up to 40mls

filter through 0.2um syringe-filter, into a 50ml Falcon tube
3) wipe down bench area with EtOH
4) heat up heat block on bench to at least 90°C
5) place in the chem. hood:
- stirring plate for preparation of blocking solution
- shaker, with secondary containment tray taped down
6) make up pre-wash solution: 1X SSC/ 0.1% SDS
for 700mls (to use for 1 rack of slides, in a 3L beaker):
658mls Milli-Q H2O
35mls of 20X SSC
7mls of 10% SDS
mix briefly with stir bar, remove stir bar, and cover beaker.

Step 1. Rehydrate the slides
This step recovers the spot's circularity, and decreases "donuting" of the spots.
a) wipe down the bench with 70% EtOH and kimwipe (*not paper towel*) to remove dust
b) put warm tap water into humid chamber, and place chamber next to heat block
c) Rehydrate slides by inverting (array side down toward steamy water) them over warm water in a slide-staining chamber – **don't use much water or it can splash up on slides**. Watch until the spots become glistening and juicy, **3-10 minutes**. (Under-hydration causes too little DNA to stay adhered to the spot during the subsequent washing, and over-hydration causes the spots to be blebby. AG says 2-3" is usually sufficient, the DeRisi protocol says 1-10".)
Be careful not to allow the water to touch the array.
d) Immediately flip them (array side up!) onto a heating block (inverted, about 90°C). Watch the steam evaporate. When the array spots dry in a rapid wave-like pattern, remove them from the heating block. This takes about **5 seconds**. Do 1 slide at a time.

Step 2. UV cross link
This step helps the DNA stick to the poly-lysine.
a) Place the slides, array side up, on a flat, dust-free board that fits into the UV cross-linker (I use a pre-cut piece of cardboard that I keep in my cupboard).
Do not put them on a seran wrap surface since the slides stick to it.
b) Irradiate with 600 uJoules UV light – press the "**ENERGY**" button and then enter **600**, then press "**Start**". It will count down and beep when done. (Andy does his at 650 and does them in the metal rack, not laying them flat.)
c) Before the next step, **etch the slides** with a diamond scribe (in the top middle drawer of my bench) to demarcate where the array is – after the pre-wash the spots will become invisible!

Step 3. Pre-Wash (a.k.a. the "shampoo method")
This step removes excess, unbound DNA to prevent "pluming" of the DNA out from the spots. Some protocols suggest skipping this step if the initial spots are small, however skipping this would then require an extra step later, which is not in this protocol.

207

a) Place the slides in the slide rack and secure them with a strip of metal wire on top, or rubber bands. If rubber bands, then position as close to the rack edge as possible. (AG uses rubber bands.)

b) rapidly plunge slides into 1x SSC / 0.1%SDS for 30 sec

c) Gently wash slides with *lots* of Milli-Q water, swish rack back and forth (something like 5 consecutive gentle rinses of 30+ seconds each). Let sit in water briefly while preparing blocking solution. (Pre-wash protocols will vary on the details, but all have the same general principle. Some also spin slides dry before blocking, but AG doesn't, and drying doesn't appear necessary or beneficial.)

Step 4. Block free lysine

wear a lab coat when working with methyl pyrrolidinone

a) In a 1L beaker, add **8.643g** of succinic anhydride into **526ml** 1-Methyl-2-pyrrolidinone while stirring with stir bar on stir plate.

b) *As soon as* the solids dissolve, (though they may not dissolve completely, some protocols warn – mine always has dissolved), quickly add **23.57ml** of 1M Na-Borate pH 8, and pour the mixed solution into a 3L beaker.

c) *Quickly* place the slides (in metal rack) into the succinic anhydride solution (**do not pour the solution over the slides**) and plunge up and down for 60 seconds. Rotate at 60 rpm for 1 hour *if possible* – as little as 30 minutes probably OK. (AG's surface chemist friends say the process doesn't go to completion for about an hour, though some protocols call for as little as 15".) While blocking, set up 3L beaker with Milli-Q water on hotplate, in hood, and heat to boiling.

d) Remove the slide rack from the organic reaction mixture and place it immediately into the boiling Milli-Q water bath (some protocols call for room temp, feel free to try this out and see which works better, just make a note of it! I haven't noticed a difference, actually) and wash thoroughly but gently by swishing rack back and forth for 90 seconds.

e) Transfer the slide rack to a 3L beaker containing approximately 575mls of **95% ethanol (do not make from 100%)**, plunge slides to mix, and then carry directly in EtOH to the tabletop centrifuge.

f) Spin dry the slides by centrifugation at **150 x g** for **2 min**. Use a counter balance with the same number (& orientation) of slides in a rack. (Balance slides are in the top right drawer of my bench).

g) Carefully transfer the slides to a dry slide box for storage in a dessicator. Make sure the slide box is appropriately labeled.

h) Collect methyl-pyrrolidinone solution as waste for periodic EHS pick-up.

Array Post-Processing Form:

Date:_____   Person:_____

Which slides being post-processed (printdate, series):_____

       Total number =_____

Notes on solution making:

Notes on re-hydration:

Notes on pre-wash:

Time in to blocking solution:_____Time out:_____

When out of blocking solution, into HOT or COLD water bath? (circle one)

Notes on blocking steps:

Amplification and Labeling of DNA for Microarray Hybridization, using the "Round A/B/C" Random DNA Amplification Protocol

Goal: To amplify and label DNA prior to hybridization to a microarray, in a relatively random way. This protocol does not give linear amplification, but as DeRisi says it "is useful to compare relative enrichment between two samples." DeRisi reports that it has been successful in their hands for amplifying less than 1ng of genomic DNA. I have obtained results from as little as several hundred picograms of environmental DNA but a safer lower limit starting amount of DNA seems to be ~6ng per slide hyb (see caveats below).

Protocol History: I adapted this protocol from that used by Joseph DeRisi's Lab at U.C. San Francisco, and theirs was adapted from Bohlander et al. *Genomics* 13 (1992).

Overview: There are three stages of this protocol. In Round A, the Sequenase polymerase extends random primers with specific ends (Primer A) that have annealed to the template DNA. In Round B, conventional PCR amplifies the templates from Round A, using the specific primer (Primer B) which matches the 3' end of Primer A. In Round C, Primer B is used again to mediate rounds of conventional PCR during which modified nucleotides are incorporated for labeling. These modified nucleotides are typically either amino-allyl-dUTP, for indirect labeling, or nucleotides that are directly coupled to Cy dyes. I use the aa-dUTP indirect labeling so that is what is described here. There is less discrimination by the polymerase against the smaller aa-dUTPs than against large, bulky Cy-dNTPs.

Understanding the **Round A/B/C**
Random Amplification and Labeling Protocol
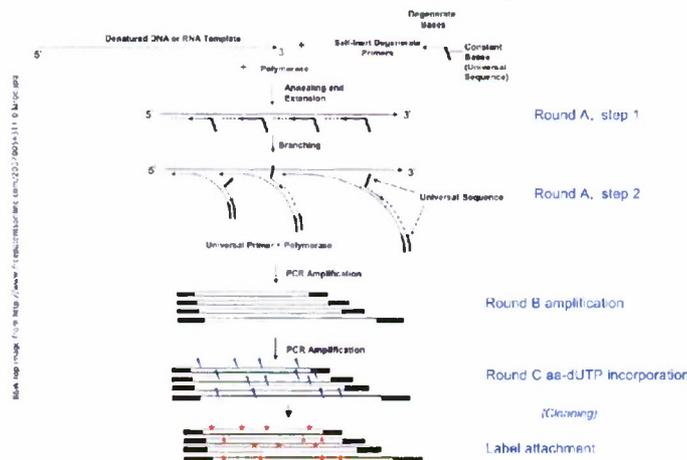
Precautions: This is a random-amplification protocol, which means that *ANY* DNA can be amplified. Therefore, use filter tips, wear gloves, and UV-sterilize your tubes along the way. Also, run negative control reactions. Also, because it's a random-primed multi-round exponential amplification, we might worry about stochastic skewing of the relative abundance of different organisms during amplification. The protocol partly accounts for this by subsampling each amplification as the template for the next round (shown to be beneficial to PCR evenness generally in Thompson *et al.*, 2004). In addition, I choose to **run triplicate amplification reactions and pool them prior to labeling**. Pooling multiple reactions has also been shown to decrease random biases introduced early during amplification (Thompson *et al.*, 2004).

Materials

Item	Supplier	Item #
General:		
Nuclease-Free Water	Applied Biosystems/Ambion	#AM9937
Positive control DNA		
In my case, *Halobacterium*	ATCC	#700922D
Round A		
Sequenase (13 units/µl)	US Biochemical	#70775
5X Sequenase Buffer	included	
Sequenase Dilution Buffer	included	

"Sequenase Version 2.0 DNA polymerase is a genetically engineered form of T7 DNA polymerase which retains extraordinary polymerase activity with virtually no 3'->5' exonuclease activity. It is highly processive, able to effectively incorporate nucleotide analogs for sequencing (dideoxy NTPs, -thio dATP, dITP, 7-deaza-dGTP, etc.) and is not easily impeded by template secondary structure."

"A" dNTPs = 3 mM each nucleotide
500 µg/ml BSA
0.1 M DTT

40 pmol/µl Primer A:	e.g. Proligo	N/A

5' - GTT TCC CAG TCA CGA TCN NNN NNN NN - 3'

Round B

10X Mg-minus PCR Buffer	to match the Taq	
(500 mM KCl, 100 mM Tris pH 8.3)		
25 or 50 mM MgCl2		
"B" dNTPs = 25 mM each nucleotide		
5 U/µl Taq polymerase	e.g. Invitrogen	
100 pmol/µl Primer B:	e.g. Proligo	N/A

5' - GTT TCC CAG TCA CGA TC - 3'

Round C

    Same as Round B except use modified "C" dNTP mix:
Their recommended recipe is:
        25 mM each dATP, dCTP and dGTP
        10 mM dTTP

15 mM aminoallyl-dUTP (or Cy-dUTP)	Ambion	#AM8439

However, they suggest that the ratio of aa-dUTP to dTTP can be altered/optimized. My optimized recipe is:
        22.5 mM each dATP, dCTP and dGTP
        9 mM dTTP
        11.75 mM aminoallyl-dUTP
For 100 µl this corresponds to:
        22.5 µl 100mM dATP, dCTP, and dGTP
        9 µl 100mM dTTP
        23.5 µl 50mM aa-dUTP

aa-dUTP structure (for L-A ☺), and the general amine-ester reaction employed for dye coupling:

Succinimidyl ester   Carboxamide

211

By Virginia Rich, graduate student in the DeLong Lab. vrich@mit.edu
Last modified 4/28/08
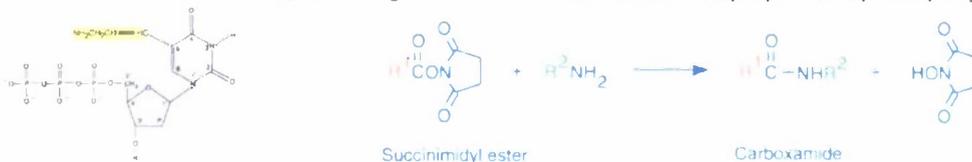
Protocol

Round A: Denature template DNA, anneal primers, and extend.
Time: At least 1 hour 20 minutes, increasing significantly based on number of reactions.

Round A, Step 1: First strand synthesis.
Each reaction receives:

Ingredient	Volume
Template DNA	7 µl
(e.g. 6 µl template DNA and 1µl positive control DNA)	
5X Sequenase Buffer	2 µl
Primer A (40 pmol/µl)	1 µl

Total Volume = 10 µl

To standardize things I prepare a master mix of 5X Sequenase buffer and Primer A, and then dispense it into my tubes – either 0.2ml PCR tubes or a PCR plate. For triplicate reactions, I dispense 3X of this master mix into the wells, add 3X of my DNA, mix, and then aliquot into three separate tubes, or three separate rows if using a PCR plate. This works well.

Use "vr-a" cycling protocol on "Goldie" thermal cycler**:
Heat 2 min at 94 °C
Rapid cool to 10 °C and hold 5 min at 10 °C.

*** I use this thermal cycle because I have programmed it to have approximately the correct ramp time for later steps. Other thermal cyclers have different ramping speeds and so will need to be programmed accordingly.*

With program paused at 10 °C and the tubes in the thermal cycler, add 5.05 µl **Reaction Mixture** to each reaction (having assembled reaction mixture in UV-hood):

Ingredient	Volume
5X Sequenase Buffer	1 µl
"A" dNTPs (3mM)	1.5 µl
0.1 M DTT	0.75 µl
500 µg/ml BSA	1.5 µl
Sequenase (13U/µl)	0.3 µl

Total Volume = 5.05 µl

Again, I make a master mix of this reaction mixture, in the UV-hood, and then dispense it at the thermal cycler into each tube or well.

Ramp from 10 °C to 37 °C over 8 min.
Hold at 37 °C for 8 min
Rapid ramp to 94 °C and hold for 2 min.
Rapid ramp to 10°C and hold for 5 min

Round A, Step 2: Second strand synthesis
Pause at 10°C while adding 1.2 µl of diluted Sequenase (1:4 dilution in Sequenase Dilution Buffer).

Ramp from 10 °C to 37 °C over 8 min.
Hold at 37°C for 8 min.
END

In PCR hood, dilute samples with Ambion Water to final volume = 60 µl (should be 60 – 10 – 5.05 – 1.2 = 43.75ul).

Round B: PCR amplification.
Time: ~2-4 hours, depending on # of cycles run.

Mix in a 0.2ml PCR tube, in the UV-hood:

Ingredient	Volume
Round A Template	6 µl
50mM MgCl2	4 µl
10X Mg-minus PCR Buffer	10 µl
"B" dNTPs (25mM)	1 µl
Primer B (100pmol/µl)	1 µl
Taq	1µl
Ambion Water	77µl
	Total Volume = 100 µl

Use "vr-b" cycling protocol on "Goldie" thermal cycler:
> 30 sec at 94 °C
> 30 sec at 40 °C
> 30 sec at 50 °C
> 2 min at 72 °C

Run 15-35 cycles, depending on the amount of starting material. I typically use 20 cycles.

If you run 5 µl of each reaction product on a 1% agarose gel, you should see a smear of DNA between 500bp -1kb. To minimize the number of cycles you run, the first time you're working with a new type of template they recommend removing aliquots of your reaction (of which you have extra to spare, don't worry) every 2 cycles or so and checking them on a gel – you want to use the minimum number of cycles that produces a visible smear of product DNA, and that still keeps the negative control lanes empty.

Round C: Incorporation of aa-dUTP.
Time: ~1-2 hours, depending on # of cycles run.

They recommend using 10-15 µl of Round B to seed the Round C reaction. I use 10 µl.

Ingredient	Volume
Round B Template	10 µl
50mM MgCl2	4 µl
10X PCR Buffer	10 µl
"C" dNTP mix	1 µl
Primer B (100pmol/µl)	1 µl
Taq	1 µl
Water	73 µl
	Total = 100ul

Use "vr-c" cycling protocol on "Goldie" thermal cycler:
> 30 sec at 94 °C
> 30 sec at 40 °C
> 30 sec at 50 °C
> 2 min at 72 °C

10-25 cycles can be run, I typically run 10 cycles.

Clean-up I:

Salts and Tris interfere with dye coupling, so before proceeding you must clean up the reactions. They recommend using a Microcon size-exclusion column to do this.

Add 400 µl water to the sample in a Microcon 30
Spin 14,000xg until liquid mostly drained. Empty collection tube.
Wash 1X with 500 µl Ambion water.
Concentrate to ~9.3 µl in Ambion water: 8 µl will be used for the labeling reaction, and ~1 µl will be used to Nanodrop the sample. I record the volume I was actually able to concentrate the sample to (it's tricky, and I can't usually get to 9.3 exactly) so that I can calculate my amplification efficiency if I want to, and also understand the comparability of my samples. Obviously, one wants the volumes to be as close as possible to one another between samples to permit the most comparability.

Also, I combine my triplicate reactions at this stage, pre-labeling. You can combine triplicates prior to Microconning but beware that if you pool the negative controls before you clean them they may be VERY slow to drain.

If doing lots of samples, instead of using Microcons, I use an ExcelaPure 96-well size-exclusion-column plate with the vacuum manifold. I run it at 10"Hg so as not to lose DNA <300bp. Note that this size exclusion cutoff is a little bigger than the Microcon-30's. SO, for any experiment or for experimental series you'd like to be able to compare, it would be advisable to consistently use one or the other clean-up method.
Wash 1 x 300 µl of Ambion water
Resuspend in ~30 µl Ambion water, transfer to a v-bottom 96 or 384-well plate.
Use the vacuum centrifuge with the plate rotor to dry down the DNA. Use e.g. the automatic spin with 2 hours vacuum spin, 1 hour at 45 deg. C. Then resuspend your DNAs directly in 0.1M NaHCO3 (allowing to sit at e.g. 60 deg C for 10", then vortex gently and spin).

Labeling:

8 µl aa-DNA
2 µl 0.5M NaHCO3 (0.1M final concentration in the DNA mixture)
    mix
        OR 10 µl aa-DNA resuspended in 0.1M NaHCO3
5 µl Cy dye (33 µg in DMSO)
    mix
incubate at room temperature in the dark for 1 hour.

Co-spot complement: If you are using the co-spot complement as well, you will have done a single separate labeling reaction of that, linked to Cy5. I've found that using ~1pmol of the co-spot complement oligo per array hybridization works fine. In my case, the co-spot complement that performed the best and that I ended up using was the "alien" complement oligo from Urisman et al., 2005.

Quenching: If using the co-spot complement, you'll combine the differently-labeled DNAs at the hybridization stage. You wouldn't want any residual uncoupled dyes to cross-label the wrong DNA. Although rinsing with TE will quench the labeling reaction, and should remove uncoupled dyes, for best practices you should ALSO use the traditional chemical quenching protocol step of adding 2 µl of 4M hydroxylamine to each reaction, mixing, and allowing them to sit in the dark an additional 15 minutes.

Clean-up II:

Now you want to remove unincorporated dye molecules. Use the single-column or 96-well size exclusion column plates, as before. Now, however, wash with TE. The TE helps inactivate the dye conjugation process.

    Add 480 µl (or 280 µl if using Excela-Pure plates) to your samples.

    Transfer to the columns.

    Spin.

    Wash columns 2x 500 µl TE (or 2x 300 µl if using Excela-Pure).

    Concentrate to ~19 µl in TE, or more if you're doing triplicate slides.

Note: do not use the Excela-Pure plate to clean the co-spot complement. This oligo is smaller than the cutoff of the Excela-pure columns.

Hybridizing Cy-labeled DNA to homemade PLL oligo microarrays

Protocol History: I adapted this protocol from that used by Joseph DeRisi's Lab at U.C. San Francisco.

Materials

Supplies:
0.5ml tubes
Lifterslips, Erie#22x40I-M-5516, available as VWR #48382-242
Heat block with 0.5ml-tube block
Hyb ovens with accurate and precise temperate
Monitoring digital thermometer
Hyb chambers (e.g. from Genetix)
Microcentrifuge
Centrifuge with plate-spinning rotor
Slide racks

Reagents:

20X SSC	Applied Biosystems/Ambion	# AM9770
HEPES, make to 1M, pH7	Sigma	#H4034-25G
Ambion H20	Applied Biosystems/Ambion	# AM9937
polyA, make 10mg/ml	Sigma-Aldrich	#P9403-25MG
10% SDS	Applied Biosystems/Ambion	# AM9822

Protocol

The total volume of your hybridization reaction will depend on the size of your lifterslip. I am currently using mid-sized lifterslips with a recommended volume of 29µl; I use 30µl. For the prototype array, I used smaller lifterslips for which my hybridization volume was 25µl.

For one reaction:	For 3.1 reactions:	For all reactions:
	Multiply column 1 values by 3.1 if you are hybridizing triplicate arrays for each sample.	Multiply column 1 or 2 values (depending on if doing triplicate or single arrays) by ~110% of the number of samples you've got.
DNA:	DNA:	
19.83µl Cy3-DNA	61.47µl Cy3-DNA	
1µl Cy5-cospot-complement	3.1µl Cy5-cospot-complement	
Mix H1:	Mix H1:	Mix H1:
4.49µl 20X SSC	13.92µl 20X SSC	20X SSC
0.62µl 1M HEPES, pH 7.0	1.92µl 1M HEPES, pH 7.0	1M HEPES, pH 7.0
2.24µl Ambion H2O	6.94µl Ambion H2O	Ambion H2O
Σ = 7.34µl	Σ = 22.78µl	
		For H2, multiply values by ~130% of the # of samples.
Mix H2:	Mix H2:	Mix H2:
1.22µl 10mg/ml polyA	3.78µl 10mg/ml polyA	10mg/ml polyA
0.62µl 10% SDS	1.92µl 10% SDS	10% SDS
Σ = 1.84µl	Σ = 5.7µl	
Σ = 30µl	Σ = 93µl	

– Make up Mix H1 and H2, mix each well

- Aliquot Mix H1 into 0.5ml tubes, one for each sample. Thus, if hybing 3 arrays per sample, make hyb mix for all three slides in same tube.
- Add Cy5-DNAs, if relevant, and add Cy3-DNAs. Mix well.
- Add H2 into each tube, and mix thoroughly by pipetting.

- Heat at 100 deg. C for 2" if 30µl, 4" if 93µl
- Spin max speed 1"

- Load samples onto arrays, quickly, and load arrays into pre-heated hyb chamber (with water in base).
    Note: If doing many hybridizations at once, I will heat, spin and load tubes 1 chamber at a time, so 9 or 10 slides at a time. You don't want your DNAs to cool off too much between when you heat them and when you load them on the array and get them into the warm chamber. So how many you do at once partly depends on how fast your technique is.

- Hyb arrays overnight, >= 12 hours.

Washing Arrays:

Prepare in bowls:

Wash Solution I:	Wash Solution II:
18ml 20X SSC	1.8ml 20X SSC
1.8ml 10% SDS	598.2ml MilliQ H2O
580.2ml MilliQ H2O	

- Remove 1 hyb chamber at a time from hyb oven. Quickly, transfer slides from hyb chamber to a slide rack submerged in Wash Solution I.
    For doing many slides at once, I have two bowls of Wash Sltn I set up, and use the first for gently swooshing off the coverslip and have the slide rack in the second (gentle coverslip removal can be tricky with the slide rack in the same bowl). To remove the coverslip, I hold the slide horizontally and submerge it into the solution, moving it down while tilting it forwards and moving it back, all at once with a swoop of the wrist. This allows the coverslip to float off cleanly with minimal chance of it scratching or touching the array as it's coming off. In theory. Experiment and find your own best way to do it - sometimes the PLL coating can be very delicate and you really want to be as gentle as possible.

- Rinse slides in Wash Sltn I vigorously for 30 sec by plunging slide rack up and down.
I use a plastic tub around the bowl for this because I always splash a lot.

- Transfer the slide rack to Wash Sltn II, blotting base of slide rack on kimwipes to remove excess SDS.
- Rinse slides in Wash Sltn II vigorously for 30 sec
- Cover the bowl with foil so it is dark, and transfer bowl to rotator. Rotate max speed allowable (keeps slides covered still and doesn't splash) for 5".
    Note: If I'm doing a lot of slides at once, certainly if doing half the rack or more, I transfer the slides to a clean bowl of Wash Sltn II after 2.5".

- Quickly transfer the slide rack to a plate-rotor in the centrifuge, into a rotor-cup lined with large kimwipes. You may blot the slide rack on other kimwipes as you transfer it. Make sure you have set up balance slides during the previous step - you want the slides to spin ASAP once they are out of the liquid. Spin the slides 90 x g for 5", or until dry. Spin the slides with the array face facing Into the direction of spin.

Note: Make sure the centrifuge is very clean before you do this. Dust is your enemy - it autofluoresces.

217

Extracting data from array tiff files in GenePix 6.1
Note: some of these features are not available in GenePix 6.0.
Presumes some prior experience with GenePix.

1. Make a master Settings file:

   a. Open first array image, (= a tiff file).

   b. Load the appropriate array list for this array, (= a gal file).

   c. In block mode, select all blocks and align overall array.

   d. Double-click on any block, and adjust spot diameter (with "apply to all" checked off) so that it is a little smaller than the average spot size on your actual arrays (in my case, 150um).

   e. Align each block individually, more precisely.

   f. Adjust the auto-alignment specs:

      Set **Options Box** (Alt + I), **Alignment Tab** as follows:
      Click on "Find circular features"
      Click on "Resize features during alignment"
                    between e.g. 70% and 150%
      Click on "Limit feature movement during alignment" to e.g. 40um
      Toggle for unfound features to be "Unflagged"
      (No CPI threshold – default)
      (Check Align Blocks, estimate warping and rotation – default)
      (Automatic Image Reg Max translation 10 – default)
      (No sub-pixel reg. allowed – default)

   g. Then save all of this by going to Save Settings, which will create a gps file. This will be your master Settings file to use on the other arrays *from the same print & hyb date.*

2. <u>Auto-align the array grids for each array scan:</u>

   a. Click on **Batch Analysis** tab in the main GenePix window.

   b. Click "Add" and select your tiff files to process

   c. Select all tiff files and click Add gps file, and choose the master gps file for this set (the file you just created above).

   d. Uncheck "Analyze", leaving "**Align**" checked only.

e. Click on "Configure Alignment" box, and within it check "Find Array" and "Align Features" ONLY - UNcheck "Find Blocks". (This is because on our high-density array, several of the blocks are very close together and so block-finding gets confused. With the master gps file tailored to each array printrun, block-finding isn't needed anyway for good gridding if the other two alignment types are used.)

f. Click "Go", with "All at once" checked. Depending on how many files you have, this can take several hours.

3. Check the new alignments.
   a. Use the results browser window (this will come up automatically during batch the alignment. If you close it accidentally, you can reopen it by clicking within the Batch Analysis tab, click on the lower right-hand array-like icon), to check the new gps alignment files it created for each tiff. Clicking on a gps file within the browser box will take you to the Image tab of the main GenePix window, and will load the tiff and its associated newly-created gps file.
   b. Manually inspect EACH gps and tiff pair: manually adjust stray features, and flag areas of surface PLL peeling, or excessive background, as "Bad", to be discounted from further analysis.
   c. Save each gps file using the same name as before – replace the previous version.

4. Extract the data (you can do this immediately or at a later time):

   a. In the **Batch Analysis** tab, delete all files.

   b. Click on "Add", then select all tiff files AND their associated gps files at once and click OK – this will link each file correctly.

   c. Check "Analysis" and uncheck "Alignment".

   d. Click "Go" and "All at once". Again, this may take a while depending on how many files you're doing.

   e. You may wish to also use a **flag feature query** – e.g., to automatically flag as bad all spots in areas of background peeling. To set this up, you must first go into the Results Tab and Click on Flag Features, and make a new query to suit your needs. E.g., I created a query called "test" which for most of my arrays successfully flagged many of my missing features, using the following syntax:

   [B532] <= 100 AND [B635] <= 100

This had the effect of removing features along the edges of the array where the PLL coating may have peeled away. Also, if there are large interior peeled sections, it removed those. HOWEVER, it was still unable to find smaller patches or scratches, or features on the edge of a patch that should have been flagged because either part of the feature itself was peeled/scratched or part of its local background was.

In addition, the "test" query I create works for most but not all of my slides – if some have unusually low background, then it artificially flags my data as "bad" even though the spots are there. So you must tailor this to your particular slides based on their background, and also judge whether to used it based on the homogeneity of your background among your slide set. For me, it actually wasn't worth using the auto-filter query since I was manually flagging each gps file for a variety of things anyway before extracting the data.

Another way to do it, computationally longer but perhaps easier depending on your particular slides, would be to have it autoalign and then immediatately analyze, using a "flag features as bad" query. THEN go through each results file, and you'll see which features have been autoflagged on the corresponding gps that loads. You can do additional flagging at that stage, re-save over the gps files, and re-run the analysis to save over the previous analysis files. Clunky, but may be worth it based on your particular specs.

Also, you CAN get a lot more sophisticated with your queries. For example, in order to auto-flag features (spots) that are partially peeled, or at the edge of a peeled region, or have some other aberration, you might like to is have a query that says e.g. "if e.g. >20% of the pixels in the feature, or background, are less than e.g. 100, flag as Bad". I asked Sandra Lew about it, and she said it's something you should be able to do with VBScript in the Results Tab under the Flag Features button. She recommends going to the GenePix Help, where there are chapters on scripting under the Index Tab, describing some commonly-used functions.

5. So, that's it - that's the full pipeline I used to get my results.

Currently I extract all the data the software will give me, in case I ever want to go back to a parameter I don't currently use but which ends up being important, but you can decrease the columns of data that you get if you so desire.

By Virginia Rich, graduate student in the DeLong Lab, vrich@mit.edu
Last modified 1/26/08

GenePix contact people:

The software technical details guru:
Sandra Lew, Sandra.Lew@moldev.com

The woman who updated our hardware and software: **Yvonne FitzGerald,**
yvonne.fitzgerald@moldev.com
She recommends that if we have any further problems with software crashes
(which we had for a while after she first installed 6.0 and 6.1 on our new
machine, before she uninstalled both and reinstalled 6.1), then we make the
noise to get a formal field engineer out here to look at the problem, because
she says we've exhausted her knowledge, so if the reinstall didn't work then
someone else needs to have a go. Since we just bought a new machine (new
computer + software package, in March 2008) we ARE under warranty for some
period, but I don't know how long – so if the problems re-occur then it would
be wise to get them seen to asap.

Our local GenePix sales rep, who has given us a loaner computer when the last
one broke: **David Micha**, david.micha@moldev.com

Interactive Excel Worksheet for Calculating Reagents Needs for Genome Proxy Array:

How many samples will you hybridize? Fill in highlighted cells (in this worksheet only), calculations are then automatic and propagated into the next worksheet, which summarizes the order you should place.

20

How many replicate A/B/C reactions will you run? (Three recommended, with pooling prior to labelling. Can then split prior to hybridization.)

3

How many replicate arrays will you hybridize?

3

Reagent/Consumable	Stock Conc	Amount per 120 slides	units	Total amount		
Making PLL slides						
Gold Seal Micro Slides	solid	120	N/A	60		
NaOH pellets	solid	240	g	120		
EtOH	95%	960	mls	480		5760 2880
PBS	1X	73.8	mls	36.9		0.5ul oper inking * 2 inkings = ~1.0ul, = 1pmoles,
Poly-L-lysine solution	0.1% (w/v)	126.72	mls	63.36		for aach of (384*15 wells = 5760 walls), par 85 arrays,
co-spot oligo	1 pmol/ul	8131.76	pmolas	4.0659	nmolas	so 5760 pmolas par 85 arrays
array oligos	40pmol/ul	56.47	pmoles	0.0282	nmolas	~1.0 ul of EACH per 85 arrays, so 40pmoles per 85 arrays
		Amount per 30 slides				Or, if losing ~ 0.68ul par wall, then for co-spot, =
Post-processing of Slides						
						2 inkings * 0.68pmolas par wall * 5760 wells = 7834 pmoles
succinic anhydride	solid	8.643	g	17.286		
1-Methyl-2-pyrrolidinona		526	mls	1052		11059.76 S.S299
Boric Acid	solid	5	g	10		And for "real" oligos, 2 inkings * 0.68ul par inking * 40pmoles/
SSC	20X	35	mls	70		= 54.4 pmoles per bed-full
SDS	10%	7	mls	14		76.80 0.04
EtOH	95%	575	mls	1150		
		Vol per one rxn				
A/B/C						
Halobactarium DNA	10ng/ul	1	uls	60		
Primar A	40pmol/ul	1	uls	60		
"A" dNTPs	3mM	1.5	uls	90		
DTT	0.1M	0.75	uls	45		
BSA	500ug/ul	1.5	uls	90		
Saquanasa	13U/ul	0.6	uls	36		
MgCl2	50mM	4 + 4	uls	480		
"B" dNTPs	25mM	1	uls	60		
Primar B	100pmol/ul	1 + 1	uls	120		
Taq	5U/ul	1 + 1	uls	120		
Ambion water	pure	43.75 + 77 + 73	uls	11625		
"C" dNTP mix	various	1	uls	120		
		Vol per sample - can pool replicate A/B/C reactions				
Labelling						
NaHCO3	0.5M	2	uls	40		
	33ug in 5ul					
Cy3 dya	DMSO	33	ug	660		
Ambion water	pure	300+ 50	uls	7000		
TE	1X	500 + 500 + 19	uls	20400		
Co-spot complement labeling:						
co-spot complement oligo	1 nmol/ul	1	pmol	60		
NaHCO3	0.5M	0.02	uls	0.33		
	33ug in 5ul					
Cy5 dye	DMSO	0.28	ug	5.5		
Ambion water	pura	2.92	uls	7000		
TE	1X	8.49	uls	20400		
Hybridization						
SSC	20X	4.49	uls	269.4		
HEPES	1M, pH7	0.62	uls	37.2		
Ambion H20	pura	2.24	uls	134.4		
TE	1X	3.5	uls	210		
polyA	10mg/ml	1.22	uls	73.2		
SDS	10%	0.62	uls	37.2		

So, for dNTPs that means:
per reaction, assuming 100mM stock solution:

Round A	Round B	Round C		
0.045	0.5	0.5	uls	1.045
		so, total:		62.7

What about aa-dUTP?
23.5 µl 50mM aa-dUTP par 100ul of C-dNTP mix

so for abova
# of samplas,
naed   14.1   uls of 50mM aa-dUTP

So, final tally of reagents' volumes required that you should order:

Reagent	Conc.	Amount		Supplier	Item #	Notes	Unit of Sale	Reagent cost per unit sold	Total cost	Cost per array	Other way to calc. cost per array

Note: for this first section, making and post-processing the PLL slides is done in batch. Thus, the "total volumes required" are under-estimates of what you'll actually use, since they represent the amount for the actual # of slides you're hybing rather than for the next larger batch size, which is how you will process them.

Making and post-processing the arrays

Reagent	Conc.	Amount		Supplier	Item #	Notes	Unit of Sale	cost per unit sold	Total cost	Cost per array	Other
Poly-L-lysine solution	0.1% (w/v)	63.36	mls	Sigma-Aldrich	P8920-500ml		500 ml	$222.00	$28.13	$0.47	
NaOH pellets	dry	120	g	Sigma-Aldrich	S8045-500G		500 g	$53.77	$12.90	$0.22	
Microscope Slides	solid	60	slides	Fisher	12-518-100A	Goldseal brand, Cat. No. 3010, 3" x 1", 1mm thick; can also buy as case of 25 x 144 for $639.07	144 slides	$31.95	$13.31	$0.22	
co-spot oligo	dry; make stocks of 0.1nmol/ul in 3X SSC	4.07	nmoles	Proligo / Sigma-Aldrich	N/A	HPLC-purified; 1.0umol starting synthesis scale results in, on average, 25.5nmoles yield; 5'- ACC TCG CTA ACC TCT GTA TTG CTT GCC GGA CGC GAG ACA AAC CTG AAC ATT GAG AGT CAC CCT CGI TGT T - 3'	25.5 nmoles	$107.00	$17.06	$0.28	$0.39 for 0.68ul per ink
oligos to array	use at 40pmol/ul	0.03	nmoles of EACH	Invitrogen / Illumina	N/A	ordered 50nmol starting synthesis scale, concentration-normalized to 40pmol/ul, made into aliquot plates of 10ul which were shipped dry, with the remainders shipped as liquid for aliquotting here. So 400pmoles per aliquot plate, x 10 plates (conservatively?) = 4000 pmoles = 4 nmoles	4 nmoles of EACH	$42,000.00	$296.47	$4.94	$6.72 for 0.68ul per ink
1-Methyl-2-pyrrolidinone	anhydrous, 99.5%	1852	mls	Sigma-Aldrich	M6762-1L		1000 mls	$17.85	$18.78	$0.31	
succinic anhydride	dry	17.29	g	Sigma-Aldrich	239690-50G		50 g	$25.60	$8.85	$0.15	
Boric acid	dry	10	g	e.g. Sigma-Aldrich	B6768-1KG		1000 g	$25.20	$0.25	$0.00	
EtOH	95%	1630	mls	e.g. stockroom		do not make from 100%!	3785 mls	$12.50	$5.38	$0.09	
PBS	1X	36.9	mls	e.g. stockroom			mls				

Amplification, labeling, and hybridization

Reagent	Conc.	Amount		Supplier	Item #	Notes	Unit of Sale	cost per unit sold	Total cost	Cost per array	Other
Halobacterium DNA	dry	600	ng	ATCC	700922D	shipped as dried genomic DNA, 10ug	10,000 ng	$199.00	$11.94	$0.20	
Primer A	40 pmol/ul OR, if 100	60	uls	e.g. Proligo	N/A	5' - GTT TCC CAG TCA CGA TCN NNN NNN NN - 3'	order as liquid, de	$12.54			
	pmol/ul	24	uls								
DTT	0.1M	45	uls	e.g. Promega	P1171	comes as 100ul, 100mM	100 uls	$12.00	$5.40	$0.09	
BSA	500 ng/ul	90	uls	e.g. NEB	890015	dilute to sold as 10mg/ml	25 mg	$10.00	negligible	negligible	
Sequenase	13 U/ul	36	uls	USB	70775Z	Sequenase 2.0, 1000U at 13U/ul	76.92 uls	$514.00	$240.56	$4.01	
MgCl2	50mM	480	uls	e.g. comes with Taq	N/A						
Primer B	100 pmol/ul	120	uls	e.g. Proligo	N/A	5' - GTT TCC CAG TCA TC - 3'	order as liquid, de	$9.90			
Taq	5 U/ul	120	uls	your favorite, e.g. the cheap stuff from th	#M829S, 500U, $119	I've liked what I got from Promega before but now they've replaced their cheap stuff with GoTaq Flexi, and it's pricier; e.g.	2500 U	$530.00	$25.44	$0.42	
NaHCO3	make 0.5M	40.33	uls	e.g. Sigma, or stockroom	S6297	make from dry, filter-sterilize	250 g	$12.10	negligible	negligible	
Cy3 dye	dry	660	ug	GE Healthcare / Amersha	PA13105	Cy3 NHS ester, 5mu	5000 ug	$834.00	$110.09	$1.83	
co-spot oligo complement, amine-modified	shipped dry	60	pmol	Proligo / Sigma-Aldrich	N/A	synthesis resulted in 75.1nmoles yield; 5' - AAC AAC GAC GG[ACGT] GAC TCT CAA [ACGT]GT TCA GGT TTG TC[ACGT] CGC GTC CGG CAA GCA A[ACGT]A CAG AGG TT[ACGT]A GCG AGG T - 3'	75.1 nmoles	$782.00	$0.62	$0.01	
Cy5 dye	dry	5.5	ug	GE Healthcare / Amersha	PA15101	Cy5 NHS ester	1 mg	$244.00	$1.34	$0.02	
aa-dUTP	50mM	14.1	uls	Applied Biosystems / Amb	AM8439	5-(3-aminoallyl)-dUTP, 50uls of a 50mM	50 uls	$126.00	$35.53	$0.59	
dNTP stocks	100mM	62.7	uls	e.g. Promega	U1420	set, 100mM each, 25umoles (=250uls) EACH	250 uls	$188.00	$47.15	$0.79	
SSC	20X	339.4	mls	Applied Biosystems / Amb	AM9770		500 mls	$31.00	$21.04	$0.35	
HEPES, pH 7.0	make up 1M, pH7	37.2	uls	e.g. Sigma	H4034-25G	sold dry, "HEPES >= 99.5%, biotech performance certified"	25 g	$14.75	negligible	negligible	
Ambion H20	pure	25.759	mls	Applied Biosystems / Amb	AM9937	10 x 50mls	500 mls	$77.00	$3.97	$0.07	
TE, pH 8.0	1X	20.82	mls	e.g. Applied Biosystems /	AM9849		500 mls	$31.00	$1.29	$0.02	
polyA	make 10mg/ml	73.2	uls	Sigma-Aldrich	P9403-25MG		25 mg	$47.80	negligible	negligible	
SDS	10%	51.7	uls	e.g. Applied Biosystems /	AM9770		500 mls	$31.00	$3.17	$0.05	

$908.70 $15.14 $17.03
winter cost   summer cost

$8.89   0.58718 $27.65 $45.43
portion in array costs   0.6085

portion in PCR costs

Appendix 2

A Primer on Microarray Design

Appendix: Primer on aspects of Microarray Design
With specific reference to techniques used in microbial ecology microarrays

From proposal defense paper, 2004.

I. Array Creation:
(A) Type of Probe: Most microarray studies characterize expression in specific organisms, and the majority of those arrays consequently use immobilized cDNAs as probes ("probes" are the DNAs spotted onto the array, while "targets" are the labeled complementary nucleotides in the sample being queried). Of the limited research employing microarrays to examine diversity (i.e., the presence/ absence/ relative abundance of probe sequences in a given environment) rather than expression, there are in general three types of probes that have been used: PCR products (Wu et al., 2001; Cho and Tiedje, 2002), short oligonucleotides (Loy et al., 2002; Bodrossey et al., 2003), and long oligonucleotides (Taroncher-Oldenburg et al., 2003).

Two issues to consider in choosing probe type are specificity and sensitivity. Short probes are generally more specific but less sensitive than long probes. This becomes intuitive in the context of hybridization kinetics; a long probe (several kb) will be less specific to a given target than a short probe because of increased cross-hybridization. However, signal intensity increases linearly with probe length; Wu et al. (2001) tested PCR products of varying sizes and found a linear increase in signal up to 1.4kb, using pure culture targets. The longer a probe is, up to a point, the more labeled target can hybridize to it, and the greater the signal intensity.

This relationship between probe length, specificity and signal has been described mathematically (e.g., Greene and Voordouw, 2003)

$$I(x) = k(x)*c(x)*f(x)$$

where the hybridization intensity $[I(x)]$ for a given spot equals the hybridization constant $[k(x)]$ of the probe sequence times the amount of probe DNA $[c(x)]$ spotted on the filter times the fractional amount (wt/wt) of the target sequence $[f(x)]$ within the community DNA. The hybridization constant is specific to a given sequence, as it is proportional to G-C content and length but also depends upon the precise sequence of bases.

To counter the confounding effects of cross-hybridization when dealing with complex natural communities, it is best to use relatively short probes. In addition, by choosing probes of uniform length and with approximately the same G-C content, we can choose a hybridization temperature roughly appropriate to the entire array; PCR product-based arrays can be more complicated in their interpretation because of their length and sequence heterogeneities.

Very short probes (in the range of 18-30mers) have their own limitations. They seem to have poorer hybridization properties than slightly longer probes (Hughes et al., 2001); this may be because they are too close to the surface of the array causing hybridization to be physically hindered. For this reason, some investigators have inserted spacers (Loy et al., 2002; Bodrossey et al., 2003), while others have increased the length of the oligonucleotide, to provide a spacer region which can also be involved in hybridization, thereby potentially increasing sensitivity as well (Hughes et al., 2001; Taroncher-Oldenburg et al., 2003).

The maximum length of oligonucleotide synthesis with high accuracy is 70nts. Due to

the increase in sensitivity and accessibility of these longer oligonucleotide probes compared to very short probes, and their better specificity when compared to longer PCR products, I will be using 70-mers for our array.

(B) Printing: There are several general options for creating the microarrays. Short oligonucleotide microarrays such as those made by Affymetrix can be made through a photolithographic process (like computer microchips), although recently they have also been synthesized in place using an ink-jet printer to arrange and control the chemistry (Hughes et al., 2001). PCR-product microarrays and those made with long-oligonucleotides are usually spotted with an "arrayer" robot. For the description of the design and construction of such an arrayer, please see Eisen and Brown (1999). The DeLong lab will have its own arrayer arriving this fall to the new lab at MIT.

The probes are suspended in a buffer during spotting, and the nature of this buffer can effect both the success of the print run (clogging of the robot's arraying pins, etc.) and the morphology of the resulting probe spots. While the majority of microarrays have been printed using 3X SSC as the printing buffer, several studies have shown that 50% DMSO provides better quality printing. Spotting short oligonucleotides attached to a spacer, Bodrossy et al. (2003) found that 50% DMSO provided lower standard deviation between replicate spots, and dried out more slowly during the spotting protocol, than 3X SSC. Spotting PCR products, Wu et al. (2001) found the same difference, with 50% DMSO providing better signal intensity and spot homogeneity, and lower evaporation during printing. Others have used betaine as the printing buffer to similarly decrease evaporation during spotting (A. Gracey, personal communication). The reason that slow drying is desired during a print-run is because using a fully-aqueous buffer dries quickly and unevenly, leading to poor spot morphology. For the prototype array I've been using 3X SSC, to match the lab whose arrayer we've used, but in the fall I may experiment with different printing buffers.

(C) Post-processing: In general, once a microarray is printed, depending on the type of probe and the type of slide used, it may need to be cross-linked, blocked, the spotted DNA may need to be denatured, and then the microarray must be dried. These steps are neither interesting nor particularly controversial, and so I will not go into any details. It is likely that I will follow the protocols at microarrays.org, although appropriate post-processing will depend on choice of printing buffer.

II: Target Preparation:
The target sequences, those complementary to the probe and present in the environmental mix being queried, can be prepared in a variety of ways. Considerations include whether any amplification step will occur, what type of fluorophore should be used for visualization, and the method of attaching the fluorophore to the target.

A key problem in existing microarray research on microbial communities is the high limit of detection. Several groups, using several different probe types and target DNA preparation methods, have found that to be detected by its probe a target must be present at ≥ 10pg of DNA – assuming a genome size of around 5Mbp, with a gene of around 1000bp, this means a species must represent ≥5% of the DNA in the community for its genes to be detected (e.g., Taroncher-Oldenberg et al., 2003; Bodrossey et al., 2003; Cho and Tiedje, 2002). Bodrossey et al. (2003) used a short oligonucleotide array, and amplified the gene of interest from their target DNA pool.

In contrast, Cho and Tiedje (2002) used longer (500-900bp) PCR products as probes but did not amplify their extracted community DNA. Taroncher-Oldenburg et al. (2003) found the same detection limit of 10pg of target DNA using long oligonucleotide probes and PCR amplifying target DNA. While this relatively high detection limit leaves microarrays useful for mapping the distribution of dominant species in a system, we know that numerically rare species can play important roles in community dynamics and biogeochemical cycling (e.g. nitrogen fixers). Cho and Tiedje (2002) propose several possible solutions to this issue of detection limitations: 1) increase the amount of probe immobilized on the array; 2) enrich the environmental samples for the genomes or genes of interest; and, 3) achieve higher sensitivity in signal detection. Solution (1) is dependent upon the probe spotting pins used during printing of the array, and will not be discussed here. Attempts at solutions (2) and (3) are discussed below.

(A) To amplify or not: In studies directed to specific functional groups, the sample is often enriched for the target sequences (solution 2 above). PCRs are commonly performed on the community DNA using primers specific to the gene(s) of interest (e.g., Bodrossey et al., 2003). While this increases the effective sensitivity, it can also skew the relative abundances of different sequences due to differential amplification through PCR (a well-documented limitation of PCR – see, for example, Suzuki & Giovannoni 1996). In addition, it limits the possible targets to those amplified by primers designed based on sequences already in the database.

Another, broader approach is to use random amplification of the target DNA. This can be a powerful way to increase the effective sensitivity by increasing the entire pool of target DNA without biasing to specific known sequences. However, during any primed amplification process there will be heterogeneity in both the binding efficiency of the primers and the polymerization efficiency, depending on the local structure and sequence involved. This will create an unpredictable distortion in the relative abundances or different possible amplicons (Suzuki & Giovannoni 1996). For this reason, several techniques for amplifying target in a uniform way have emerged, though it is not clear yet which technique is consistently most robust across studies. This fall I will be experimenting with amplification of small DNA amounts to assess which amplification method is best for our application.

Due to PCR's inherent potential bias and stochasticity, the ideal would therefore be to avoid PCR-based amplification of the target altogether. One possible solution is to collect more target DNA during the sampling process, obviating the need for amplification. While organismal or soil-based studies are limited in the quantity of DNA that is practical to collect, aquatic research can employ filtration to greatly increase DNA yields. Using tangential flow filtration the DeLong lab has previously collected sufficient water column DNA to create BAC libraries (Béjà et al., 2000) and in microarray experiments the same technique can be used. The typical tangential flow procedure concentrates 500L of glass fibre pre-filtered seawater into a final resuspension volume of 0.5mls (Béjà et al., 2000), which represents an 1000-fold concentration of the community DNA. Therefore because the marine microbial habitat is amenable to concentration of cells, a good strategy may be to avoid amplification entirely. However, one of the goals of microarray development is to allow sampling at small spatial and temporal scales, and from a practical standpoint ship time limitations will mean that samples will be collected during cruises that have other primary goals. For this reason, it will not be practical to concentrate large amounts of water during every sampling effort. Critical locations may be periodically sampled intensively by collecting large numbers of cells, for extraction and labeling

without amplification, to validate the chosen amplification technique, but this cannot be the standard collection method.

Some studies have successfully queried un-amplified extracted community nucleic acids, without collecting large volumes of sample. Small et al. (2001) used 1μg of extracted community RNA per slide and could reproducibly assign presence or absence of the targeted groups, although they were unable to reliably quantify their targets. Practically, without dramatic increases in sensitivity through detection abilities, some form of amplification is likely to remain necessary. To increase the sensitivity of the detection of hybridized target (solution 3 in the preceding discussion of detection limits), one can improve the visualization method and/or the fluorophore.

(B) Choice of fluorophore: The next consideration regarding the target DNA is which fluorophore should be used. The fluorophore is the fluorescent molecule that is attached to the target DNA, so that the target's hybridization to any of the probe spots on the array can be visualized. Issues surrounding fluorophore choice include the relative intensity of the fluorophore, its susceptibility to bleaching, and to quenching. The vast majority of microarray studies use the rhodamine-derivative Cy dyes, Cy3 and Cy5. However, several recent studies have suggested that Alexa dyes (Molecular Probes) may be better, increasing sensitivity by 2-3-fold (Appendix1 Fig 1; and, e.g., DeRisi, 2003). There are seven different Alexa dyes, one of which can already be bought in the esterified form (see next section for why this is required). The Alexa dyes are less effected by pH and more resistant to photobleaching compared to the Cy dyes (Fig. 1; and DeRisi, 2003). A caveat when attempting to reproduce results: researchers have found that different Cy dye batches can have quite different levels of sensitivity (Wu et al., 2001).

(C) Labeling the target with the fluorophore: Once the appropriate fluorophore has been chosen, the next step is the labeling of the target DNA. There are two types of protocols for labeling. The first Is "direct" labeling, where the fluorophore is conjugated directly to one of the nucleotides used in a replication or transcription step of the target preparation. The second approach is "indirect" labeling, which incorporates a non-labeled but modified nucleotide in the replication or transcription, which is then conjugated to the fluorophore after the polymerization reaction. Direct labeling is faster and simpler, however the incorporation efficiency of the labeled nucleotides is lower than for unlabelled nucleotides (DeRisi, 2003). Indirect labeling avoids the problems of differential incorporation of the Cy3- and Cy5-labeled dUTPs, gives a lower background fluorescence, and increases sensitivity (Dennis et al., 2003; DeRisi, 2003). In indirect labeling, amino-allyl dUTPs are used in the polymerization step. The products are then conjugated to an N-hydroxylsuccinamidyl ester form of the desired fluorophore. As opposed to dye-labeled dNTPs, the incorporation of amino-allyl dUTPs is approximately the same as that of unmodified dUTPs (DeRisi, 2003).

To prevent secondary structure from forming in the target sequence and interfering with its hybridization to probes, the labeled target is often fragmented. While many groups use labeled DNA as the target, RNA can be chemically fragmented in a random manner (Bodrossy et al., 2003). In addition, for those using a direct-labeling approach, the Incorporation of Cy-labeled nucleotides into RNA is more efficient than it is in DNA. For these reasons, in vitro transcription has been used during target preparation (e.g., Bodrossey et al., 2003). However, even chemical fragmentation of RNA may not always provide uniformly small pieces – longer transcripts may not

may not fragment thoroughly; Koizumi et al. (2002) had difficulty with some of their probes never showing signal even when they knew there was a perfect match transcript in the target mix, and they suggested that one factor responsible was incomplete fragmentation of the target.

III: Hybridization:

As with more traditional blot-based hybridization, hybridization to microarrays is affected by a range of factors. The stringency of hybridization is affected both by the conditions during the hybridization itself, and in the subsequent wash steps. The temperature used for hybridization cannot be empirically tailored to each individual probe when there are hundreds or thousands of probes on the same substrate, although when designing short probes of uniform length it is possible to select for probes with a close-to-uniform melting temperature (within a narrow range). Although some researchers do empirically hone their hybridization temperature (Wu et al., 2001), many use a low- or even room-temperature hybridization and rely on wash steps to achieve the desired specificity. Wash buffers typically include a detergent, such as SDS, and ionic components such as SSC (which is made from sodium chloride and sodium citrate), both of which stabilize the hydrogen bonding interactions of hybridization. By decreasing ionic or detergent concentrations in the wash buffer and by increasing the temperature or duration, the wash stringency can be increased. For this reason, several groups have invested significant effort empirically determining the most appropriate wash conditions for their microarrays to achieve the best trade-off between specificity and sensitivity (Wu et al., 2001; Koizumi et al., 2002; Taroncher-Oldenburg et al., 2003).

The most appropriate hybridization and washing conditions for a given array will depend not only on the specific probes and target involved, but also on the design of the array and on questions being asked. Has the array been created with some redundancy in probes, so that there is more than one probe for a given gene or organism of interest? If there is some redundancy, then it may be less crucial to prevent closely related target sequences from binding probes, since the degree of similarity between two different species' genes or genomes will vary with region – as was demonstrated graphically in the Wang et al. (2002) virochip viral "barcoding" results. In addition, the questions addressed may focus more on family- and genus-level changes in community composition, rather than changes in single species. Even within species, different strains can have different genes and different possible niches (REF); using multiple probes specific to a given species or strain, it is becoming possible to use microarrays to examine microheterogeneity of strains or closely related species (dubbed "genomotyping") and to pick out evidence for lateral gene transfer (Murray et al., 2001; Urakawa et al., 2003). Thus, the appropriate degree of stringency will depend entirely on the microarray involved and on the questions being asked.

If one were addressing questions that depended on precise, incontrovertible identification of perfect probe-match in the target mix, would it possible to achieve a sophisticated enough level of resolution to resolve perfect matches and single-base-pair mismatches? Current research indicated that it is not possible for every case. While the level of discrimination being created by the hybridization and wash conditions should be tailored to the questions being asked, it does not appear possible to ensure that all probes on an array, or even any given probe, only produces signal from a perfect match. However, it is possible to use hybridization and wash conditions to effectively remove double- and greater mismatches. With

optimization, one can also achieve exclusion of internal single-base-pair mismatches. However, for some probes terminal or penultimate single-base-pair mismatches hybridize as well or even slightly better than perfect matches (for example, see Urikawa et al., 2003; Taroncher-Oldenburg, 2003). Stahl's group regularly collects melting profiles for all the spots on their microarrays, in order to analyze dissociation temperatures of targets from probes. Recently they trained a neural network to inspect the signal data from their microarray at a given optimal discrimination temperature, determined by the melting profiles, and make the judgment of whether or not a given signal represented a perfect- or mis-match. The R2 for the ability of the neural network to correctly call a perfect match from a mismatch based on signal intensity was only 0.70 (Urakawa et al., 2003). It seems that this is not a limitation of the neural network analysis, but rather an oddity of the hybridization kinetics of certain mismatches, and therefore will likely not be surmountable by improved analytical tools. However, the good news is that by creating an array with redundancy, one can safeguard against misinterpretation based on a single faulty data point. For studies specifically of microheterogeneity, this internal-single-base-pair mismatch discrimination represents the current limit of discrimination.

Hybridization of microarrays is still poorly understood. In Loy et al.'s (2001) microarray studying sulfate-reducing bacteria they saw up to a dramatic 56-fold difference in the signal intensity of perfect matches among their 136 probes. They suggest that this difference may be due either to secondary structure in the labeled target DNA or to steric hindrance from hybrids formed on the array during the hybridization process. For a detailed discussion of the hybridization behavior of oligonucleotides in the context of microarrays, see Bodrossey at al (2003).

When using microarrays to track gene expression, investigators are looking for differences in signal among different stages or cell types, and so will competitively hybridize target from both or from a range or conditions in relation to one "standard" condition. With non-expression microarray studies, competitive hybridization has continued to be used, because absolute quantification through hybridization is not possible and so some form of relative quantification must be used. With longer PCR products as probes, Cho and Tiedje (2002) used lambda DNA in their microarray design. By spotting equal amounts of lambda DNA to probe in each spot on the array, they could then spike their target with lambda DNA, labeled with the other fluorophore. This provided an internal standard to quantify each spot's signal in relation to, and also allowed for normalization across the array for differences in spotting or hybridization efficiency, as well as representing a positive control. With longer PCR products as probes this is a smart approach, because the hybridization kinetics of the probe and the lambda DNA will be reason ably similar when averaged over their entire length, allowing the lambda DNA to act as a standard for whatever probe is being used. However, with oligonucleotide probes, the hybridization kinetics can be markedly different depending on the precise sequence. So unfortunately, lambda DNA would not be a meaningful internal standard on an oligonucleotide array (Bodrossey et al., 2003).

An interesting but very labor-intensive approach used by Bodrossey et al. (2003) in their oligonucleotide array was to use as a reference an artificial mixture of the PCR products represented by their short oligonucleotide probes. They would first do a one-color hybridization of the community DNA of interest to their array to get an idea of which sequences were present, and their rough relative abundances. Then they would then make up an appropriate reference mix of those sequences present using the appropriate PCR products, in a known ratio, and use that to competitively

hybridize against the target community DNA to refine their quantification of relative abundances. While this seems tractable in a smaller microarray (they had only 59 probes) one can imagine it becoming unwieldy quickly as the number of probes on the microarray increased. Bodrossey et al. (2003) acknowledged the limitations of this reference DNA approach, since it requires a reference set as similar as possible to the sequences present in the target. This limits the utility of this approach to, for example, studies looking at a single community over time or under different conditions.

However, for our BAC-derived microarray, there may be a way to competitively hybridize a reference set for less precise quantitative purposes. By amplifying, labeling, and fragmenting the BAC inserts used for the creation of the array, it should be possible to develop a standard reference mix to be used identically in all hybridization reactions. It has been previously suggested that the ideal reference for a complicated sample is equal portions of each component mixed together (Eisen and Brown, 1999). This would provide relative quantification for all samples targeted, in relation to this reference mix, and would us Indirectly compare multiple water samples taken at differing times. For comparisons with other labs in the long run, it would be ideal to develop a more precise means of standardization. (To be clear, hybridization depends not only on the target's absolute quantity but also on its relative abundance in the total target DNA, which might be quite different than that in the reference mix. This is why creating a reference mix with the same components present and in the same relative proportions as the target is the best way to actually get the most precise quantification – but this ideal reference mix will change over time, with the community, and so is not a practical solution for ecological studies.)

V: Data Analysis:
        * this section is just a brief introduction to a few of the considerations surrounding microarray data analysis, and will be expanded in the future.

Several studies have compared the results of microarray analyses to standard methods of assessing microbial diversity and relative abundances, such as PCR-DGGE (Koizumi et al.., 2002), sequencing of PCR product clone libraries (Loy et al., 2002), and Northern blots (Koizumi et al., 2002). In general, these studies have found a good agreement among techniques, with the caveat that microarrays have a comparatively high detection-limit (see start of Materials and Methods).

An important consideration is how to decide whether to count a probe's signal as "on" or not, or in our case "present" or not. Many different approaches to microarray data standardization have been explored (e.g., Dennis et al., 2003). Many researchers use an arbitrary cut-off for defining a signal as "on", for example in Loy et al. (2002) the cut-off for considering a given spot "positive" was if its signal-to-noise ratio, calculated using their unique formula, was greater than 2.0.

Once the number of spots exceeding the defined cut-off signal intensity has been determined, the next step is to interpret the remaining data. To date, several approaches exist to ordering the data to interpret meaning, look for patterns, and assess the significance of differences seen among treatments.
Hierarchical clustering is a common tool for looking at microarray data and can reveal informative patterns (Brown and Botstein, 1999). For example, in the results of an experiment using the BAC-based microarray proposed here, all the probes from a given BAC might cluster together, implying the presence of that genome or its close relatives in the sample. Alternatively, the probes to several homologues of a

given operon from several different BACs might cluster together – this could be interpreted in two ways: either the operon has a very high degree of conservation, to the exclusion of the rest of the host genomes, or there are novel genomes present which contain a highly conserved version of that operon. This example shows how important probe design can be – in highly conserved genes, it may be appropriate to include two different probes, one in the heart of the conserved region, and one in the most divergent region. While hierarchical clustering is an extremely useful tool, many have reservations about it because of the sheer volume of data involved in microarrays studies. Clustering can be unreliable when dealing with so much data because it is often impossible to achieve high bootstrap values (Tilstone, 2003; and K. Pollard, personal communication).

A true statistical analysis of microarrays is a difficulty problem, and a single solution has not been embraced in the community. An add-in to Excel has been developed called SAM, the Significance Analysis of Microarrays, which is superior to a t-test at the low replication numbers typical of microarray studies (Piper et al., 2002). Several groups are working on robust tools for statistically analyzing microarray data, including Duke University's CAMDA, the Critical Assessment of Microarray Data Analysis (www.camda.duke.edu).

Appendix Figure 1. Comparison of Alexa and Cy dyes for labeling targets in microarray hybridizations (figures taken from Molecular Probes' website)

A. They have similar absorption and emission spectra
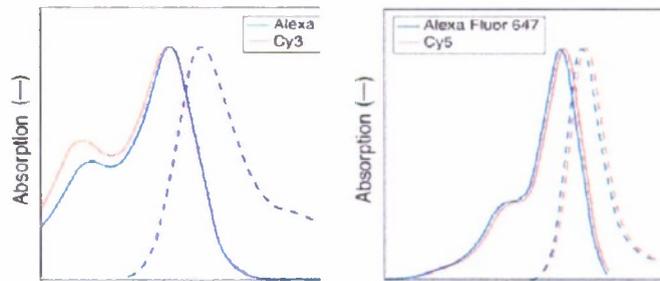
B. But look at the emission intensity
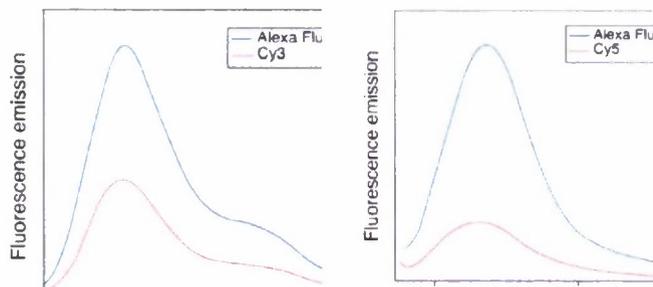
C. Alexas also have less bleaching over time, less sensitivity to pH change, and less quenching as the number of dye molecules per target increases
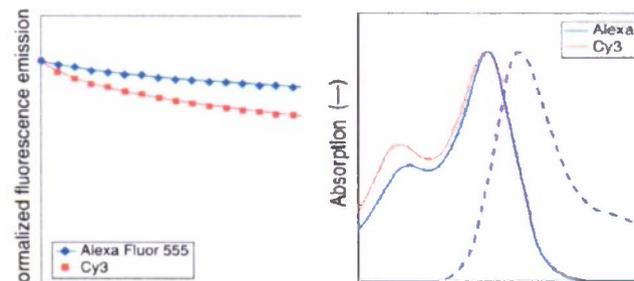
Appendix 3

Methane-Oxidizing Bacteria in a California Upland Grassland Soil: Diversity and Response to Simulated Global Change.

Authors: Hans-Peter Horz, Virginia Rich, Sharon Avrahami, and Brendan Bohannan.

Methane-Oxidizing Bacteria in a California Upland Grassland Soil: Diversity and Response to Simulated Global Change

Hans-Peter Horz,[1,2] Virginia Rich,[1]† Sharon Avrahami,[1] and Brendan J. M. Bohannan[1]*

Department of Biological Sciences, Stanford University, Stanford, California 94305,[1] and Division of Oral Microbiology and Immunology, RWTH Aachen University Hospital, 52074 Aachen, Germany[2]

We investigated the diversity of methane-oxidizing bacteria (i.e., methanotrophs) in an annual upland grassland in northern California, using comparative sequence analysis of the *pmoA* gene. In addition to identifying type II methanotrophs commonly found in soils, we discovered three novel *pmoA* lineages for which no cultivated members have been previously reported. These novel *pmoA* clades clustered together either with clone sequences related to "RA 14" or "WB5FH-A," which both represent clusters of environmentally retrieved sequences of putative atmospheric methane oxidizers. Conservation of amino acid residues and rates of nonsynonymous versus synonymous nucleotide substitution in these novel lineages suggests that the *pmoA* genes in these clades code for functionally active methane monooxygenases. The novel clades responded to simulated global changes differently than the type II methanotrophs. We observed that the relative abundance of type II methanotrophs declined in response to increased precipitation and increased atmospheric temperature, with a significant antagonistic interaction between these factors such that the effect of both together was less than that expected from their individual effects. Two of the novel clades were not observed to respond significantly to these environmental changes, while one of the novel clades had an opposite response, increasing in relative abundance in response to increased precipitation and atmospheric temperature, with a significant antagonistic interaction between these factors.

Methane-oxidizing bacteria (methanotrophs) are a unique group of aerobic, gram-negative bacteria that use methane as their sole source of energy. They are ubiquitous in nature and, as the major biological sink for the greenhouse gas methane, they are involved in the mitigation of global warming. Methanotrophs are also of special interest to environmental microbiologists because of their capability to degrade various environmental contaminants, their potential for single cell protein production, and other novel aspects of their biochemistry (19).

Based on physiological and biochemical characteristics, cultured members of the methanotrophs are traditionally divided into two main groups: type I methanotrophs, which are members of the class *Gammaproteobacteria* (e.g., *Methylomonas*, *Methylococcus*, *Methylomicrobium*, *Methylothermus*, *Methylohalobium*, *Methylocaldum*, and *Methylobacter*) and type II methanotrophs, which are in the class *Alphaproteobacteria* (e.g., *Methylosinus*, *Methylocella*, *Methylocapsa*, and *Methylocystis*) (14, 15, 19).

However, this picture of methanotrophic diversity has become much more complex recently. The genera *Methylocella* and *Methylocapsa*, although considered members of the type II methanotrophs, are phylogenetically distinct from the classical representatives of type II methanotrophs and differ physiologically in many aspects from all other known methanotrophs (13–16). In addition, methanotrophic isolates from some Arc-

tic soils have been shown to possess highly divergent *pmoA* genes; this gene encodes the active site polypeptide of particulate methane monooxygenase, a key enzyme in methane oxidation (39). The *pmoA* gene has been used as a molecular marker in numerous environmental studies of methanotroph diversity (18, 20, 28, 36) and is an ideal marker because it codes for an enzyme that is central to methane oxidation, is present in all known methanotrophs (with the exception of *Methylocella*), and there is no evidence of horizontal transfer of *pmoA* among methanotrophs (i.e., the *pmoA* phylogeny is generally consistent with the 16S rRNA-based phylogeny of methanotrophs) (12, 36). Unique *pmoA* gene sequences (for which no isolates are known) have also been identified in a number of culture-independent studies of environmental samples (4, 21, 25, 28, 32). Among the most interesting of these unique sequences are those suggested to belong to specialized methanotrophs adapted to the trace levels of methane found in the atmosphere (25, 32).

For example, in forest soils that are sinks for atmospheric methane, novel *pmoA* sequence types (the clade containing type sequence "RA 14") distantly related to *Methylocapsa acidiphila* have been described frequently, providing evidence for the existence of a distinct group of "specialized" methanotrophs (10, 21, 25). It has also been suggested that another group of methanotrophs represented by a novel *pmoA* lineage (the clade containing type sequence "WB5FH-A") that groups distantly to type I methanotrophs might be involved in atmospheric methane consumption in some soils as well (32). None of these putative atmospheric methane consumers has yet been isolated.

Consumption of atmospheric methane has the potential to play an important role in climate change. Methane is 20 to 25 times more effective per molecule than CO_2 as a greenhouse

* Corresponding author. Mailing address: Department of Biological Sciences, 371 Serra Mall, Stanford University, Stanford, CA 94305. Phone: (650) 723-3344. Fax: (650) 723-0589. E-mail: bohannan@stanford.edu.

† Present address: MIT/Woods Hole Oceanographic Institution Joint Program in Biological Oceanography, Massachusetts Institute of Technology, Cambridge, MA 02139.

gas (5, 44). Consumption of atmospheric methane is estimated to account for about 6% (about 30 Tg/year) of the global atmospheric methane sink (41). Furthermore, the environmental changes associated with greenhouse scenarios (e.g., increased temperature, precipitation, and nitrogen deposition) have the potential to interact with methane consumption and cause positive feedbacks between methane flux and climate change (31, 51). These interactions have been attributed to changes in the activity of methanotrophs and/or alterations in the structure of the methanotroph community in response to these environmental changes (31). However, it is unknown whether realistic global changes have the potential to alter the structure of the methanotroph community.

We investigated the response of soil methanotrophs to simulated multifactorial global change, including elevated atmospheric CO_2, higher atmospheric temperatures, increased precipitation, and increased nitrogen deposition, manipulated on the ecosystem level in a Californian annual grassland. The aim of our study was twofold. The first goal was to assess the methanotrophic diversity of the Californian annual grassland. This was accomplished by amplifying, cloning, and sequencing *pmoA* genes. Our second goal was to monitor shifts in methanotroph community composition in response to simulated global change. This was accomplished by creating genetic community profiles of methanotrophs from soils exposed to different combinations of simulated global changes. These profiles were based on terminal restriction fragment length polymorphism (T-RFLP) analyses of *pmoA* genes, an automated and sensitive approach that has been used for the characterization of methanotrophs in various environments (23, 27, 28, 40).

We observed that our grassland soil harbored a remarkable diversity of known and novel *pmoA* gene types and that the community structure of methanotrophs in this soil changed in response to simulated global change.

MATERIALS AND METHODS

Field experiment. The impact of individual and multiple, simultaneous global changes on methanotroph community composition was investigated using the Jasper Ridge Global Change Experiment (JRGCE). The JRGCE is located on the Jasper Ridge Biological Preserve, which lies in the eastern foothills of the Santa Cruz Mountains in northern California. The climate, vegetation, and soil parameters, as well as the experimental design, have been described in detail previously (43, 48). In brief, the JRGCE was established in a grassland ecosystem dominated by annual grasses (*Avena barbata* and *Bromus hordeaceus*) and forbs (*Geranium dissectum* and *Erodium botrys*), growing on a sandstone-derived soil with an average pH of 6.31 ± 0.3. Four global change factors, CO_2 (ambient and 680 ppm), temperature (ambient and ambient plus 80 W m^{-2} of thermal radiation), precipitation (ambient and 50% above ambient), and nitrogen deposition (ambient and ambient plus 7 g N m^{-2} in the form of calcium nitrate), were applied to different plots in a full factorial design (leading to a total of 16 different treatments). Each treatment was replicated eight times. The treatments were applied as a split-plot design with 32 circular plots, each divided into four 0.78 m^2 quadrants, separated by solid belowground and mesh aboveground partitions. Infrared heat lamps were suspended over the centers of the warming plots, heating the plants in all quadrants of a plot by 0.8 to 1°C. Atmospheric CO_2 concentrations were elevated with a ring of free-air emitters surrounding the plots. Ambient precipitation events were augmented with drip irrigation and overhead sprinklers; the precipitation treatment increased the average soil moisture from 19.8% to 26.6% (measured at the time of soil sampling). Warming and CO_2 treatments were applied on the whole-plot level, and precipitation and nitrogen treatments were applied on the subplot level. Manipulations started in the autumn of 1998, at the beginning of coastal California's rainy season.

Soil sampling. The analysis of microbial communities was initiated in May 2000. Soil cores from all replicate treatments were taken from a depth of 15 cm

with a 2.2-cm-diameter corer. Each core was placed in a plastic bag, cooled on ice in the field, and homogenized thoroughly by hand in the laboratory prior to storing at −80°C.

Extraction of total DNA. Extraction of DNA from 0.5 g of soil was performed using the Ultra soil DNA extraction kit (MoBio Laboratories, Solana Beach, CA) according to the manufacturer's instructions, with the exception that the final purification step was repeated to increase the purity of the DNA. The DNA was resuspended in a final volume of 50 μl and stored at −80°C. DNA quantification was performed with the PicoGreen assay (Molecular Probes, Eugene, OR) according to the manufacturer's directions. The DNA yield was approximately 5 to 20 ng/μl.

Primer evaluation. To characterize the methanotrophic diversity, we tested five different primer combinations for their suitability to amplify *pmoA* gene types in the Jasper Ridge grassland soils. For this preliminary test, we chose soil samples from two plots with elevated CO_2, temperature, precipitation, and nitrogen (plots 1D5 and 1D60). For each soil and primer combination, one clone library was generated, and we sequenced 15 clones per library. The primer combinations tested were (i) A189F-682R (24), (ii) A189F-650R (10), (iii) A189F-mb661R (12), (iv) A189F-682R (seminested: 650R), and (v) A189F-682R (seminested: mb661R). All clones sequenced from clone libraries generated by use of the A189F-682R primer system were *amoA* sequence types closely related to the ammonia-oxidizer *Nitrosospira multiformis*. Clone libraries that were generated based on the A189F-650R and A189F-mb661R primer systems contained some *pmoA* sequences. However, up to 50% of the randomly selected clones contained nonspecific inserts. In contrast, all clones sequenced from clone libraries generated using the two seminested PCR approaches, A189F-682R (seminested: 650R) and A189F-682R (seminested: mb661R), were *pmoA* sequence types. Therefore the seminested PCR approach was subsequently used for the study of methanotrophic diversity and for generating community profiles by T-RFLP analysis.

PCR amplification. As described above, the amplification of *pmoA* genes was performed via a seminested PCR approach using the 5′ primer A189 and the 3′ primer A682 (24). The temperature profile (Table 1) was identical to the previously described "touch-down" PCR protocol (28). Aliquots of the first round of PCR (0.25 μl) were used as the template in the second round of PCR using the 5′ primer A189 and the two 3′ primers mb661R and 650R in a multiplex PCR setting (i.e., both reverse primers were present in the same reaction). This approach allowed simultaneous amplification of a broad range of *pmoA* targets. The reverse primer mb661R was designed for the detection of type I and type II methanotrophs (12), while the reverse primer 650R was designed for the specific detection of putative atmospheric methane oxidizers from the "RA 14" clade (10). Each reaction mixture contained 12.5 μl of MasterAmp PCR premix F (Epicentre Technologies, Madison, WI), 0.5 μM of (each) primer (QIAGEN, Alameda, CA), 1.25 U of *Taq* DNA polymerase Low DNA (AmpliTaq, Applied Biosystems, Foster City, CA), and 0.25 μl of template DNA. Amplification was performed in a total volume of 25 μl in 0.2-ml reaction tubes, using a DNA Engine thermal cycler (MJ Research, San Francisco, CA). The PCR amplifications of environmental DNA resulted in amplicons of the expected size (approximately 500 bp). The first round reaction and the second round reaction were each performed in triplicate. Aliquots from the first round (three independent reactions in three different tubes) were pooled before going into the second round (which was itself done in triplicate). These final reactions were pooled prior to digestion. Aliquots of the amplicons (5 μl) were checked by electrophoresis on a 1% agarose gel.

Cloning and sequencing. PCR products were cloned using a TOPO TA cloning kit (Invitrogen Corp., San Diego, CA) following the protocol of the manufacturer. The preparation of plasmid DNA of randomly selected clones, PCR amplification of cloned inserts, and nonradioactive sequencing were carried out as described previously (28).

Phylogenetic analysis. The identities of the *pmoA* gene sequences were confirmed by searching the international sequence databases using the BLAST programs (http://www.ncbi.nlm.nih.gov/BLAST/). The currently available database of *pmoA* gene sequences was integrated within the ARB program package (33), and DNA sequences were analyzed and edited using the alignment tools implemented in ARB. We constructed phylogenetic trees using the maximum likelihood approach (with the default settings), the Fitch-Margoliash approach (using global rearrangement and randomized input order with three jumbles), and the neighbor-joining approach (with the Felsenstein correction) in ARB. The robustness of the tree topology was verified through calculating bootstrap values for the neighbor-joining tree and through comparison of the topology of the trees constructed using the different approaches.

Analysis of molecular evolution of the novel *pmoA* lineages. The molecular evolution of the novel *pmoA* lineages was investigated using the codeml execut-

TABLE 1. Primer description and thermal profiles for PCR

Primer pair	Sequence (5'–3')	No. of PCR round	Thermal profile[a]	Molecular analysis
A189[b] A682[b]	GGNGACTGGGACTTCTGG GAASGCNGAGAAGAASGC	1	94°C, 45s; 62–52°C, 60s; 72°C, 180s (30 cycles)[c]	T-RFLP
A189[d] mb661R, A650R[e]	GGNGACTGGGACTTCTGG[b] CCGGMGCAACGTCYTTACC, ACGTCCTTACCGAAGGT	2	94°C, 45s; 56°C, 60s; 72°C, 60s (22 cycles)	
M13F M13R	GTAAAACGACGGCCAG CAGGAAACAGCTATGAC		94°C, 45s; 55°C, 60s; 72°C, 60s; (25 cycles)	Sequencing

[a] All PCR profiles began with an initial denaturation at 94°C for 3 min and ended with a final elongation step at 72°C for 10 min, prior to holding temperature at 4°C.
[b] Reference 24; A189 is the forward primer, A682 the reverse.
[c] Touch-down PCR was used from 62 to 52°C. After each cycle, the annealing temperature was decreased by 0.5°C until it reached 52°C (28).
[d] Primer labeled with 5-carboxyfluorescein.
[e] For mb661R, see reference 12; for A650R, see reference 10.

able of the Phylogenetic Analysis by Maximum Likelihood (PAML) program (58). The input nucleotide files contained a 453-nucleotide portion of all *pmoA* sequences shown in Fig. 1 (with the exception of sequence "E5FB-b" [AJ579668] from the "WB5FH-A" clade, which was too short to include, and "LOPA 12.6" [AF358043], "1Y-6.48" [AY236518], and "RA 14" [AF148521], which were added to Fig. 1 during final revisions of the manuscript; in addition, "VIp9" [AY37258] was removed from Fig. 1 during final revisions, but was present in the PAML analysis). To reduce the level of sequence divergence to within recommended levels (Z. Yang, personal communication), the PAML analyses were run on the two halves of the *pmoA* phylogenetic tree separately. One half contained the type I "WB5FH-A," JR2, and JR3 *pmoA* clades along with the two *Nitrosococcus amoA* sequences. The other half contained the type II "RA 14" and JR1 clades as well as *Methylocapsa acidiphila*. Due to the divergence of the two *amoA* sequences from the *pmoA* sequences in members of the class *Gammaproteobacteria*, the type I side of the tree was still at the limits of acceptable divergence for the PAML program, and so analyses for that side of the tree were also run without the *amoA* sequences. All sequences were in frame and aligned (using MacClade 4.03 PPC; Sinauer Associates, Inc., Sunderland, MA), and the few ambiguous sites were assigned the nucleotide of their nearest phylogenetic neighbors. The input tree files were created using PAUP 4.0b10 (49), using analysis by distance, neighbor joining with Jukes-Cantor correction, and ties broken randomly. Their topology matched that of the tree (Fig. 1) presented in this paper.

Branch lengths were estimated by the PAML program using the one-ratio model, and then those branch lengths were used as the initial values for branch length estimation in further models performed. In the codeml control file, the majority of parameters were left in their default specifications, with the following exceptions: runmode = 0, seqtype = 1, CodonFreq = 2, Model = 0 or 2, and, for the multiratio models, fix_blength = 1.

The one-ratio model was run to provide an estimation of a single nonsynonymous-to-synonymous substitution ("dN/dS") ratio for each half of the tree. A series of two-ratio models were then run, to allow the dN/dS ratio of the three novel lineages to vary in turn. Lastly, the dN/dS of each major branch and clade (as denoted in Fig. 2) was allowed to vary simultaneously under the freely varying model, generating maximum likelihood estimates for all dN/dS values across the tree (57).

To test the robustness of the parameter estimates, all analyses were also run on various subsets of the taxa, with little variation in the results; this is consistent with other studies that have shown that codeml is robust to sampling (57, 59). In addition, all analyses were run at least twice to ensure that parameter estimates were likely global rather than local optima.

The likelihood ratio test was used to assess the goodness of fit of the two-ratio models to the data and to compare it with that of the one-ratio model. This allowed us to test whether the dN/dS ratios on the branches leading to the three novel clades were significantly different from the background dN/dS ratio in the remainder of the tree (57).

T-RFLP analysis. The creation of terminal restriction fragments (T-RFs) from *pmoA* genes was carried out as previously described (28). After purification with QIAquick spin columns (QIAGEN, Alameda, CA), approximately 100 ng of the amplicons was digested separately with 20 U of the restriction endonuclease MspI (New England BioLabs, Beverly, MA). The digestions were carried out in a total volume of 10 μl for 3 h at 37°C according to the instructions of the

manufacturer. Enzyme inactivation was carried out by incubation at 65°C for 20 min. The subsequent T-RFLP analysis was performed at the Genomics Technology Support Facility (http://genomics.msu.edu/; Michigan State University, East Lansing, Michigan). Briefly, the T-RFs were separated by capillary electrophoresis on an ABI Prism 3700 DNA analyzer. The DNA bands were automatically identified and sized using GeneScan software (Applied Biosystems, Foster City, CA) and comparison to internal lane standards. The relative abundances of individual T-RFs in a given *pmoA* PCR product were calculated based on the peak height of the individual T-RFs in relation to the total peak height of all T-RFs detected in the respective T-RFLP community fingerprint pattern. The peak heights were automatically quantified by the GeneScan software. To verify the assignments of T-RFs to our detected *pmoA* gene types, we also tested individual clones by T-RFLP analysis.

The T-RFLP results were highly reproducible. The coefficient of variation of the relative signal intensity of the T-RFs between different DNA isolations from the same soil sample ranged from 3 to 10.1%. The coefficient of variation of the relative signal intensity of the T-RFs between different PCRs from a single DNA sample ranged from 1 to 6.5%. Those variations are in the same range as those previously reported (28). The variations between different digests from the same PCR product and between different electrophoretic runs from the same digest were negligible. This is consistent with previous systematic evaluations of the T-RFLP method (38).

Statistical analysis. The relative abundance data were analyzed with a split-plot analysis of variance performed using the MIXED procedure in SAS (SAS Institute, Inc., Cary, NC). Means were estimated as least-square means, and the degrees of freedom were estimated using the Satterthwaite approximation. The data were arcsine-transformed before analysis.

Nucleotide sequence accession numbers. The partial *pmoA* gene sequences determined in this study have been deposited in the EMBL, GenBank, and DDBJ nucleotide sequence databases under the accession numbers AY654669 through AY654732.

RESULTS

Characterization of *pmoA* genes. Clone libraries were constructed using *pmoA* PCR products from three experimental plots: two with elevated CO_2, temperature, precipitation, and nitrogen (plots ID 5 and ID 60) and one with ambient levels of CO_2, temperature, precipitation, and nitrogen (plot ID 107). In total, 64 clones were analyzed (11 clones for ID 5, 21 clones for ID 60, and 32 clones for ID 107). Figure 1 shows the phylogenetic affiliation of all clone sequences analyzed in this study.

Five sequences formed a distinct clade (JR1) that was related to the "RA 14" clade, environmental sequence types that have been hypothesized to represent uncultured "high-affinity" methanotrophs capable of oxidizing methane at atmospheric concentrations (21, 25). The similarity in DNA sequence between JR1 and the "RA 14" clade was approximately 80%.

0.1
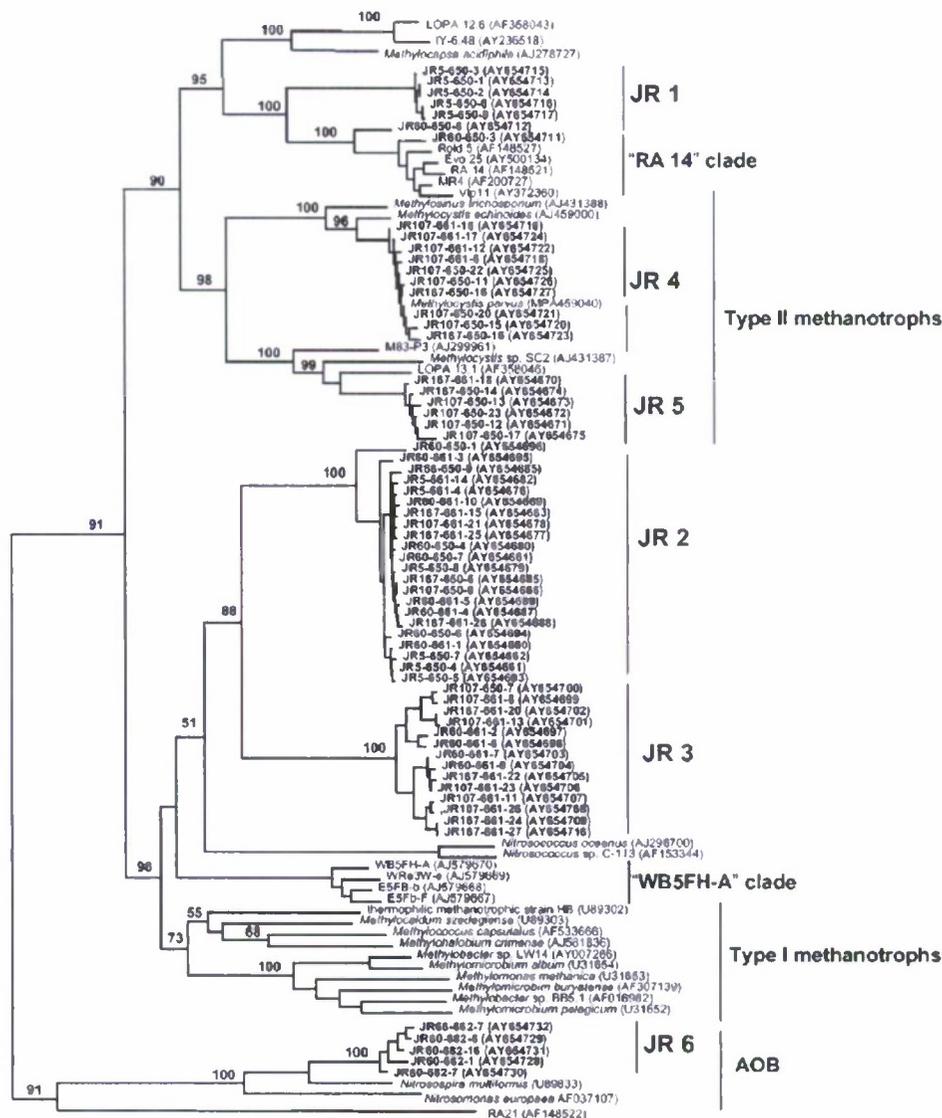
FIG. 1. Phylogenetic relationships among *pmoA* gene types identified in the Jasper Ridge Global Change Experiment and *pmoA* and *amoA* gene types available in public databases (2, 3, 6–9, 17, 21, 25, 28–30, 32, 35, 37, 45, 53). Sequences obtained in this study are shown in boldface type with the prefix "JR" and are designated clades JR1 to JR6. The environmental *pmoA* sequences used for reference were retrieved from various habitats, as follows: forest soils (AF148527, AF148521 [25], AF148522 [25], AF200727 [21], AY500134, AY372360 [29]), rice fields (AJ299961 [28]), peat soil (AF358043, AF358046 [35]), and upland grassland soils (AJ579670, AJ579669, AJ579668, AJ579667 [32]). The scale bar corresponds to 0.1 substitutions per nucleotide. The tree was calculated using 475 nucleotide positions and the neighbor-joining approach (with the Felsenstein correction), via the ARB program package (33). The tree topology was confirmed using the maximum likelihood approach. Bootstrap values were calculated using 1,000 replications. AOB, ammonia-oxidizing bacteria.
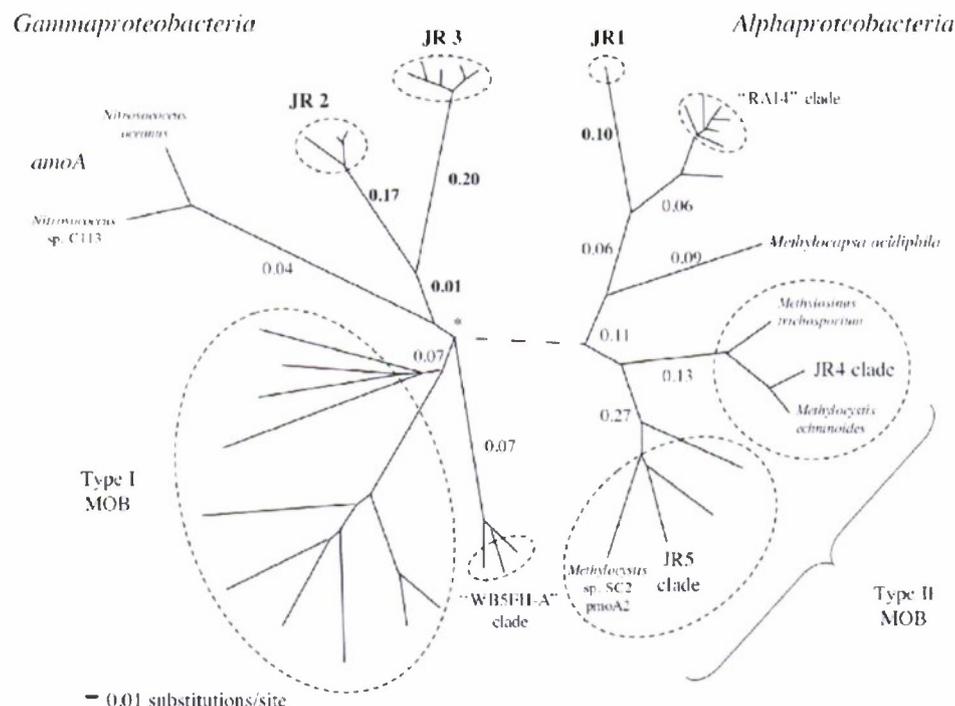
FIG. 2. The dN/dS values of the major lineages of *pmoA* as estimated using the codeml executable of the PAML program. The numbers at each branch are the dN/dS ratios estimated by the program under the freely varying model, which allowed the dN/dS of each major branch and clade (as denoted with dashed circles) to vary simultaneously. (The asterisk at the branch connecting the type I and "WB5FH-A" clades to the other *Alphaproteobacteria* clades indicates a noncomputable dN/dS ratio, where dN = 0.03 and dS = 0). Novel clades are shown in boldface type, as are the dN/dS ratios of the branches leading to these clades. The dN/dS ratios within each clade are not shown. Analyses were run on each half of the *pmoA* tree independently; the split between the *Alphaproteobacteria* and *Gammaproteobacteria* sections is indicated by a dashed line. MOB, methane-oxidizing bacteria.

Two sequences ("JR60-650-3" and "JR60-650-8") grouped tightly with the "RA 14" clade.

Two further clades (JR2 with 22 sequences and JR3 with 14 sequences) were moderately related to each other but showed no close relationship to any cultivated methanotroph species. They grouped distantly to type I methanotrophs, with a DNA sequence similarity of approximately 72%. While representing distinct lineages, JR2 and JR3 branched together with the "WB5FH-A" clade, the second novel *pmoA* lineage suggested to represent atmospheric methane oxidizers (32). Although supported by a bootstrap value below 50%, the common branching point of JR2, JR3, and the "WB5FH-A" clade was supported by a tree calculated using the maximum-likelihood approach. In contrast, a neighbor-joining tree calculated from deduced amino acid sequences favored a common branching point of the "WB5FH-A" clade with type I methanotrophs; however, the bootstrap value was again below 50%. Phylogenetic analysis consistently suggests a common evolutionary origin for the sequence clusters JR2 and JR3 and the *amoA* gene of *Nitrosococcus oceani* (an ammonia-oxidizing bacterium that is capable of using methane as a carbon source and whose prevalence is thought to be restricted to aquatic systems).

Eleven *pmoA* sequence types (clade JR4) were closely re-lated to *Methylocystis parvus*, a relatively well-characterized member of the type II methanotrophs. We also found six sequences (clade JR5) that grouped together with a novel *pmoA* lineage previously described as a diverged second *pmoA* gene copy present in various strains of type II methanotrophs (50). JR5 and the novel *pmoA* copy of one of the representative species (*Methylocystis* sp. strain SC2) had a DNA sequence similarity of 85%.

In summary, we identified *pmoA* gene types belonging to five different lineages within the phylogenetic radiation of the *pmoA/amoA* family. Three of these clades (clades JR1, JR2, and JR3) have DNA sequence similarities of 80% or less with previously described *pmoA* variants.

Nonsynonymous/synonymous substitution rates. The ratio of nonsynonymous to synonymous nucleotide substitution rates (dN/dS) was determined for each novel clade. The overall dN/dS ratio (as calculated with the one-ratio model of codeml) was 0.11 for the type I side of the *pmoA* tree and 0.10 for the type II side (data not shown). The dN/dS ratios (as calculated with the freely varying model of codeml) along the branches leading to the three novel lineages (JR1, JR2, and JR3) were 0.10, 0.17, and 0.20, respectively (Fig. 2). The likelihood ratio test showed that these dN/dS ratios were not

significantly different ($P < 0.05$) from the background dN/dS ratios in their respective sides of the tree (data not shown).

These dN/dS ratios could result from two possible scenarios. Clades JR1, JR2, and JR3 could have diverged recently, with insufficient time for their dN/dS ratios to reflect any change in functional state, or the three clades could have diverged earlier, with the low dN/dS ratios reflecting continued purifying selection. Using codeml, we estimated the relative divergence of the PmoA clades by estimating the likely numbers of synonymous changes along each of the three lineages, based on the assumption that synonymous changes are not acted upon by selection and accumulate steadily over time (55) (divergence times can also be roughly estimated by inspection of branch lengths in Fig. 1; however, these lengths reflect both synonymous and nonsynonymous changes). There were approximately 65 synonymous changes along the branch to JR1, 58 to JR2, 54 to JR3, and 109 on the branch leading to JR2 and 3. This is not substantially less than, for example, the 47 synonymous changes leading to *Methylocapsa acidiphila*, the 119 leading to the "WB5FH-A" clade, and the 55 leading to Type I methanotrophs (data not shown). Thus, the clades JR1, JR2, and JR3 do not appear to have diverged recently compared to other known *pmoA* clades, and so their estimated dN/dS ratios suggest that they are undergoing purifying selection, encoding functionally active proteins.

Conservation of amino acid residues. The pMMO and AMO genes are evolutionarily related (24), and at the amino acid level they share a number of highly conserved residues (25, 42, 52). Based on alignments of the predicted peptide sequences of the α subunits of 112 particulate methane monooxygenases (PmoAs) and 349 ammonia monooxygenases (AmoAs), Tukhvatullin et al. (52) identified residues common to both proteins. Ricke et al. (42) extended this analysis to include the second PmoA gene copy, PmoA2, present in many Type II methanotrophs (50). The inferred translation of the region amplified by the primers used in our study spans 16 of these highly conserved residues (Table 2). All members of novel clades JR2 and JR3 each had all 16 of these conserved residues. All members of JR1, JR4, and JR5 had 15 of the 16 conserved residues, with all but one member in each group also having the 16th residue. Among the residues common to both PmoA and AmoA, Tukhvatullin et al. (52) proposed a subset of seven that could potentially be the metal ligands of the active site. The translation of our amplified *pmoA* region spans three of these (residues E100, Y157, and H169), which are conserved in all Jasper Ridge sequences. A further set of four residues were identified as potential non-active-site metal ligands, which could additionally stabilize the peptide structure (52); our amplified region spans two of these (residues D182 and Y196), which are also conserved in all Jasper Ridge sequences.

In addition, Holmes et al. identified 21 residues that could distinguish PmoA from most AmoA sequences (25). Our *pmoA* amplicons spanned 16 of the putative PmoA/AmoA diagnostic residues. All of the clades we detected (with the exception of JR6) shared a high percentage of amino acid residues typical of PmoA (Table 2). JR4 had all 16 of the PmoA-specific residues, while JR5 and JR1 had 13 and 11, respectively, of the PmoA residues, and in all cases the mismatches were amino acids belonging to the same amino acid

similarity group (22) as the conserved PmoA residue (Table 2). Both JR2 and JR3 shared 14 of the 16 PmoA residues. The two mismatches in JR2 and one in JR3 were in the same amino acid similarity groups as the conserved residues, while the other mismatch in JR3 was a perfect match in half of the sequences in this clade.

T-RFLP community profiles. Figure 3 shows a representative community T-RFLP-profile and the assignment of the T-RFs to the sequence clusters detected in our study. All clones produced the T-RFs that were predicted based on sequence information (data not shown). All *pmoA* clades determined by comparative sequence analysis could be consistently recovered by T-RFLP community analysis. JR2, JR3, and JR5 exhibited specific T-RFs (208 bp, 373 bp, and 349 bp, respectively), confirmed by in silico analysis of the publicly available *pmoA* gene sequences (combined with the sequences generated in this study). JR4 produced a T-RF of 245 bp as anticipated (this is the specific T-RF for the type II methanotrophs) (28). However, no specific T-RF could be generated for clade JR1 by use of MspI (i.e., the 80-bp T-RF generated by JR1 can also be produced by digestion of *pmoA* sequences from *Methylococcus capsulatus* and related species, as well as *M. acidiphila*). A T-RF of 34 bp was indicative for sequences belonging to the "RA 14" clade.

Although not confirmed by cloned sequences, our T-RFLP community profiles indicated the presence of various members of type I methanotrophs (e.g., T-RFs of 440 bp, 505 bp, and 511 bp, with the latter two representing undigested *pmoA* sequence types without the MspI recognition site) (28), although in low abundance (generally less than 4% of the total). We can think of at least two possible explanations for the absence of *pmoA* sequences related to type I methanotrophs in our clone libraries, namely: (i) low relative abundance of the type I methanotrophs combined with nonexhaustive clone sampling, and (ii) cloning biases against type I sequence types. Recently reported discrepancies between the community composition of *pmoA* clone libraries and *pmoA*-based T-RFLP analysis (28, 40) suggests that such biases can be present. Given this possibility, we did not attempt to determine the response of type I methanotrophs to simulated global change in our study.

The response of methanotrophs to simulated global change. We generated T-RFLP community profiles of methanotrophs from all replicate treatments of our multifactorial climate change experiment (8 replicates of 16 treatments, for a total of 128 soil samples). Simulated global change did not significantly alter the number of T-RFs present (the phylogenetic richness of the methanotroph community) or the magnitude of Shannon, Simpson, or Berger-Parker diversity indices (34) calculated from the T-RFLP data. However, the simulated global changes did alter community composition. The relative abundance of type II methanotrophs (clade JR4) significantly decreased under elevated precipitation ($F_{1,24} = 7.89$; $P = 0.0068$) (Fig. 4) and elevated temperature ($F_{1,24} = 4.12$; $P = 0.0469$) (Fig. 4). However, these effects were not additive; i.e., there was a significant antagonistic interaction between precipitation and temperature ($F_{1,24} = 8.31$; $P = 0.0055$) (Fig. 4) such that the effect of both treatments together was less than that expected from their individual effects. In contrast, the relative abundance of the novel methanotroph clade JR2 responded to simulated global change very differently (Fig. 5). Elevated pre-

241

TABLE 2. Presence of conserved and diagnostic AA residues in PmoA and AmoA across taxa[a]

[a] The amino acids (AA) are numbered according to the published sequence for *M. capsulatus* PmoA (47). Uppercase letters are residues conserved in >95% of the reference data set; lowercase letters are residues conserved in >80% of the reference set. Letters in parentheses indicate conservation within AA similarity groups (A, PAGST; D, QNEDBZ; H, HKR; I, LIVM, F, FYW). Ties are indicated by both letters with a slash between them. Residues 100, 157, 168, 182, and 196 (with ^ below) are putative metal-binding residues as described by Tukhvatullin et al. (52). Residue columns containing gray backgrounds are AmoA/PmoA diagnostic sites described by Holmes at al. (25). Residues on a black background are generally agreed-upon AmoA/PmoA conserved sites (25, 42, 52). Residues in bold type and framed are AmoA diagnostic sites for ammonia oxidizers from the *Gammaproteobacteria*.

[b] Type II PmoA (*n* = 6), including *M. parvus*, *Methylocystis echinoides*, *Methylocystis trichosporium*, uncultured bacterium AF358046 (35), *Methylocystis* sp. strain SC2 (17), and uncultured bacterium M84 P3 (AJ299961) (28).

[c] *Nitrosococcus* clade AmoA (*n* = 8), including *N. oceani* (U96611) (37), *N. oceani* strain AFC27 (AF509001) (53), strain SW (AF509003) (53), strain AFC (AF508999) (53), strain AFC12 (AF508996) (53), strain AFC36 (AF508995) (53), *Nitrosococcus* sp. strain C113 (AF153344) (2), and uncultured bacterium BAC6 (AF070987) (45).

cipitation and temperature increased the relative abundance of this clade, and there was a significant antagonistic interaction between elevated precipitation and temperature ($F_{1,24}$ = 13.48; P = 0.0012) (Fig. 5) as well.

DISCUSSION

PmoA-based approach for methanotroph community analysis. The aim of the present study was to explore the methanotrophic diversity of a Californian upland grassland and to assess whether a shift in the methanotrophic community structure in response to simulated global change was detectable. We assessed methanotroph diversity in this study using a cultivation-independent approach, with *pmoA* as a molecular marker. To date, most studies involving *pmoA*-based analysis of methanotrophic populations have used the primer system A189F-682R. These primers also amplify *amoA*, which encodes the homologous subunit of the ammonia monooxygenase in nitrifying bacteria. Reverse primers that discriminate against the *amoA* (e.g., mb661) and highly specific primers with intended target specificity for the "RA 14" clade (e.g., 650R) have been applied as alternative methods for studying methanotroph diversity. Bourne et al. (10) tested these three primer sets in various soils and found that one primer combination alone was not sufficient to explore methanotrophic diversity.

We tested five different primer combinations (three single-round PCR assays and two nested PCR assays) in order to determine their potential to detect a broad range of methanotrophs in our grassland soil. When primer set A189F-682R was used, clone libraries created from the single-round PCR amplicons showed a high representation of *amoA* inserts. When primer set A189F-mb661R or A189F-650R was used, the clone libraries contained a large number of nonspecific inserts. Nested PCR, however, using the A189F-682R primer set in the first round and either reverse primer mb661 or reverse primer 650R in the second round, generated consistently high yields of *pmoA* amplicons, even in some soils for which single-round PCRs produced little or no *pmoA* amplification. In fact, all analyzed clones derived from nested PCRs were "*pmoA* positive." The reverse primers we used detected different components of the methanotroph community. Primer 650R detected the clades JR1 and sequences from the "RA 14" clade, clade JR3 could be detected only with reverse primer mb661, and clades JR2, JR4, and JR5 were detectable with both reverse primers. Therefore, we used both reverse primers together in a multiplex (i.e., in the same reaction), nested PCR approach for the T-RFLP community analysis. This enabled us to simultaneously recover a broad range of distinct *pmoA* clades in single electrophoretic profiles for each sample.
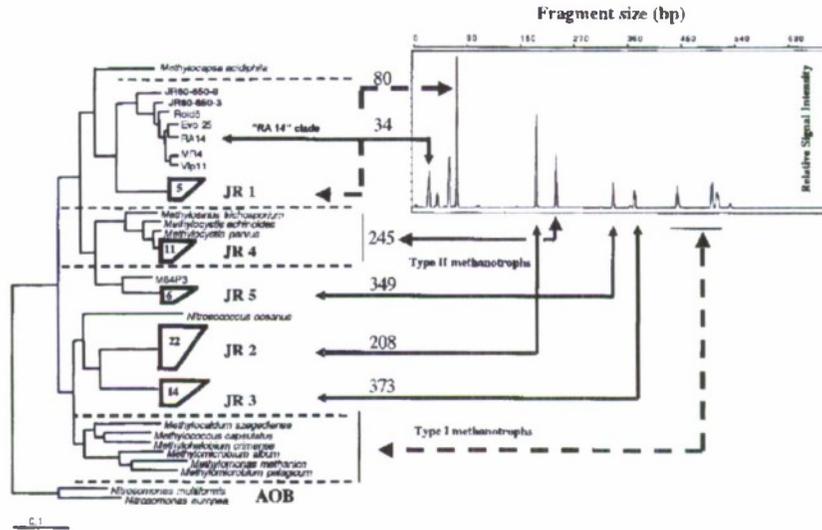
FIG. 3. Representative T-RFLP profile of the methanotroph community and the assignment (arrows) of the T-RFs to known methanotrophs-sublineages and to *pmoA* gene types determined in this study. The phylogenetic tree was graphically modified from Fig. 1. Arrows with dashed lines indicate the existence of multiple sequence types that potentially can produce the respective T-RFs according to the sequence information of the *pmoA* database (i.e., T-RFs of 80 bp, 440 bp, 503 bp, and 511 bp). AOB, ammonia-oxidizing bacteria.

Methanotrophic diversity. We discovered a remarkably high diversity of *pmoA* gene types in our study (Fig. 1), including those closely related to the *pmoA* of known members of the class *Alphaproteobacteria* as well as gene types distinct from known species forming hitherto undescribed *pmoA* lineages. Within type II methanotrophs, we found sequences closely related to *M. parvus* (clade JR4), as well as the recently characterized type II *pmoA* gene copy (50) of *Methylocystis* sp. (clade JR5). Interestingly, the relative abundance of the T-RFs

was consistently higher for JR4 than for JR5 in our T-RFLP profiles (data not shown), which agrees with the findings of Tchawa Yimga et al. (50) that not all type II methanotrophs possess this additional gene copy. We also discovered the clade JR1, which forms a distinct subgroup of the "RA 14" clade, the clade that has been putatively identified as atmospheric methane consumers (21, 25). This finding considerably expands the known depth of the "RA 14" clade and demonstrates that methanotrophs possessing this gene type are not restricted to forest soils. We did not detect the other putative atmospheric methane consumers, the "WB5FH-A" clade (32),
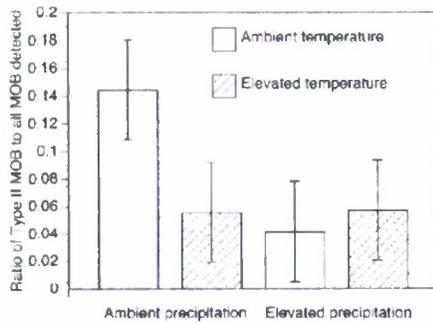
FIG. 4. Effect of temperature and precipitation on *pmoA* clade JR4 (type II methanotrophs) in the JRGCE. The mean relative abundance of JR4 is depicted for all samples, grouped by temperature and precipitation treatments. For example, the first bar depicts the mean relative abundance of JR4 from all experimental plots under ambient temperature and precipitation, including those under both ambient and elevated CO_2 and ambient and elevated nitrogen treatments ($n = 32$). Error bars are 95% confidence limits. MOB, methane-oxidizing bacteria.
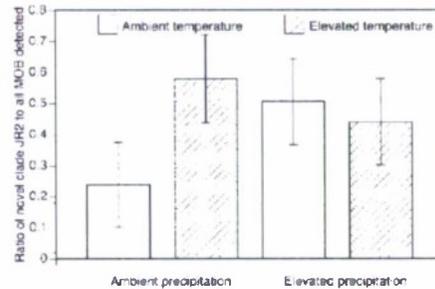
FIG. 5. Effect of temperature and precipitation on novel *pmoA* clade JR2 in the JRGCE. The mean relative abundance of JR2 is depicted for all samples, grouped by temperature and precipitation treatments. For example, the first bar depicts the mean relative abundance of JR2 from all experimental plots under ambient temperature and precipitation, including those under both ambient and elevated CO_2 and ambient and elevated nitrogen treatments ($n = 32$). Error bars are 95% confidence limits. MOB, methane-oxidizing bacteria.

although we did discover two novel clades (JR2 and JR3) which are distantly related to the "WB5FH-A" clade.

There are several lines of evidence that suggest that the three novel *pmoA* clades we discovered (JR1, JR2, and JR3) encode functional monooxygenases, with a primary substrate of methane rather than ammonia. All three of the novel clades had dN/dS ratios well below 1 (Fig. 2), evidence for purifying selection (55, 56). If these genes were nonfunctional copies, a lack of selection would result in nonsynonymous changes occurring at the same rate as synonymous changes, pushing the overall dN/dS ratio towards 1; how closely it approached 1 would depend on the divergence time of these clades. The dN/dS of the branches leading to all three novel clades, however, are not statistically different from the "background" dN/dS in the rest of each respective half of the *pmoA* phylogeny. In addition, the high number of synonymous changes along the branches leading to the three novel clades suggests that they did not diverge recently (see Results above), and thus their low dN/dS ratios suggest that their encoded proteins are expressed and functional.

The conservation of functionally diagnostic amino acid residues provides further evidence for retained function in the novel clades and for their substrate specificity for methane rather than ammonia. The novel sequences contain a very high percentage of those amino acid residues conserved in both methane and ammonia monooxygenases (42, 52). These conserved residues include those proposed to bind metal ions within the active site and at secondary stabilization sites (42, 52), as well as a majority of the previously identified PmoA-specific residues (25). Among the mismatched residues, almost all are in the same amino acid similarity groups as the PmoA-specific residues. The novel *Alphaproteobacteria* clade JR1 has the lowest number of perfect matches to putatively PmoA-specific residues (11 of 16) (Table 2) and has several putatively AmoA-diagnostic residues. However, two of these "AmoA-like" residues are, in fact, shared by several other PmoA clades. Furthermore, JR1 robustly clusters in the *Alphaproteobacteria*, within which there are no known *amoA*-containing members. Thus, the total evidence suggests that JR1 likely binds methane rather than ammonia. The novel *Gammaproteobacteria* clades JR2 and JR3 did not contain any AmoA-diagnostic residues. However, this picture is complicated somewhat by the fact that the only known ammonia-oxidizing bacteria within the class *Gammaproteobacteria*, the *N. oceani*-like clade, also lack many of the AmoA-diagnostic residues, and they are the closest phylogenetic relatives of JR2 and JR3 (Fig. 1). However, based on protein and inferred-translation alignments, there appear to be six sites that distinguish the *N. oceani*-like AmoA from the *Gammaproteobacteria* PmoA (Table 2) and from the enzyme encoded by JR2 and JR3. At position 71, the *N. oceani*-like clade contains an AmoA-diagnostic residue present in no known PmoAs. This residue is not present in JR2 or JR3. At five other sites, the *N. oceani*-like clade contains conserved residues distinct from known PmoAs and AmoAs; two residues are at PmoA-/AmoA-diagnostic positions, and three others are at positions conserved in all other PmoAs and AmoAs examined (Table 2), strongly suggesting functional relevance. None of these residues is present in JR2 or JR3. Finally, hydrophobicity plots of the consensus protein sequence of JR2 and JR3 show four transmembrane domains

at positions identical to those of the *Gammaproteobacteria* PmoA consensus; in contrast, the fourth domain of the consensus for *N. oceani*-like AmoA is shifted 12 residues towards the C terminus, exactly matching the position of the corresponding hydrophobic domain of AmoA found within the class *Betaproteobacteria* (data not shown). Together, these sequence analyses suggest strongly that JR2 and JR3 are more likely to preferentially bind methane than ammonia.

Response to simulated global change. It has been suggested that feedback between methane flux and climate change may be due to changes in the structure of the methanotroph community (31, 51); however, it is unknown whether realistic global changes have the potential to alter the community structure of methanotrophs. We used T-RFLP analyses of *pmoA* to provide a molecular profile of the methanotroph community and to determine if shifts in community structure occurred in response to simulated global change. We observed shifts in the relative abundance of both type II methanotrophs and the novel methanotroph clade JR2.

Type II methanotrophs decreased in relative abundance in response to increased precipitation (under ambient temperature) (Fig. 4, compare the open bars). Previous studies have reported decreased methane oxidation rates under increased soil moisture (1, 11, 54), possibly due to limitations on the diffusive transport of methane through the soil gas phase when soil moisture is high (31, 46). It is reasonable that reduced oxidation rates could result in the reduced relative abundance that we observed here, although this was not directly tested in our study. We also observed a significant decrease in the relative abundance of type II methanotrophs in response to increased temperature (under ambient precipitation) (Fig. 4, compare the open and hatched bars on the left). Although the diffusion of methane can be altered by temperature (31), and rates of methane oxidation are known to vary with temperature, the effect we observed is unlikely to be caused by the direct effects of temperature on methane supply or oxidation. The change in soil temperature in our plots due to the temperature treatment is negligible. However, the temperature treatment in our experiment has been reported to significantly increase soil moisture at the time of year at which we sampled, due to effects on the plant community that alter water loss from plant transpiration in the spring (60). It is thus plausible that the decrease in relative abundance we observed with increased temperature is due ultimately to the same mechanism as the decrease we observed with increased precipitation: an increase in soil water content. Indeed, soil water content was significantly correlated with the relative abundance of type II methanotrophs ($P = 0.0178$), while other factors (ammonium, nitrate, plant biomass, net primary productivity) were not.

In addition, we observed a significant interaction between precipitation and temperature, such that the combined effect of increased precipitation and temperature on type II methanotrophs was less than that expected by their individual effects. It is unclear why this might be. It is not due to nonadditive effects of temperature and precipitation on soil water content; there was not a significant interaction between these two factors in regard to soil water content in our study (data not shown). One possible explanation is that the negative effects of soil moisture on methane diffusivity are ameliorated at higher water contents by an increase in the proportion of anoxic

microsites in the soil, leading to a net increase in methanogenesis. This could result in the combined effects of temperature and precipitation being less than that expected by their individual effects, if when combined they raise the soil water content to a level where the proportion of anoxic microsites is increased. This hypothesis could be tested in future work by comparing the relative abundance of type II methanotrophs at our site across years that vary naturally in precipitation.

The relative abundance of the novel methanotroph clade JR2 also responded to elevated precipitation and temperature, although in a manner opposite of that of type II methanotrophs. The relative abundance of JR2 increased in response to elevated precipitation and temperature (Fig. 5), rather than decreased, as observed for type II methanotrophs. Why might JR2 have responded so differently from classical type II methanotrophs? If the methanotrophs in JR2 are atmospheric methane "specialists," as suggested by their association (although distant) with the "WB5FH-A" clade, then they might be expected to out-compete other methanotrophs under low-methane conditions. Such conditions could be present under conditions of relatively high soil water content, such as those resulting from increased precipitation or temperature, which would reduce the diffusion of methane into the soil. We observed a significant antagonistic interaction between elevated precipitation and temperature, such that the combined effect on the relative abundance of JR2 was less than that expected from their individual effects. Although it is unclear why this might be, it is possible that it could be due to same mechanism suggested for type II methanotrophs: simultaneous increases in temperature and precipitation increase soil moisture such that methanogenesis increases, increasing the methane supply and reducing the competitive advantage of JR2. Again, this is a testable hypothesis.

Our observations of significant interactions among global change factors are consistent with previous studies of global change. For example, Shaw and colleagues observed that antagonistic interactions among global changes could alter plant biomass at our site (48). Furthermore, Horz et al. (26) observed that both the abundance and community structure of ammonia-oxidizing bacteria at our site were altered by antagonistic interactions among global change factors, including interactions between temperature and precipitation.

Final conclusions. Our study expands our understanding of the diversity of naturally occurring *pmoA* gene types. The high number of novel *pmoA* clades we detected was possible because of our use of combinations of different PCR primers in a nested and multiplex manner. Since most other *pmoA*-based studies have relied on the use of only one primer set, it is plausible that the novel clades we observed are present in other environments as well but have been overlooked due to the primer set used.

Using this approach, we not only discovered novel *pmoA* clades, which evolutionary and sequence analyses suggest are functional, but we also observed that at least one such clade responded to simulated multifactorial global change in a very different manner than classic type II methanotrophs. To our knowledge, this is the first study that demonstrates that significant changes in the community structure of methanotrophs can occur in response to multifactorial global change. It is not yet known how widespread such responses are, how such re-

sponses may vary through time, or the relationship between such changes and ecosystem function. Nonetheless, our results demonstrate that methanotrophs can be altered by global changes and that multifactorial experimental approaches may be necessary to fully assess the complexity of these responses.

REFERENCES

1. **Adamsen, A. P. S., and G. M. King.** 1993. Methane consumption in temperate and sub-arctic forest soils—rates, vertical zonation, and responses to water and nitrogen. Appl. Environ. Microbiol. **59:**485–490.
2. **Alzerreca, J. J., J. M. Norton, and M. G. Klotz.** 1999. The *amo* operon in marine, ammonia-oxidizing γ-proteobacteria. FEMS Microbiol. Lett. **180:**21–29.
3. **Auman, A. J., S. Stolyar, A. M. Costello, and M. E. Lidstrom.** 2000. Molecular characterization of methanotrophic isolates from freshwater lake sediment. Appl. Environ. Microbiol. **66:**5259–5266.
4. **Baker, P. W., H. Futamata, S. Harayama, and K. Watanabe.** 2001. Molecular diversity of pMMO and sMMO in a TCE-contaminated aquifer during bioremediation. FEMS Microbiol. Ecol. **38:**161–167.
5. **Blake, D. R., and F. S. Rowland.** 1988. Continuing worldwide increase in tropospheric methane, 1978 to 1987. Science **239:**1129–1131.
6. **Bodrossy, L., E. M. Holmes, A. J. Holmes, K. L. Kovacs, and J. C. Murrell.** 1997. Analysis of 16S rRNA and methane monooxygenase gene sequences reveals a novel group of thermotolerant and thermophilic methanotrophs, *Methylocaldum* gen. nov. Arch. Microbiol. **168:**493–503.
7. **Bodrossy, L., K. L. Kovacs, I. R. McDonald, and J. C. Murrell.** 1999. A novel thermophilic methane-oxidizing gamma-proteobacterium. FEMS Microbiol. Lett. **170:**335–341.
8. **Bodrossy, L., J. C. Murrell, H. Dalton, M. Kalman, L. G. Puskas, and K. L. Kovacs.** 1995. Heat-tolerant methanotrophic bacteria from the hot water effluent of a natural gas field. Appl. Environ. Microbiol. **61:**3549–3555.
9. **Bodrossy, L., N. Stralis-Pavese, J. C. Murrell, S. Radajewski, A. Weilharter, and A. Sessitsch.** 2003. Development and validation of a diagnostic microbial microarray for methanotrophs. Environ. Microbiol. **5:**566–582.
10. **Bourne, D. G., I. R. McDonald, and J. C. Murrell.** 2001. Comparison of *pmoA* PCR primer sets as tools for investigating methanotroph diversity in three Danish soils. Appl. Environ. Microbiol. **67:**3802–3809.
11. **Castro, M. S., P. A. Steudler, J. M. Melillo, J. D. Aber, and R. D. Bowden.** 1995. Factors controlling atmospheric methane consumption by temperate forest soils. Global Biogeochem. Cycles **9:**1–10.
12. **Costello, A. M., and M. E. Lidstrom.** 1999. Molecular characterization of functional and phylogenetic genes from natural populations of methanotrophs in lake sediments. Appl. Environ. Microbiol. **65:**5066–5074.
13. **Dedysh, S. N., H. P. Horz, P. F. Dunfield, and W. Liesack.** 2001. A novel *pmoA* lineage represented by the acidophilic methanotrophic bacterium *Methylocapsa acidiphila* [correction of *acidophila*] B2. Arch. Microbiol. **177:**117–121.
14. **Dedysh, S. N., V. N. Khmelenina, N. E. Suzina, Y. A. Trotsenko, J. D. Semrau, W. Liesack, and J. M. Tiedje.** 2002. *Methylocapsa acidiphila* gen. nov., sp. nov., a novel methane-oxidizing and dinitrogen-fixing acidophilic bacterium from *Sphagnum* bog. Int. J. Syst. Evol. Microbiol. **52:**251–261.
15. **Dedysh, S. N., W. Liesack, V. N. Khmelenina, N. E. Suzina, Y. A. Trotsenko, J. D. Semrau, A. M. Bares, N. S. Panikov, and J. M. Tiedje.** 2000. *Methylocella palustris* gen. nov., sp. nov., a new methane-oxidizing acidophilic bacterium from peat bogs, representing a novel subtype of serine-pathway methanotrophs. Int. J. Syst. Evol. Microbiol. **50:**955–969.
16. **Dedysh, S. N., N. S. Panikov, W. Liesack, R. Grosskopf, J. Zhou, and J. M. Tiedje.** 1998. Isolation of acidophilic methane-oxidizing bacteria from northern peat wetlands. Science **282:**281–284.
17. **Dunfield, P. F., M. T. Yimga, S. N. Dedysh, U. Berger, W. Liesack, and J. Heyer.** 2002. Isolation of a *Methylocystis* strain containing a novel *pmoA*-like gene. FEMS Microbiol. Ecol. **41:**17–26.
18. **Fjellbirkeland, A., V. Torsvik, and L. Ovreas.** 2001. Methanotrophic diversity in an agricultural soil as evaluated by denaturing gradient gel electrophoresis profiles of *pmoA*, *mxaF* and 16S rDNA sequences. Antonie Leeuwenhoek **79:**209–217.
19. **Hanson, R. S., and T. E. Hanson.** 1996. Methanotrophic bacteria. Microbiol. Rev. **60:**439–471.
20. **Henckel, T., M. Friedrich, and R. Conrad.** 1999. Molecular analyses of the methane-oxidizing microbial community in rice field soil by targeting the

genes of the 16S rRNA, particulate methane monooxygenase, and methanol dehydrogenase. Appl. Environ. Microbiol. 65:1980–1990.

21. Henckel, T., U. Jaekel, S. Schnell, and R. Conrad. 2000. Molecular analyses of novel methanotrophic communities in forest soil that oxidize atmospheric methane. Appl. Environ. Microbiol. 66:1801–1808.

22. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89:10915–10919.

23. Hoffmann, T., H. P. Horz, D. Kemnitz, and R. Conrad. 2002. Diversity of the particulate methane monooxygenase gene in methanotrophic samples from different rice field soils in China and the Philippines. Syst. Appl. Microbiol. 25:267–274.

24. Holmes, A., A. Costello, M. Lidstrom, and J. Murrell. 1995. Evidence that particulate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. FEMS Microbiol. Lett. 132:203–208.

25. Holmes, A. J., P. Roslev, I. R. McDonald, N. Iversen, K. Henriksen, and J. C. Murrell. 1999. Characterization of methanotrophic bacterial populations in soils showing atmospheric methane uptake. Appl. Environ. Microbiol. 65:3312–3318.

26. Horz, H.-P., A. Barbrook, C. B. Field, and B. J. M. Bohannan. 2004. Ammonia-oxidizing bacteria respond to multifactorial global change. Proc. Natl. Acad. Sci. USA 101:15136–15141.

27. Horz, H.-P., A. Raghubanshi, E. Heyer, C. Kammann, R. Conrad, and P. Dunfield. 2002. Activity and community structure of methane-oxidizing bacteria in a wet meadow soil. FEMS Microbiol. Ecol. 41:247–257.

28. Horz, H.-P., M. Tchawa Yimga, and W. Liesack. 2001. Detection of methanotroph diversity on roots of submerged rice plants by molecular retrieval of pmoA, mmoX, mxaF, and 16S rRNA and ribosomal DNA, including pmoA-based terminal restriction fragment length polymorphism profiling. Appl. Environ. Microbiol. 67:4177–4185.

29. Jaatinen, K., C. Knief, P. F. Dunfield, K. Yrjälä, and H. Fritze. 2004. Methanotrophic bacteria in boreal forest soil after fire. FEMS Microbiol. Ecol. 50:195–202.

30. Juretschko, S., G. Timmerman, M. Schmid, K. H. Schleifer, A. Pommerening-Roser, H. P. Koops, and M. Wagner. 1998. Combined molecular and conventional analyses of nitrifying bacterium diversity in activated sludge: Nitrosococcus mobilis and Nitrosospira-like bacteria as dominant populations. Appl. Environ. Microbiol. 64:3042–3051.

31. King, G. M. 1997. Responses of atmospheric methane consumption by soils to global climate change. Global Change Biol. 3:351–362.

32. Knief, C., A. Lipski, and P. F. Dunfield. 2003. Diversity and activity of methanotrophic bacteria in different upland soils. Appl. Environ. Microbiol. 69:6703–6714.

33. Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, A. Yadhukumar, T. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer. 2004. ARB: a software environment for sequence data. Nucleic Acids Res. 32:1363–1371.

34. Magurran, A. E. 1988. Ecological diversity and its measurement. Princeton University Press, Princeton, N.J.

35. Morris, S. A., S. Radajewski, T. W. Willison, and J. C. Murrell. 2002. Identification of the functionally active methanotroph population in a peat soil microcosm by stable-isotope probing. Appl. Environ. Microbiol. 68:1446–1453.

36. Murrell, J. C., I. R. McDonald, and D. G. Bourne. 1998. Molecular methods for the study of methanotroph ecology. FEMS Microbiol. Ecol. 27:103–114.

37. Norton, J. M., J. J. Alzerreca, J. Suwa, and M. G. Klotz. 2002. Diversity of ammonia monooxygenase operon in autotrophic ammonia-oxidizing bacteria. Arch. Microbiol. 177:139–149.

38. Osborn, A. M., E. R. B. Moore, and K. N. Timmis. 2000. An evaluation of terminal-restriction fragment length polymorphism (T-RFLP) analysis for the study of microbial community structure and dynamics. Environ. Microbiol. 2:39–50.

39. Pacheco-Oliver, M., I. R. McDonald, D. Gruleau, J. C. Murrell, and C. B.

Miguez. 2002. Detection of methanotrophs with highly divergent pmoA genes from Arctic soils. FEMS Microbiol. Lett. 209:313–319.

40. Pester, M., M. W. Friedrich, B. Schink, and A. Brune. 2004. pmoA-based analysis of methanotrophs in a littoral lake sediment reveals a diverse and stable community in a dynamic environment. Appl. Environ. Microbiol. 70:3138–3142.

41. Reeburgh, W. S., S. C. Whalen, and M. J. Alperin. 1993. The role of methylotrophy in the global methane budget, p. 1–14. In J. C. Murrell and D. P. Kelly (ed.), Microbial growth on C1 compounds. Intercept, Andover, United Kingdom.

42. Ricke, P., M. Erkel, R. Kube, R. Reinhardt, and W. Liesack. 2004. Comparative analysis of the conventional and novel pmo (particulate methane monooxygenase) operons from Methylocystis strain SC2. Appl. Environ. Microbiol. 70:3055–3063.

43. Rillig, M., S. Wright, M. Shaw, and C. Field. 2002. Artificial climate warming positively affects arbuscular mycorrhizae but decreases soil aggregate water stability in an annual grassland. Oikos 97:52–58.

44. Rodhe, H. 1990. A comparison of the contribution of various gases to the greenhouse-effect. Science 248:1217–1219.

45. Sakano, Y., and L. Kerkhof. 1998. Assessment of changes in microbial community structure during operation of an ammonia biofilter with molecular tools. Appl. Environ. Microbiol. 64:4877–4882.

46. Schnell, S., and G. M. King. 1996. Responses of methanotrophic activity in soils and cultures to water stress. Appl. Environ. Microbiol. 62:3203–3209.

47. Semrau, J. D., A. Chistoserdov, J. Lebron, A. Costello, J. Davagnino, E. Kenna, A. J. Holmes, R. Finch, J. C. Murrell, and M. E. Lidstrom. 1995. Particulate methane monooxygenase genes in methanotrophs. J. Bacteriol. 177:3071–3079.

48. Shaw, M. R., E. S. Zavaleta, N. R. Chiariello, E. E. Cleland, H. A. Mooney, and C. B. Field. 2002. Grassland responses to global environmental changes suppressed by elevated CO₂. Science 298:1987–1990.

49. Swofford, D. L. 2002. PAUP: phylogenetic analysis using parsimony, 4.0b10 ed. Sinauer Associates, Sunderland, Mass.

50. Tchawa Yimga, M., P. F. Dunfield, P. Ricke, J. Heyer, and W. Liesack. 2003. Wide distribution of a novel pmoA-like gene copy among type II methanotrophs, and its expression in Methylocystis strain SC2. Appl. Environ. Microbiol. 69:5593–5602.

51. Torn, M. S., and J. Harte. 1996. Methane consumption by montane soils: implications for positive and negative feedback with climatic change. Biogeochemistry 32:53–67.

52. Tukhvatullin, I. A., R. I. Gvozdev, and K. K. Anderson. 2001. Structural and functional model of methane hydroxylase of membrane-bound methane monooxygenase from Methylococcus capsulatus (Bath). Russ. Chem. Bull. 50:1867–1876.

53. Ward, B. B., and G. D. O'Mullan. 2002. Worldwide distribution of Nitrosococcus oceani, a marine ammonia-oxidizing γ-proteobacterium, detected by PCR and sequencing of 16S rRNA and amoA genes. Appl. Environ. Microbiol. 68:4153–4157.

54. Whalen, S. C., W. S. Reeburgh, and K. A. Sandbeck. 1990. Rapid methane oxidation in a landfill cover soil. Appl. Environ. Microbiol. 56:3405–3411.

55. Yang, Z. 2001. Adaptive molecular evolution. In D. J. Balding, M. Bishop, and C. Cannings (ed.), Handbook of statistical genetics. John Wiley and Sons, Ltd., Hoboken, N.J.

56. Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15:496–503.

57. Yang, Z. H. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573.

58. Yang, Z. H. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

59. Yang, Z. H., and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–917.

60. Zavaleta, E. S., B. D. Thomas, N. R. Chiariello, G. P. Asner, M. R. Shaw, and C. B. Field. 2003. Plants reverse warming effect on ecosystem water balance. Proc. Natl. Acad. Sci. USA 100:9892–9893.

Appendix 4

Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior.

Authors:
Edward F. DeLong, Christina M. Preston, Tracy Mincer, Virginia Rich, Steven J. Hallam, Niels-Ulrik Frigaard, Asuncion Martinez, Matthew B. Sullivan, Robert Edwards, Beltran Rodriguez Brito, Sallie W. Chisholm, David M. Karl.

32. Materials and methods are available as supporting material on *Science* Online.
33. J. R. Abo-Shaeer, C. Raman, W. Ketterle, *Phys. Rev. Lett.* **88**, 070409 (2002).
34. O. Guéry-Odelin, *Phys. Rev. A* **62**, 033607 (2000).
35. Y. Kagan, E. L. Surkov, G. V. Shlyapnikov, *Phys. Rev. A* **55**, R18 (1997).
36. C. Menotti, P. Pedri, S. Stringari, *Phys. Rev. Lett.* **89**, 250402 (2002).
37. G. B. Partridge, W. Li, R. I. Kamar, Y.-a. Liao, R. G. Hulet, *Science*, **311**, 503 (2006); published online 22 December 2005 (10.1126/science.1122876).
38. T. Mizushima, K. Machida, M. Ichioka, *Phys. Rev. Lett.* **94**, 060404 (2005).
39. P. Castorina, M. Grasso, M. Oertel, M. Urban, D. Zappalà, *Phys. Rev. A* **72**, 025601 (2005).
40. A. A. Abrikosov, L. P. Gorkov, I. E. Dzyaloshinski, *Methods of Quantum Field Theory in Statistical Physics* (Dover, New York, 1975).
41. G. Bertsch, INT Workshop on Effective Field Theory in Nuclear Physics (Seattle, WA, February 1999).
42. T. D. Cohen, *Phys. Rev. Lett.* **95**, 120403 (2005).
43. We thank G. Campbell for critical reading of the manuscript and X.-G. Wen, E. Demler, and S. Sachdev for stimulating

32. Materials and methods are available as supporting material on *Science* Online.
33. J. R. Abo-Shaeer, C. Raman, W. Ketterle, *Phys. Rev. Lett.* **88**, 070409 (2002).
34. O. Guéry-Odelin, *Phys. Rev. A* **62**, 033607 (2000).
35. Y. Kagan, E. L. Surkov, G. V. Shlyapnikov, *Phys. Rev. A* **55**, R18 (1997).
36. C. Menotti, P. Pedri, S. Stringari, *Phys. Rev. Lett.* **89**, 250402 (2002).
37. G. B. Partridge, W. Li, R. I. Kamar, Y.-a. Liao, R. G. Hulet, *Science*, **311**, 503 (2006); published online 22 December 2005 (10.1126/science.1122876).
38. T. Mizushima, K. Machida, M. Ichioka, *Phys. Rev. Lett.* **94**, 060404 (2005).
39. P. Castorina, M. Grasso, M. Oertel, M. Urban, D. Zappalà, *Phys. Rev. A* **72**, 025601 (2005).
40. A. A. Abrikosov, L. P. Gorkov, I. E. Dzyaloshinski, *Methods of Quantum Field Theory in Statistical Physics* (Dover, New York, 1975).
41. G. Bertsch, INT Workshop on Effective Field Theory in Nuclear Physics (Seattle, WA, February 1999).
42. T. D. Cohen, *Phys. Rev. Lett.* **95**, 120403 (2005).
43. We thank G. Campbell for critical reading of the manuscript and X.-G. Wen, E. Demler, and S. Sachdev for stimulating

Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior

Edward F. DeLong,[1]* Christina M. Preston,[2] Tracy Mincer,[1] Virginia Rich,[1] Steven J. Hallam,[1] Niels-Ulrik Frigaard,[1] Asuncion Martinez,[1] Matthew B. Sullivan,[1] Robert Edwards,[3] Beltran Rodriguez Brito,[3] Sallie W. Chisholm,[1] David M. Karl[4]

Microbial life predominates in the ocean, yet little is known about its genomic variability, especially along the depth continuum. We report here genomic analyses of planktonic microbial communities in the North Pacific Subtropical Gyre, from the ocean's surface to near–sea floor depths. Sequence variation in microbial community genes reflected vertical zonation of taxonomic groups, functional gene repertoires, and metabolic potential. The distributional patterns of microbial genes suggested depth-variable community trends in carbon and energy metabolism, attachment and motility, gene mobility, and host-viral interactions. Comparative genomic analyses of stratified microbial communities have the potential to provide significant insight into higher-order community organization and dynamics.

Microbial plankton are centrally involved in fluxes of energy and matter in the sea, yet their vertical distribution and functional variability in the ocean's interior is still only poorly known. In contrast, the vertical zonation of eukaryotic phytoplankton and zooplankton in the ocean's water column has been well documented for over a century (*1*). In the photic zone, steep gradients of light quality and intensity, temperature, and macronutrient and trace-metal concentrations all influence species distributions in the water column (*2*). At greater depths, low temperature, increasing hydrostatic pressure, the disappearance of light, and dwindling energy supplies largely determine vertical stratification of oceanic biota.

For a few prokaryotic groups, vertical distributions and depth-variable physiological properties are becoming known. Genotypic and phenotypic properties of stratified *Prochlorococcus* "ecotypes" for example, are suggestive of depth-variable adaptation to light intensity and nutrient availability (*3–5*). In the abyss, the vertical zonation of deep-sea piezophilic bacteria can be explained in part by their obligate growth requirement for elevated hydrostatic pressures (*6*). In addition, recent cultivation-independent (*7–15*) surveys have shown vertical zonation patterns among specific groups of planktonic *Bacteria*, *Archaea*, and *Eukarya*. Despite recent progress however, a comprehensive description of the biological properties and vertical distributions of planktonic microbial species is far from complete.

Cultivation-independent genomic surveys represent a potentially useful approach for characterizing natural microbial assemblages (*16, 17*). "Shotgun" sequencing and whole genome assembly from mixed microbial assemblages has been attempted in several environments, with varying success (*18, 19*). In addition, Tringe *et al.* (*20*) compared shotgun sequences of several disparate microbial assemblages to identify community-specific patterns in gene distributions. Metabolic reconstruction has also been attempted with environmental genomic approaches (*21*). Nevertheless, integrated genomic surveys of microbial communities along well-defined environmental gradients (such as the ocean's water column) have not been reported.

To provide genomic perspective on microbial biology in the ocean's vertical dimension, we cloned large [~36 kilobase pairs (kbp)] DNA fragments from microbial communities at different depths in the North Pacific Subtropical Gyre (NPSG) at the open-ocean time-series station ALOHA (*22*). The vertical distribution of microbial genes from the ocean's surface to abyssal depths was determined by shotgun sequencing of fosmid clone termini. Applying identical collection, cloning, and sequencing strategies at seven depths (ranging from 10 m to 4000 m), we archived large-insert genomic libraries from each depth-stratified microbial community. Bidirectional DNA sequencing of fosmid clones (~10,000 sequences per depth) and comparative sequence analyses were used to identify taxa, genes, and metabolic pathways that characterized vertically stratified microbial assemblages in the water column.

Study Site and Sampling Strategy

Our sampling site, Hawaii Ocean Time-series (HOT) station ALOHA (22°45' N, 158°W), represents one of the most comprehensively characterized sites in the global ocean and has been a focal point for time series–oriented oceanographic studies since 1988 (*22*). HOT investigators have produced high-quality spatial and time-series measurements of the defining physical, chemical, and biological oceanographic parameters from surface waters to the seafloor. These detailed spatial and temporal datasets present unique opportunities for placing microbial genomic depth profiles into appropriate oceanographic context (*22–24*) and leverage these data to formulate meaningful ecological hypotheses. Sample depths were selected, on the basis of well-defined physical, chemical, and biotic characteristics, to represent discrete zones in the water column (Tables 1 and 2, Fig. 1; figs. S1 and S2). Specifically, seawater samples from the upper euphotic zone (10 m and 70 m), the base of the chlorophyll maximum (130 m), below the base of the euphotic zone (200 m), well below the upper mesopelagic (500 m), in the core of the dissolved oxygen minimum layer (770 m), and in the deep abyss, 750 m above the seafloor (4000 m), were collected for preparing microbial community DNA libraries (Tables 1 and 2, Fig. 1; figs. S1 and S2).

The depth variability of gene distributions was examined by random, bidirectional end-sequencing of ~5000 fosmids from each depth, yielding ~64 Mbp of DNA sequence total from the 4.5 Gbp archive (Table 1). This represents raw sequence coverage of about 5 (1.8 Mbp sized) genome equivalents per depth. Because we surveyed ~180 Mbp of cloned DNA (5000 clones by

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Monterey Bay Aquarium Research Institute, Moss Landing, CA 95064, USA. [3]San Diego State University, San Diego, CA 92182, USA. [4]University of Hawaii, Honolulu, HI 96822, USA.

*To whom correspondence should be addressed. E-mail: delong@mit.edu

~36 kbp/clone per depth), however, we directly sampled ~100 genome equivalents at each depth. We did not sequence as deeply in each sample as a recent Sargasso Sea survey (19), where from 90,000 to 600,000 sequences were obtained from small DNA insert clones, from each of seven different surface-water samples. We hypothesized, however, that our comparison of microbial communities collected along well-defined environmental gradients (using large-insert DNA clones), would facilitate detection of ecologically meaningful taxonomic, functional, and community trends.

Vertical Profiles of Microbial Taxa

Vertical distributions of bacterial groups were assessed by amplifying and sequencing small subunit (SSU) ribosomal RNA (rRNA) genes from complete fosmid library pools at each depth (Fig. 2; fig. S3). Bacterial phylogenetic distributions were generally consistent with previous polymerase chain reaction based cultivation-independent rRNA surveys of marine picoplankton (8, 15, 25). In surface-water samples, rRNA-containing fosmids included those from *Prochlorococcus*; *Verrucomicrobiales*; *Flexibacteraceae*; Gammaproteobacteria (SAR92, OM60, SAR86 clades); Alphaproteobacteria (SAR116, OM75 clades); and Deltaproteobacteria (OM27 clade) (Fig. 2). Bacterial groups from deeper waters included members of *Deferribacteres*; *Planctomycetaceae*; *Acidobacteriales*; *Gemmatamonadaceae*; *Nitrospina*; *Alteromonadaceae*; and SAR202, SAR11, and Agg47 planktonic bacterial clades (Fig. 2; fig. S2). Large-insert DNA clones previously recovered from the marine environment (9, 10) also provide a good metric for taxonomic assessment of indigenous microbes. Accordingly, a relatively large proportion of our shotgun fosmid sequences most closely matched rRNA-containing bacterioplankton artificial clones previously recovered from the marine environment (fig. S3).

Taxonomic bins of bacterial protein homologs found in randomly sequenced fosmid ends (Fig. 2; fig. S4) also reflected distributional patterns generally consistent with previous surveys in the water column (8, 15). Unexpectedly large amounts of phage DNA were recovered in clones, particularly in the photic zone. Also unexpected was a relatively high proportion of Betaproteobacteria-like sequences recovered at 130 m, most sharing highest similarity to protein homologs from *Rhodoferax ferrireducens*. As expected, representation of *Prochlorococcus*-like and *Pelagibacter*-like genomic sequences was high in the photic zone. At greater depths, higher proportions of *Chloroflexi*-like sequences, perhaps corresponding to the co-occurring SAR202 clade, were observed (Fig. 2). *Planctomycetales*-like genomic DNA sequences were also highly represented at greater depths.

All archaeal SSU rRNA containing fosmids were identified at each depth, quantified by macroarray hybridization, and their rRNAs sequenced

Table 1. HOT samples and fosmid libraries. Sample site, 22°45′ N, 158°W. All seawater samples were pre-filtered through a 1.6-μm glass fiber filter, and collected on a 0.22-μm filter. See (35) for methods.

Depth (m)	Sample date	Volume filtered (liters)	Total fosmid clones	Total DNA (Mbp)	
				Archived	Sequenced
10	10/7/02	40	12,288	442	7.54
70	10/7/02	40	12,672	456	11.03
130	10/6/02	40	13,536	487	6.28
200	10/6/02	40	19,008	684	7.96
500	10/6/02	80	15,264	550	8.86
770	12/21/03	240	11,520	415	11.18
4,000	12/21/03	670	41,472	1,493	11.10

Table 2. HOT sample oceanographic data. Samples described in Table 1. Oceanographic parameters were measured as specified at (49); values shown are those from the same CTD casts as the samples, where available. Values in parentheses are the mean ± 1 SD of each core parameter during the period October 1988 to December 2004, with the total number of measurements collected for each parameter shown in brackets. The parameter abbreviations are Temp., Temperature; Chl a, chlorophyll a; DOC, dissolved organic carbon; N+N, nitrate plus nitrite; DIP, dissolved inorganic phosphate; and DIC, dissolved inorganic carbon. The estimated photon fluxes for upper water column samples (assuming a surface irradiance of 32 mol quanta m^{-2} d^{-1} and a light extinction coefficient of 0.0425 m^{-1}) were: 10 m = 20.92 (65% of surface), 70 m = 1.63 (5% of surface), 130 m = 0.128 (0.4% of surface), 200 m = 0.07 (0.02% of surface). The mean surface mixed-layer during the October 2002 sampling was 61 m. Data are available at (50). *Biomass derived from particulate adenosine triphosphate (ATP) measurements assuming a carbon:ATP ratio of 250. ND, Not determined.

Depth (m)	Temp. (°C)	Salinity	Chl a (μg/kg)	Biomass* (μg/kg)	DOC (μmol/kg)	N+N (nmol/kg)	DIP (nmol/kg)	Oxygen (μmol/kg)	DIC (μmol/kg)
10	26.40	35.08	0.08	7.21 ± 2.68	78	1.0	41.0	204.6	1,967.6
	(24.83 ± 1.27)	(35.05 ± 0.21)	(0.08 ± 0.03)	[78]	(90.6 ± 14.3)	(2.6 ± 3.7)	(56.0 ± 33.7)	(209.3 ± 4.5)	(1,972.1 ± 16.4)
	[2,104]	[1,611]	[320]		[140]	[126]	[146]	[348]	[107]
70	24.93	35.21	0.18	8.51 ± 3.22	79	1.3	16.0	217.4	1,981.8
	(23.58 ± 1.00)	(35.17 ± 0.16)	(0.15 ± 0.05)	[86]	(81.4 ± 11.3)	(14.7 ± 60.3)	(43.1 ± 25.1)	(215.8 ± 5.4)	(1,986.9 ± 15.4)
	[1,202]	[1,084]	[363]		[79]	[78]	[104]	[144]	[84]
130	22.19	35.31	0.10	5.03 ± 2.30	69	284.8	66.2	204.9	2,026.5
	(21.37 ± 0.96)	(35.20 ± 0.10)	(0.15 ± 0.06)	[90]	(75.2 ± 9.1)	(282.9 ± 270.2)	(106.0 ± 49.7)	(206.6 ± 6.2)	(2,013.4 ± 13.4)
	[1,139]	[980]	[350]		[86]	[78]	[68]	[173]	[69]
200	18.53	35.04	0.02	1.66 ± 0.24	63	1,161.9 ± 762.5	274.2 ± 109.1	198.8	2,047.7
	(18.39 ± 1.29)	(34.96 ± 0.18)	(0.02 ± 0.02)	[2]	(64.0 ± 9.8)	[7]	[84]	(197.6 ± 7.1)	(2,042.8 ± 10.5)
	[662]	[576]	[97]		[113]			[190]	[125]
500	7.25	34.07	N0	0.48 ± 0.23	47	28,850	2,153	118.0	2197.3
	(7.22 ± 0.44)	(34.06 ± 0.03)		[107]	(47.8 ± 6.3)	(28,460 ± 2210)	(2,051 ± 175.7)	(120.5 ± 18.3)	(2,200.2 ± 17.8)
	[1,969]	[1,769]			[112]	[326]	[322]	[505]	[134]
770	4.78	34.32	ND	0.29 ± 0.16	39.9	41,890	3,070	32.3	2323.8
	(4.86 ± 0.21)	(34.32 ± 0.04)		[107]	(41.5 ± 4.4)	(40,940 ± 500)	(3,000 ± 47.1)	(27.9 ± 4.1)	(2,324.3 ± 6.1)
	[888]	[773]			[34]	[137]	[135]	[275]	[34]
4,000	1.46	34.69	ND	ND	37.5	36,560	2,558	147.8	2325.5
	(1.46 ± 0.01)	(34.69 ± 0.00)			(42.3 ± 4.9)	(35,970 ± 290)	(2,507 ± 19)	(147.8 ± 1.3)	(2,329.1 ± 4.8)
	[262]	[245]			[83]	[108]	[104]	[210]	[28]

(figs. S5 and S6). The general patterns of archaeal distribution we observed were consistent with previous field surveys (15, 25, 26). Recovery of "group II" planktonic *Euryarchaeota* genomic DNA was greatest in the upper water column and declined below the photic zone. This distribution corroborates recent observations of ion-translocating photoproteins (called proteorhodopsins), now known to occur in group II *Euryarchaeota* inhabiting the photic zone (27). "Group III" *Euryarchaeota* DNA was recovered at all depths, but at a much lower frequency (figs. S5 and S6). A novel crenarchaeal group, closely related to a putatively thermophilic *Crenarchaeota* (28), was observed at the greatest depths (fig. S6).

Vertically Distributed Genes and Metabolic Pathways

The depths sampled were specifically chosen to capture microbial sequences at discrete biogeochemical zones in the water column encompassing key physicochemical features (Tables 1 and 2, Fig. 1; figs. S1 and S2). To evaluate sequences from each depth, fosmid end sequences were compared against different databases including the Kyoto Encyclopedia of Genes and Genomes (KEGG) (29), National Center for Biotechnology Information (NCBI)'s Clusters of Orthologous Groups (COG) (30), and SEED subsystems (31). After categorizing sequences from each depth in BLAST searches (32) against each database, we identified protein categories that were more or less well represented in one sample versus another, using cluster analysis (33, 34) and bootstrap resampling methodologies (35).

Cluster analyses of predicted protein sequence representation identified specific genes and metabolic traits that were differentially distributed in the water column (fig. S7). In the photic zone (10, 70, and 130 m), these included a greater representation in sequences associated with photosynthesis; porphyrin and chlorophyll metabolism; type III secretion systems; and aminosugars, purine, propanoate, and vitamin B6 metabolism, relative to deep-water samples (fig. S7). Independent comparisons with well-annotated subsystems in the SEED database (31) also showed similar and overlapping trends (table S1), including greater representation in photic zone sequences associated with alanine and aspartate; metabolism of aminosugars; chlorophyll and carotenoid biosynthesis; maltose transport; lactose degradation; and heavy metal ion sensors and exporters. In contrast, samples from depths of 200 m and below (where there is no photosynthesis) were enriched in different sequences, including those associated with protein folding; processing and export; methionine metabolism; glyoxylate, dicarboxylate, and methane metabolism; thiamine metabolism; and type II secretion systems, relative to surface-water samples (fig. S7).

COG categories also provided insight into differentially distributed protein functions and categories. COGs more highly represented in photic zone included iron-transport membrane receptors,

deoxyribopyrimidine photolyase, diaminopimelate decarboxylase, membrane guanosine triphosphatase (GTPase) with the lysyl endopeptidase gene product LepA, and branched-chain amino acid–transport system components (fig. S8). In contrast, COGs with greater representation in deep-water samples included transposases, several dehydrogenase categories, and integrases (fig. S8). Sequences more highly represented in the deep-water samples in SEED subsystem (31) comparisons included those associated with respiratory dehydrogenases, polyamine adenosine triphosphate (ATP)–binding cassette (ABC) transporters, polyamine metabolism, and alkylphosphonate transporters (table S1).

Habitat-enriched sequences. We estimated average protein sequence similarities between all depth bins from cumulative TBLASTX high-scoring sequence pair (HSP) hitscores, derived from BLAST searches of each depth against every other (Fig. 3). Neighbor-joining analyses of a normalized, distance matrix derived from these cumulative hitscores joined photic zone and deeper samples together in separate clusters (Fig. 3). When we compared our HOT sequence datasets to previously reported Sargasso Sea microbial sequences (19), these datasets also clustered according to their depth and size fraction of origin (fig. S9). The clustering pattern in Fig. 3 is consistent with the expectation that randomly sampled photic zone microbial sequences will tend on average to be more similar to one another, than to those from the deep-sea, and vice-versa.

We also identified those sequences (some of which have no homologs in annotated databases)

that track major depth-variable environmental features. Specifically, sequence homologs found only in the photic zone unique sequences (from 10, 70, and 130 m), or deepwater unique sequences (from 500, 770, and 4000 m) were identified (Fig. 3). To categorize potential functions encoded in these photic zone unique (PZ) or deep-water unique (DW) sequence bins, each was compared with KEGG, COG, and NCBI protein databases in separate analyses (29, 30, 36).

Some KEGG metabolic pathways appeared more highly represented in the PZ than in DW sequence bins, including those associated with photosynthesis; porphyrin and chlorophyll metabolism; propanoate, purine, and glycerphospholipid metabolism; bacterial chemotaxis; flagellar assembly; and type III secretion systems (Fig. 4A). All proteorhodopsin sequences (except one) were captured in the PZ bin. Well-represented photic zone KEGG pathway categories appeared to reflect potential pathway interdependencies. For example the PZ photosynthesis bin [3% of the total (Fig. 4A)] contained *Prochlorococcus*-like and *Synechococcus*-like photosystem I, photosystem II, and cytochrome genes. In tandem, PZ porphyrin and chlorophyll biosynthesis sequence bins [~3.9% of the total (Fig. 4A)] contained high representation of cyanobacteria-like cobalamin and chlorophyll biosynthesis genes, as well as photoheterotroph-like bacteriochlorophyll biosynthetic genes. Other probable functional interdependencies appear reflected in the corecovery of sequences associated with chemotaxis (mostly methyl-accepting chemotaxis proteins), flagellar biosynthesis (predominant-
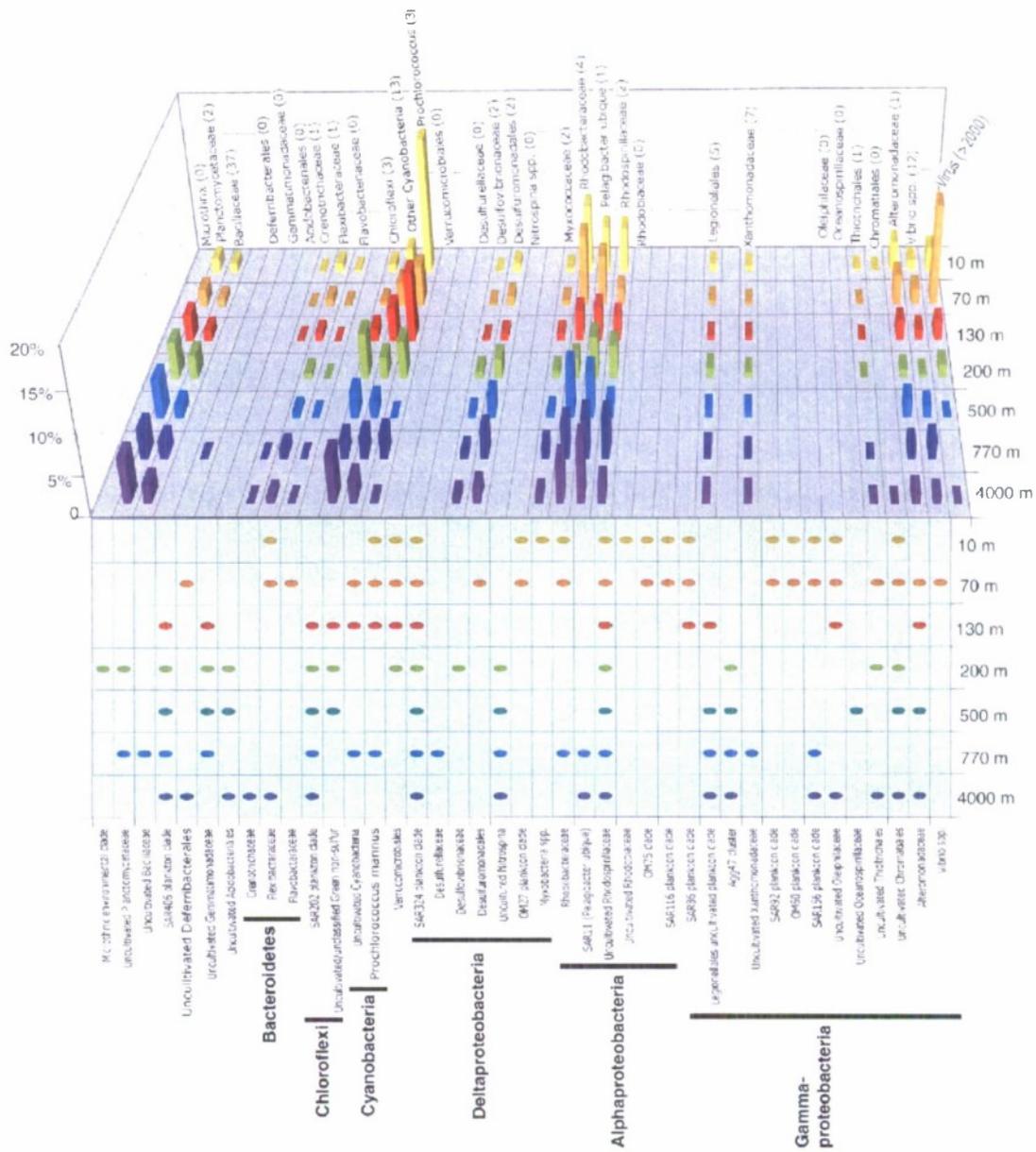

Fig. 1. Temperature versus salinity (T-S) relations for the North Pacific Subtropical Gyre at station ALOHA (22°45′N, 158°W). The blue circles indicate the positions, in T-S "hydrospace" of the seven water samples analyzed in this study. The data envelope shows the temperature and salinity conditions observed during the period October 1988 to December 2004 emphasizing both the temporal variability of near-surface waters and the relative constancy of deep waters.

Fig. 2. Taxon distributions of top HSPs. The percent top HSPs that match the taxon categories shown at expectation values of $\leq 1 \times 10^{-60}$. Values in parentheses indicate num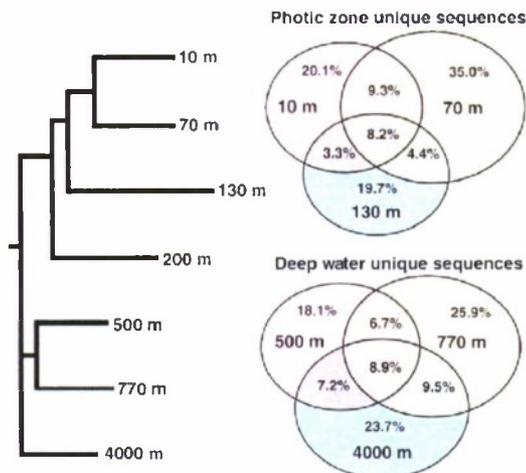ber of genomes in each category, complete or draft, that were in the database at the time of analysis. The dots in the lower panel tabulate the SSU rRNAs detected in fosmid libraries from each taxonomic group at each depth (35) (figs. S3 and S6).

ly flagellar motor and hook protein-encoding genes), and type III secretory pathways (all associated with flagellar biosynthesis) in PZ (Fig. 4A).

DW sequences were enriched in several KEGG categories, including glyoxylate and dicarboxylate metabolism (with high representation of isocitrate lyase– and formate dehydrogenase–

like genes); protein folding and processing (predominantly chaperone and protease like genes); type II secretory genes (~40% were most similar to pilin biosynthesis genes); aminophospho-

251

Photic zone unique sequences

Deep water unique sequences

nate, methionine, and sulfur metabolism; butanoate metabolism; ion-coupled transporters; and other ABC transporter variants (Fig. 4B). The high representation in DW sequences of type II secretion system and pilin biosynthesis genes, polysaccharide, and antibiotic synthesis suggest a potentially greater role for surface-associated microbial processes in the deeper-water communities. Conversely, enrichment of bacterial motility and chemotaxis sequences in the photic zone indicates a potentially greater importance for mobility and response in these assemblages.

Similar differential patterns of sequence distribution were seen in COG categories (Fig. 4B). COGs enriched in the PZ sequence bin included photolyases, iron-transport outer membrane proteins, Na^+-driven efflux pumps, ABC-type sugar-transport systems, hydrolases and acyl transferases, and transaldolases. In deeper waters, transposases were the most enriched COG category (~4.5% of the COG-categorized DW), increasing steadily in representation with depth from 500 m to their observed maximum at 4000 m (Fig. 4B; fig. S9). Transposases represented one of the single-most overrepresented COG categories in deep waters, accounting for 1.2% of all fosmids sequenced from 4000 m (fig. S8). Preliminary analyses of the transposase variants and mate-pair sequences indicate that they represent a wide variety of different transposase families and originate from diverse microbial taxa. In contrast, other highly represented COG categories appeared to reflect specific taxon distribution and abundances. For example, the enrichment of transaldolases at 70 m (Fig. 4B; fig. S9) were mostly derived from abundant cyanophage DNA that was recovered at that depth (see discussion below).

Sargasso Sea surface-water microbial sequences (19) shared, as expected, many more homologous sequences with our photic zone sequences than those from the deep sea (fig. S10). There were 10 times as many PZ than DW sequences shared in common with Sargasso Sea samples 5 through 7 (19) (fig. S10). In contrast, PZ-like sequences were only three times higher in DW when compared with sequences from Sargasso Sea sample 3 (fig. S10). The fact that Sargasso sample 3 was collected during a period of winter deep-water mixing likely contributes to this higher representation of DW-like homologs. Sargasso Sea homologs of our PZ sequence bin included, as expected, sequences associated with photosynthesis; amino acid transport; purine, pyrimidine and nitrogen metabolism; porphyrin and chlorophyll metabolism; oxidative phosphorylation; glycolysis; and starch and sucrose metabolism (fig. S10).

Tentative taxonomic assignments of PZ or DW sequences (top HSPs from NCBI's nonredundant protein database) were also tabulated (fig. S11). As expected, a high percentage of Prochlorococcus-like sequences was found in PZ (~5% of the total), and a greater representation of Deltaproteobacteria-like, Actinobacteria-like and Planctomycete-like sequences were recovered in DW. Unexpectedly, the single most highly represented taxon category in PZ (~21% of all identified sequences in PZ) was derived from viral sequences that were captured in fosmid clones (fig. S11).

Community Genomics and Host-Virus Interactions

Viruses are ubiquitous and abundant components of marine plankton, and influence lateral gene transfer, genetic diversity, and bacterial

mortality in the water column (37–40). The large number of viral DNA sequences in our dataset was unexpected (Fig. 5; fig. S12), because we expected planktonic viruses to pass through our collection filters. Previous studies using a similar approach found only minimal contributions from viral sources (19, 40). The majority of viral DNA we captured in fosmid clone libraries apparently originates from replicating viruses within infected host cells (35). Viral DNA recovery was highest in the photic zone, with cyanophage-like sequences representing 1 to 10% of all fosmid sequences (Fig. 5), and 60 to 80% of total virus sequences there. Below 200 m, viral DNA made up no more than 0.3% of all sequences at each depth. Most photic zone viral sequences shared highest similarity to T7-like and T4-like cyanophage of the Podoviridae and Myoviridae. This is consistent with previous studies (40–42), suggesting a widespread distribution of these phage in the ocean.

Analyses of 1107 fosmid mate pairs provided further insight into the origins of the viral sequences. About 67% of the viruslike clones were most similar to cyanophage on at least one end, and half of these were highly similar to cyanophage at both termini. Many of the cyanophage clones showed apparent synteny with previously sequenced cyanophage genomes (fig. S12). About 11% of the cyanophage paired-ends contained a host-derived cyanophage "signature" gene (43) on one terminus. The frequency and genetic-linkage of phage-encoded (but host-derived) genes we observed, including virus-derived genes involved in photosynthesis (psbA, psbD, hli), phosphate-scavenging genes (phoH, pstS), a cobalamin biosynthesis gene (cobS), and carbon metabolism (transaldolase) supports their widespread distribution in natural viral populations and their probable functional importance to cyanophage replication (43, 44).

If we assume that the cyanophages' DNA was derived from infected host cells in which phage were replicating, the percentage of cyanophage-infected cells was estimated to range between 1 and 12% (35). An apparent cyanophage infection maxima was observed at 70 m, coinciding with the peak virus:host ratio (Fig. 5). Although these estimates are tentative, they are consistent with previously reported ranges of phage-infected picoplankton cells in situ (38, 45).

About 0.5% of all sequences were likely prophage, as inferred from high sequence similarity to phage-related integrases and known prophage genes (35). Paired-end analyses of viral fosmids indicated that ~2.5% may be derived from prophage integrated into a variety of host taxa. A few clones also appear to be derived from temperate siphoviruses, and a number of putative eukaryotic paired-end viral sequences shared highest sequence identity with homologs from herpes viruses, mimiviruses, and algal viruses.
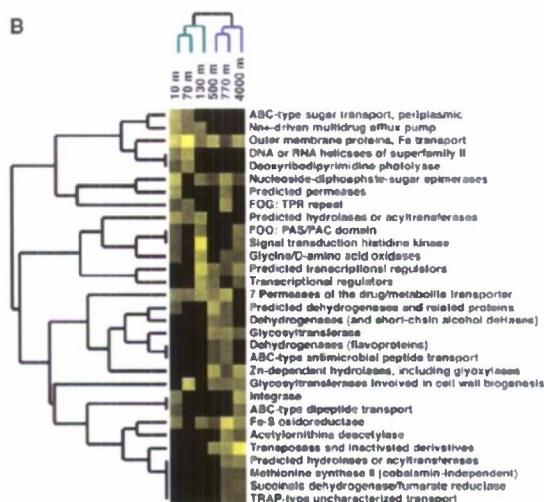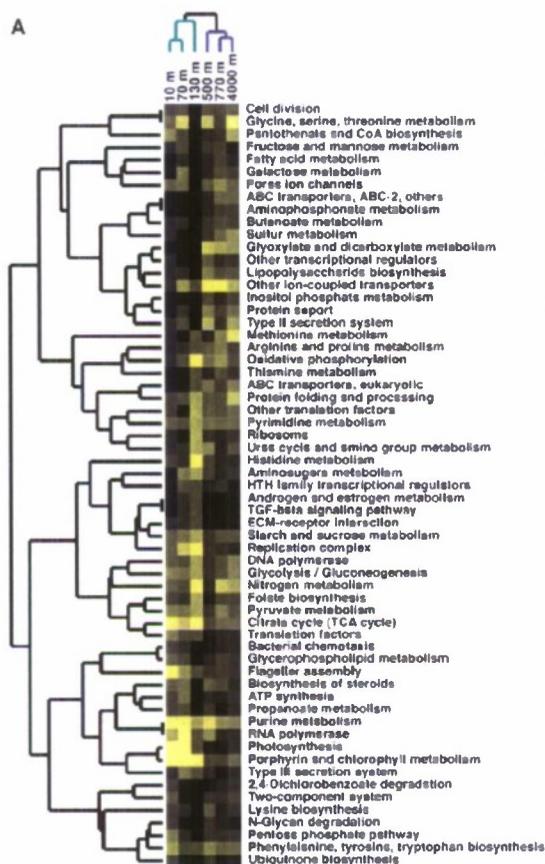
Fig. 4. Cluster analyses of KEGG and COG annotated PZ and DW sequence bins versus depth. Sequence homologs unique to or shared within the photic zone (10, 70, and 130 m) and those unique to or shared in DW (500, 770, and 4000 m) were annotated against the KEGG or COG databases with TBLASTX with an expectation threshold of 1×10^{-5}. Yellow shading is proportional to the percentage of categorized sequences in each category. Cluster analyses of gene categories (left dendrograms) were performed with the Kendall's tau nonparametric distance metric, and the Pearson correlation was used to generate the top dendrograms relating the depth series (33, 34). Dendrograms were displayed by using self-organizing mapping with the Pearson correlation metric (33, 34). Green lines in top dendrograms show PZ sequences, blue lines DW sequences. (**A**) KEGG category representation versus depth. KEGG categories with a standard deviation greater than 0.4 of observed values, having at least two depths $\geq 0.6\%$ of the total KEGG-categorized genes at each depth, are shown. For display purposes, categories >8% in more than two depths are not shown. (**B**) COG category representation versus depth. COG categories with standard deviations greater than 0.2 of observed values, having at least two depths $\geq 0.3\%$ of the total COG-categorized genes at each depth, are shown.

Ecological Implications and Future Prospects

Microbial community sampling along well-characterized depth strata allowed us to identify significant depth-variable trends in gene content and metabolic pathway components of oceanic microbial communities. The gene repertoire of surface waters reflected some of the mechanisms and modes of light-driven processes and primary productivity. Environmentally diagnostic sequences in surface waters included predicted proteins associated with cyanophage, motility, chemotaxis, photosynthesis, proteorhodopsins, photolyases, carotenoid biosynthesis, iron-transport systems, and host restriction-modification systems. The importance of light energy to these communities as reflected in their gene content was obvious. More subtle ecophysiological trends can be seen in iron transport, vitamin synthesis, flagella synthesis and secretion, an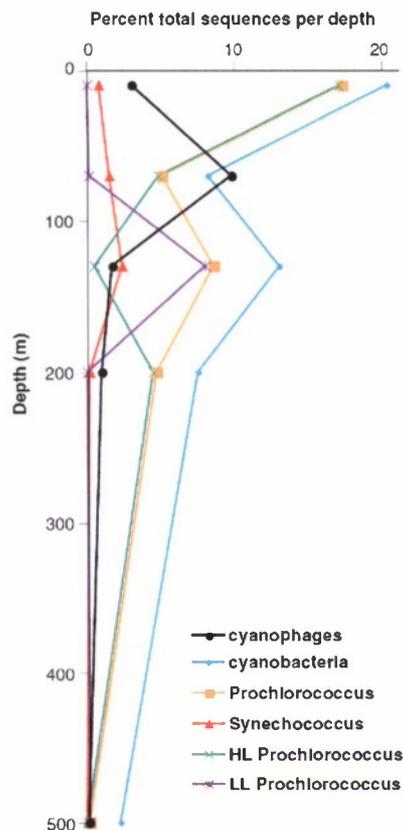d chemotaxis gene distributions. These data support hypotheses about potential adaptive strategies of heterotrophic bacteria in the photic zone that may actively compete for nutrients by swimming toward nutrient-rich particles and algae (46). In contrast to surface-water assemblages, deep-water microbial communities appeared more enriched in transposases, pilus synthesis, protein export, polysaccharide and antibiotic synthesis, the glyoxylate cycle, and urea metabolism gene sequences. The observed enrichment in pilus, polysaccharide, and antibiotic synthesis genes in deeper-water samples suggests a potentially greater role for a surface-attached life style in deeper-water microbial communities. Finally, the apparent enrichment of phage genes and restriction-modification systems observed in the photic zone may indicate a greater role for phage parasites in the more productive upper water column, relative to deeper waters.

At finer scales, sequence distributions we observed also reflected genomic "microvari-ability" along environmental gradients, as evidenced by the partitioning of high- and low-light *Prochlorococcus* ecotype genes observed in different regions of the photic zone (Fig. 5). Higher-order biological interactions were also evident, for example in the negative correlation of cyanophage versus *Prochlorococcus* host gene sequence recovery (Fig. 5). This relation between the abundance of host and cyanophage DNA probably reflects specific mechanisms of cyanophage replication in situ. These host-parasite sequence correlations we saw demonstrate the potential for observing community-level interspecies interactions through environmental genomic datasets.

Obviously, the abundance of specific taxa will greatly influence the gene distributions observed, as we saw, for example, in *Prochlorococcus* gene distribution in the photic zone. Gene sequence distributions can reflect more than just relative abundance of specific taxa, however,

Fig. 5. Cyanophage and cyanobacteria distributions in microbial community DNA. The percentage of total sequences derived from cyanophage, total cyanobacteria, total *Prochlorococcus* spp., high-light *Prochlorococcus*, low-light *Prochlorococcus* spp., or *Synechococcus* spp., from each depth. Taxa were tentatively assigned according to the origin of top HSPs in TBLASTX searches, followed by subsequent manual inspection and curation.

Percent total sequences per depth

- ● cyanophages
- ◆ cyanobacteria
- ■ Prochlorococcus
- ▲ Synechococcus
- ✛ HL Prochlorococcus
- ✱ LL Prochlorococcus

enabled ecological studies matures, it should become possible to model microbial community genomic, temporal, and spatial variability with other environmental features. Significant future attention will no doubt focus on interpreting the complex interplay between genes, organisms, communities and the environment, as well as the properties revealed that regulate global biogeochemical cycles. Future efforts in this area will advance our general perspective on microbial ecology and evolution and elucidate the biological dynamics that mediate the flux of matter and energy in the world's oceans.

References and Notes

1. E. Forbes, in *Physical Atlas of Natural Phenomena*, A. K. Johnston, Ed. (William Blackwood & Sons, London and Edinburgh, 1856).
2. P. W. Hochachka, G. N. Somero, *Biochemical Adaptation* (Princeton Univ. Press, Princeton, NJ, 1984), pp. 450–495.
3. G. Rocap et al., *Nature* **424**, 1042 (2003).
4. Z. I. Johnson et al., manuscript submitted.
5. N. J. West et al., *Microbiology* **147**, 173† (2001).
6. A. A. Yayanos, *Annu. Rev. Microbiol.* **49**, 777 (1995).
7. N. R. Pace, *Science* **276**, 734 (1997).
8. M. S. Rappé, S. J. Giovannoni, *Annu. Rev. Microbiol.* **57**, 369 (2003).
9. M. T. Suzuki et al., *Microb Ecol* (2005).
10. O. Béjà et al., *Environ. Microbiol.* **2**, S16 (2000).
11. R. M. Morris, M. S. Rappé, E. Urbach, S. A. Connon, S. J. Giovannoni, *Appl. Environ. Microbiol.* **70**, 2836 (2004).
12. S. Y. Moon-van der Staay et al., *Nature* **409**, 607 (2001).
13. J. A. Fuhrman, K. McCallum, A. A. Davis, *Nature* **356**, 148 (1992).
14. E. F. Delong, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 5685 (1992).
15. M. B. Karner et al., *Nature* **409**, 507 (2001).
16. J. Handelsman, *Microbiol. Mol. Biol. Rev.* **68**, 669 (2004).
17. E. F. Delong, *Nat Rev Microbiol* (2005).
18. G. W. Tyson et al., *Nature* **428**, 37 (2004).
19. J. C. Venter et al., *Science* **304**, 66 (2004).
20. S. G. Tringe et al., *Science* **308**, 554 (2005).
21. S. J. Hallam et al., *Science* **305**, 1457 (2004).
22. D. M. Karl, R. Lukas, *Deep-Sea Res. II* **43**, 129 (1996).
23. D. M. Karl et al., *Deep-Sea Res. II* **48**, 1449 (2001).
24. R. M. Letelier et al., *Limnol. Oceanogr.* **49**, 508 (2004).
25. E. F. Delong et al., *Appl. Environ. Microbiol.* **65**, 5554 (1999).
26. A. Pernthaler et al., *Appl. Environ. Microbiol.* **68**, 661 (2002).
27. N. U. Frigaard et al., *Nature*, in press.
28. S. M. Barns, C. F. Delwiche, J. D. Palmer, N. R. Pace, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 9188 (1996).
29. M. Kanehisa et al., *Nucleic Acids Res.* **32**, D277 (2004).
30. R. L. Tatusov et al., *BMC Bioinformatics* **4**, 41 (2003).
31. R. Overbeek et al., *Nucleic Acids Res.* **33**, 569† (2005).
32. S. F. Altschul et al., *J. Mol. Biol.* **215**, 403 (1990).
33. M. J. L. de Hoon et al., *Bioinformatics* **12**, 1453 (2004).
34. M. B. Eisen, P. T. Spellman, P. O. Brown, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 14863 (1998).
35. Materials and methods are available as supporting material on *Science* Online.
36. K. O. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **33**, D501 (2005).
37. M. G. Weinbauer, *FEMS Microbiol. Rev.* **28**, 127 (2004).

Some depth-specific gene distributions we observed [e.g., transposases found predominantly at greater depths (Fig. 4B; fig. S8)], appear to originate from a wide variety of gene families and genomic sources. These gene distributional patterns seem more indicative of habitat-specific genetic or physiological trends that have spread through different members of the community. Community gene distributions and stoichiometries are differentially propagated by vertical and horizontal genetic mechanisms, dynamic physiological responses, or interspecies interactions like competition. The overrepresentation of certain sequence types may sometimes reflect their horizontal transmission and propagation within a given community. In our datasets, the relative abundance of cyanobacteria-like *psbA*, *psbD*, and transaldolase genes were largely a consequence of their horizontal transfer and subsequent amplification in the viruses that were captured in our samples. In contrast, the increase of transposases from 500 to 4000 m, regardless of community composition, reflected a different mode of gene propagation, likely related to the slower growth, lower

productivity, and lower effective population sizes of deep-sea microbial communities. In future comparative studies, similar deviations in environmental gene stoichiometries might be expected to provide even further insight into habitat-specific modes and mechanisms of gene propagation, distribution, and mobility (*27*, *47*). These "gene ecologies" could readily be mapped directly on organismal distributions and interactions, environmental variability, and taxonomic distributions.

The study of environmental adaptation and variability is not new, but our technical capabilities for identifying and tracking sequences, genes, and metabolic pathways in microbial communities is. The study of gene ecology and its relation to community metabolism, interspecies interactions, and habitat-specific signatures is nascent. More extensive sequencing efforts are certainly required to more thoroughly describe natural microbial communities. Additionally, more concerted efforts to integrate these new data into studies of oceanographic, biogeochemical, and environmental processes are necessary (*48*). As the scope and scale of genome-

38. J. Waterbury, F. Valois, *Appl. Environ. Microbiol.* **59**, 3393 (1993).
39. M. B. Sullivan *et al.*, *Nature* **424**, 1047 (2003).
40. R. A. Edwards, F. Rohwer, *Nat. Rev. Microbiol.* **3**, 504 (2005).
41. J. Filee, F. Tetart, C. A. Suttle, H. M. Krisch, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12471 (2005).
42. M. Breitbart, J. H. Miyake, F. Rohwer, *FEMS Microbiol. Lett.* **236**, 249 (2004).
43. M. B. Sullivan *et al.*, *PLoS Biol.* **3**, e144 (2005).
44. D. Lindell *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11013 (2004).
45. M. G. Weinbauer, I. Brettar, M. G. Hofle, *Limnol. Oceanogr.* **48**, 1457 (2003).
46. F. Azam, *Science* **280**, 694 (1998).
47. C. R. Woese, *Microbiol. Mol. Biol. Rev.* **68**, 173 (2004).
48. E. F. DeLong, D. M. Karl, *Nature* **437**, 336 (2005).
49. (http://hahana.soest.hawaii.edu/hot/parameters.html).
50. (http://hahana.soest.hawaii.edu/hot/hot-dogs).
51. this work was supported by a grant from the Gordon and Betty Moore Foundation, NSF grants MCB-0084211, MCB-0348001, and MCB-0509923 to E.F.D., and OCE-0326616 to O.M.K., and sequencing support from the U.S. Department of Energy Microbial Genomics Program. We thank Dennis Ryan for help with scripting and sequence analyses, the officers and crew of the R/V Ka'imikai-O-Kanaloa and the HOT team for assistance at sea, and L. Fujieki for help with oceanographic data display. J. Chapman, A. Salamov, and P. Richardson of the DOE Joint Genome Institute provided advice and assistance in DNA sequencing and analyses. Sequences have been deposited in GenBank with accession numbers DU731018-DU796676 and DU800850-DU800864 corresponding to fosmid end sequences, and accession numbers DQ300508-DQ300926 corresponding to SSU rRNA gene sequences.

Pairing and Phase Separation in a Polarized Fermi Gas

Guthrie B. Partridge, Wenhui Li, Ramsey I. Kamar, Yean-an Liao, Randall G. Hulet*

We report the observation of pairing in a gas of atomic fermions with unequal numbers of two components. Beyond a critical polarization, the gas separates into a phase that is consistent with a superfluid paired core surrounded by a shell of normal unpaired fermions. The critical polarization diminishes with decreasing attractive interaction. For near-zero polarization, we measured the parameter $\beta = -0.54 \pm 0.05$, describing the universal energy of a strongly interacting paired Fermi gas, and found good agreement with recent theory. These results are relevant to predictions of exotic new phases of quark matter and of strongly magnetized superconductors.

Fermion pairing is the essential ingredient in the Bardeen, Cooper, and Schrieffer (BCS) theory of superconductivity. In conventional superconductors, the chemical potentials of the two spin states are equal. There has been great interest, however, in the consequences of mismatched chemical potentials that may arise in several important situations, including, for example, magnetized superconductors [1–3] and cold dense quark matter at the core of neutron stars [4]. A chemical potential imbalance may be produced by several mechanisms, including magnetization in the case of superconductors, mass asymmetry, or unequal numbers. Pairing is qualitatively altered by the Fermi energy mismatch, and there has been considerable speculation regarding the nature and relative stability of various proposed exotic phases. In the Fulde-Ferrel-Larkin-Ovchinnikov (FFLO) phase [2, 3], pairs possess a nonzero center-of-mass momentum that breaks translational invariance, whereas the Sarma [1], or the breached pair [5], phase is speculated to have gapless excitations. A mixed phase has also been proposed [6–8] in

which regions of a paired BCS superfluid are surrounded by an unpaired normal phase. Little is known experimentally, however, because of the difficulty in creating magnetized superconductors. Initial evidence for an FFLO phase in a heavy-fermion superconductor has only recently been reported [9, 10]. Opportunities for experimental investigation of exotic pairing states have expanded dramatically with the recent realization of the Bose-Einstein condensate (BEC)–BCS crossover in a two spin state mixture of ultracold atomic gases. Recent experiments have demonstrated both superfluidity [11–13] and pairing [14–17] in atomic Fermi gases. We report the observation of pairing in a polarized gas of ^{6}Li atoms. Above an interaction-dependent critical polarization, we observed a phase separation that is consistent with a uniformly paired superfluid core surrounded by an unpaired shell of the excess spin state. Below the critical polarization, the spatial size of the gas was in agreement with expectations for a universal, strongly interacting paired Fermi gas.

Our methods for producing a degenerate gas of fermionic ^{6}Li atoms [18, 19] and the realization of the BEC-BCS crossover at a Feshbach resonance [17] have been described previously [20]. An incoherent spin mixture of the $F = \frac{1}{2}$, $m_F = \frac{1}{2}$ (state $|1\rangle$) and the $F = \frac{1}{2}$,

$m_F = -\frac{1}{2}$ (state $|2\rangle$) sublevels (where F is the total spin quantum number and m_F is its projection) is created by radio frequency (rf) sweeps, where the relative number of the two states can be controlled by the rf power [20]. The spin mixture is created at a magnetic field of 754 G, which is within the broad Feshbach resonance located near 834 G [21, 22]. The spin mixture is evaporatively cooled by reducing the depth of the optical trap that confines it, and the magnetic field is ramped adiabatically to a desired field within the crossover region. States $|1\rangle$ and $|2\rangle$ are sequentially and independently imaged in the trap by absorption [20]. Analysis of these images provides measurement of N_i and polarization $P = (N_1 - N_2)/(N_1 + N_2)$, where N_i is the number of atoms in state $|i\rangle$. We express the Fermi temperature, T_F, in terms of the majority spin state, state $|1\rangle$, as $k_B T_F = \hbar\bar\omega (6N_1)^{1/3}$, where $\bar\omega = 2\pi (v_r^2 v_z)^{1/3}$ is the mean harmonic frequency of the cylindrically symmetric confining potential with radial and axial frequencies v_r and v_z, respectively. For $P \approx 0$, we find that $N_1 \approx N_2 \approx 10^5$, giving $T_F \approx 400$ nK for our trap frequencies. Because of decreasing evaporation efficiency with increasing polarization, there is a correlation between P and total atom number (fig. S1).

For fields on the low-field (BEC) side of resonance, real two-body bound states exist, and molecules are readily formed by three-body recombination. For the case of $P = 0$, a molecular Bose-Einstein condensate (MBEC) is observed to form with no detectable thermal molecules [17]. On the basis of an estimated MBEC condensate fraction of >90%, we place an upper limit on the temperature $T < 0.1 T_F$ at a field of 754 G [17]. However, the gas is expected to be cooled further during the adiabatic ramp for final fields greater than 754 G [17]. By using similar experimental methods, we previously measured the order parameter of the gas in the BCS regime and found good agreement with $T = 0$ BCS theory [17], indicating that the gas was well below the critical temperature for pairing.

Images of states $|1\rangle$ and $|2\rangle$ at a field of 830 G are shown (Fig. 1) for relative numbers

Department of Physics and Astronomy and Rice Quantum Institute, Rice University, Houston, TX 77251, USA.

*To whom correspondence should be addressed. E-mail: randy@rice.edu

Corrected 30 January 2006. This file now includes the supplemental figures.

Supporting Online Material for

Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior

Edward F. DeLong,* Christina M. Preston, Tracy Mincer, Virginia Rich,
Steven J. Hallam, Niels-Ulrik Frigaard, Asuncion Martinez, Matthew B. Sullivan,
Robert Edwards, Beltran Rodriguez Brito, Sallie W. Chisholm, David M. Karl

*To whom correspondence should be addressed. E-mail: delong@mit.edu

This PDF file includes

> Materials and Methods
> Figures S1 to S12
> Tables S1
> References and Notes

Supporting Online Material

Comparative genomics of microbial communities in the ocean's interior

Edward F. DeLong, Christina M. Preston, Tracy Mincer, Virginia Rich, Steven J. Hallam, Niels-Ulrik Frigaard, Asuncion Martinez, Matt Sullivan, Robert Edwards, Beltran Rodriguez Brito, Sallie W. Chisholm and David M. Karl

Materials and methods
Sampling and library preparation

Seawater from selected depths (Table 1, main text) were collected at the Hawaii Ocean Time Series (HOT) station ALOHA (22.45°N, 158°W). Multiple hydrocasts for sampling and measurement used a Conductivity, Temperature, Depth (CTD) rosette water sampler equipped with 24, 12 l polyvinyl chloride sample bottles aboard the R/V Ka'imikai-o-Kanaloa. Sample depths were selected based on physical (temperature, pressure), chemical (salinity, dissolved oxygen) and biotic (chlorophyll fluorescence) characteristics in real time from CTD data. Seawater samples from seven depths were collected from multiple hydrocasts using a CTD system equipped with 24, 12 l polyvinyl chloride sample bottles. Samples from 10 m to 500 m were collected on October 2002, and those from depths of 770 m and 4000 m on December 2003. The seawater was pre-filtered in line through a 47 mm Whatman glass fiber GFA filter (Millipore, Bedford, MA) before final collection onto 0.22 m Sterivex-GV filter (Millipore) using a Masterflex peristaltic pump (Cole Parmer Instrument Company, Vernon Hills, IL). From one to four Sterivex filters were used at each depth, depending on the volume of sample filtered (Table 1, main text). The glass fiber GFA prefilters were replaced after each 20 liters filtered. After seawater collection, the Sterivex filters were covered with 0.5 ml of lysis buffer (50 mM Tris•Hcl, pH 8.3, containing 40mM EDTA and 0.75M sucrose) and frozen at –80°C. Samples were transported back to the laboratory on dry ice, and stored at –80°C until DNA extraction. From one to four filters were extracted and the total DNA pooled for subsequent fosmid library construction from each depth.

DNA extraction and fosmid library construction was conducted as previously described, with minor modifications (*S1*). Briefly, a solution of proteinase K in sterile water was added to a final concentration of 0.5 mg·ml^{-1} into the Sterivex filter cartridge (Fisher, Fairlawn, NJ), followed by addition of SDS to a final concentration of 1% (Sigma, St Louis, MO). The filter cartridges were sealed and incubated at 55°C for 20 minutes, followed by further incubation at 70°C for 5 minutes to further promote cell lysis. The lysate was remove from the filter cartridge, and nucleic acids were extracted twice with phenol:chloroform:IAA (25:24:1, Sigma) and once with chloroform:isoamyl alcohol (24:1, Sigma). After concentration of the crude nucleic acids by spin dialysis using a Centricon 100 filter (Millipore), the DNA was further purified by CsCl buoyant equilibrium centrifugation, as previously

described.

Environmental DNA was cloned using the CopyControl™ Fosmid Library Production Kit (Epicentre, Madison, WI) following the manufacturer's protocol. Briefly, purified DNA was end repaired according to the manufacturer instructions, and size-fractionated by pulsed field gel electrophoresis (PFGE) on a CHEF-DR-II system (Bio-Rad, Hercules, CA) using a 1% SeaPlaque GTG agarose (Cambrex, Baltimore, MD) under the following conditions: 12°C, 6 V·cm⁻¹ for 16 hrs and 20-40 s pulse time in 1X TAE (40 mM Tris-acetate, 1 mM EDTA, pH 8.0) buffer. The gel was subsequently stained with SYBR gold (Molecular Probes, Eugene, OR) and viewed on a Dark Reader transilluminator (Clare Chemical Research, Dolores, CO). Gel regions containing genomic DNA in 40-50 kbp regions were excised. The end-repaired and size-selected DNA from gel slices was recovered by gelase treatment and concentrated and washed three time with an equal volume of TE buffer on a Centricon 100 (Millipore). DNA was ligated into the CopyControl™ pCC1FOS™ vector, packaged *in vitro* using MaxPlax™ Lambda Packaging Extracts, and transduced into *E. coli* EPI300™ according to the manufacturer's instructions (Epicentre, Madison, WI).

Bacterial and archaeal small subunit rRNA screens

To survey bacterial small subunit rRNA diversity, *E.coli* host chromosomal DNA was first removed from the fosmid library clone pools (12,000-18,000 individually grown and subsequently pooled clones for each depth) by two rounds of CsCl density gradient centrifugation (*S2*). CsCl purified clone pool DNA from each library was nuclease treated using Plasmid Safe exonuclease™ (Epicentre, Madison, WI) following the manufacturers recommendations. Aliquots (250-300 ng) of the *E. coli*-free, pooled library DNA was subsequently used as template in the downstream bacterial SSU rRNA gene amplification. Reaction mixtures for amplification of SSU rRNA gene sequences consisted of the following: 250 ng template DNA, 0.2 mM dNTPs each, 0.5 uM each forward primer 27F (5′ AGAGTTTGATCMTGGCTCAG) and reverse primer 1492R (5′ TACGGYTACCTTGTTACGACTT), 5 U "Easy A" thermostable proofreading polymerase (Stratagene, La Jolla, CA), in a total of 50 L reaction volume. Polyermase chain reaction cycles were as follows: an initial denaturation step of 2 minutes at 94°C; 30 seconds at 94°C, 30 seconds at 55°C, and 90 seconds at 72°C for a total of 15 amplification cycles. Reconditioning PCR was carried out to reduce heteroduplex formation (*S3*) as follows: initial reaction products were diluted ten-fold, and re-amplified using parameters identical to the above, except that only three thermal cycles were performed.

Triplicate PCR reactions were pooled and cloned using a TOPO TA cloning kit (Invitrogen, Carlsbad, CA). From each library 192 clones were picked and plasmid DNA was purified with an automated DNA purification system (AutoGen, Holliston, MA) using parameters recommended for high-copy plasmid DNA. Clone inserts were sequenced using primers 27F and

907R (5' CCGTCAATTCMTTTRAGTTT) with ABI PRISM BigDye Terminator v3.1 cycle sequencing kit (Applied Biosystems, Foster City, CA). A range of 41-81 bacterial SSU rRNA gene sequences were sequenced from each library. Sequences were subsequently aligned to a database in ARB (version 2.5b) (S4) and assigned to the nearest taxonomic affiliation to environmental and cultivated isolates. In total 351 bacterial 16S rRNA gene sequences were analyzed and assigned to a taxonomic bin. Sequences were analyzed for chimeras with the Bellerophon server (S5) using the Huber-Hugenholtz correction and a 300-bp window size.

Archaeal small subunit rRNA-containing fosmids were identified directly by colony hybridization (S2). All fosmid clones from each library were arrayed onto positively charged nylon membranes using a Genetix QPix2Xt automated robot and processed according to manufacturer's recommendations (Genetix, Hampshire, UK). Hybridization was carried out at 60°C with PCR-generated, non-isotopically labeled archaeal rRNA-targeted probes using AlkPhos Direct Labeling and ECF Chemifluorescent Detection kits (Amersham Biosciences, Piscataway, NJ). Positive clones were visualized on a Fuji FLA-5100 fluorescent image analyzer (Fuji Life Science, USA), and all hybridization positive archaeal rRNA-containing fosmids were picked from each library, and sequenced as described above, using Ar20F (5' TTCCGGTTGATCCYGCCRG) and U1390R (5' GACGGGCGGTGTGTRC) PCR amplification primers, and) and internal sequencing primers U530F (5' GTGCCAGCMGCCGCGG) and Ar958R (5' YCCGGCGTTGAMTCCAATT. Reagent concentrations were identical to the bacterial amplifications above using forward and reverse primers Ar20F (5' TTCCGGTTGATCCYGCCRG) and U1390R (5' GACGGGCGGTGTGTRC) and template concentration ranged from 50100 ng/reaction. Cycling parameters were as follows: an initial denaturation step of 2 minutes at 94°C; 30 seconds at 94°C, 30 seconds at 60°C, 90 seconds at 72°C, for a total of 25 amplification cycles. PCR products were purified using a Montage 96-well vacuum system (Millipore) according to the manufacturers protocol. PCR amplicons from archaeal rRNA clones from each library were sequenced directly, using Ar20F (5' TTCCGGTTGATCCYGCCRG) and U1390R (5' GACGGGCGGTGTGTRC) PCR amplification primers, and internal sequencing primers U530F (5' GTGCCAGCMGCCGCGG) and Ar958R (5' YCCGGCGTTGAMTCCAATT) using methods as stated above, yielding on average 1200 bp unambiguous DNA sequence. Phylogenetic trees were generated using ARB (version 2.5b) (S4) and PAUP 4.0 (Sinauer Associates, Sunderland. MA) using a neighbor-joining method with 1000 bootstrap replicates.

Cluster analysis of cumulative bitscore comparisons for pairwise depth comparisons

Blast searches (TBLASTX) of all sequences from one depth versus all from every other, were used to estimate cumulative protein sequence differences existing in all possible depth comparisons. The bitscores of the top high-scoring pairs (HSPs) from every single sequence from one depth versus

another were summed, to yield a cumulative pairwise bitscore value. The pairwise cumulative bitscore values from all possible pairwise sequence comparisons of one depth versus another were then used to construct a distance martrix as follows: Cumulative pairwise bitscore values were normalized by dividing each by: a) the cumulative bitscore value derived from the sum of bitscore values in self-self TBLASTX comparisons; and b) the total number of HSPs in any given comparison. The normalized, cumulative bitscore "similarities" were then each subtracted from one to derive pseudo-distance values, and construct a distance matrix. Distance matrices were analyzed using Phylip, v 3.61 by neighbor joining analysis (S6). To compare our datasets with recently reported shotgun data from the Sargasso Sea, 10,000 sequences from each Sargasso Sea sample bin were randomly selected (to normalize to our target-query size of 10,000 sequences), and identical analyses were conducted (fig. S8). Clustering patterns were consistent with the depth of origin and filtered size fraction of each sample (fig. S7), with sub-clustering differentiating the Pacific from Atlantic ocean photic zone datasets.

Analysis of photic zone and deep-water unique sequence bins
To identify sequences characteristic of either photic zone or deep water microbial assemblages, we conducted reciprocal BLAST (S7) comparisons between each individual photic zone dataset (10 m, 70 m, and 130 m) and a pooled deep-water dataset (e.g., all 500 m, 770 m, 4000 m combined). The annotation tool in Pymood (Allometra, Davis, CA) software was used to parse and identify shared sequence bins. All sequences unique to, or shared between, any of the photic zone samples to the exclusion of all deep-water samples identified in TBLASTX searches (expectation cutoff of $1x10^{-5}$) were tabulated and pooled using the Pymood annotation tool. These are the photic zone unique sequences (PZ in Fig. 4) were then analyzed by comparison to well curated databases (fig S7A). Similarly, each individual deep-water sequence dataset (500m, 770m, 4000m) were reciprocally compared to one another, and the pooled photic zone dataset (all 10 m, 70 m, 130 m sequences combined). Deep-water unique sequences (DW) were identified in TBLASTX (expectation cutoff of $1x10^{-5}$), and similarly pooled for subsequent analyses (as in fig. S7b). Likewise, all those "core" sequences that were present in and shared significant similarity (e-values < $1x10^{-5}$) in all six data sets (10 m, 70 m, 130 m, 500 m, 770 m, 4000 m), were identified and pooled as described above. (In these analyses, the transition depth 200 m between the photic zone and deeper waters was not included).

Once identified, the PZ, and DW, and "core" sequence bins were each compared to the KEGG, COG, NCBI non-redundant protein, and Sargasso sequence databases using BLASTX, TBLASTX, or BLASTN (S7). Data were parsed and ranked according to top HSPs and the functional annotations, expectation values, and taxonomic origins. Sequences associated with specific functional, COG, or taxonomic categories at specified expectation value thresholds (Fig. 3, figs. S7, S9-S12) were then plotted as a function of

their fractional representation.

Statistical analysis of protein category representation

Each of the protein sequence collections within specific categories [based on comparison to KEGG pathways, COG gene families, or SEED Subsystems (S8)], were analyzed to identify protein categories statistically more likely to be found in any one sample, versus any another, or between PZ and DW sequence bins. For speed and reproducibility we adopted a bootstrap sampling method (S9). First, the difference between median instances of any KEGG subsystem, KEGG, or COG category in the dataset was calculated: 10,000 proteins were sampled from each sequence bin, and for each pairwise comparison the difference in the number of subsystems, pathways, or gene families was calculated. This was repeated 20,000 times, and the median differences calculated. To identify those median differences that were statistically unlikely to have occurred by random chance, this process was repeated for each pairwise bin, except the 10,000 proteins were sampled from a bin at random. Again, 20,000 repeat calculations were performed, and the data organized from least difference to most difference. The confidence intervals were provided by the appropriate percentile differences, that is for 99% confidence intervals the 1% limit was provided by the 200th difference and the 99% limit was provided by the 19,801st difference from the ordered list. If the difference of medians was outside these limits, the subsystem, pathway, or protein family was considered to have a statistically significantly different distribution in one versus another dataset.

This method allowed for rapid calculation of the differences between subsystems, pathways, or protein families, and does not require a normal distribution of the data. Furthermore, the sample size and repeat size can be modified to approximate the size of the datasets involved in the analysis.

Viral sequence analyses

Shotgun sequences were ranked by the expectation values of their top scoring HSPs from blast searches using blastx (S7, S8). Less stringent expectation cutoff values ($<10^{-3}$) were initially used, due to the significant sequence divergence of viral genes, to identify potential virus sequences.

To estimate the number of phage genomes integrated into cellular hosts (i.e., prophage), the number of sequences with top scoring HSPs to known temperate phages (e.g., lambdoid siphoviruses), prophages and phage-related integrase genes were tabulated. Additionally, paired end analyses using 1,107 viral fosmid sequence mate pairs were conducted using relatively stringent blastx criteria (e-values better than 10^{-8} for phage and 10^{-10} for cellular hits). In this analysis, fosmids were interpreted to be derived from prophages when one end was similar to known temperate phages and the other end was similar to a cellular gene. Fosmids with both

termini similar to known temperate phages were binned as temperate phages, or with both ends similar to herpes viruses, binned as herpes viruses.

Our sampling filter fractionation procedures targeted cells and not free phage. Nevertheless, a large proportion of fosmid ends were derived from lytic phage DNA. The lytic virus DNA in fosmid clone libraries has two possible origins: intracellular phage DNA recovered from infected cells, or free phage particles that adhered to particulate material on the collection filters. Available data suggest the majority of cloned phage from photic zone samples originated from infected cells : First, approximately the same number of cells were collected at each depth (Table 1), so enhanced phage recovery due to a putative increased particle loading at different depths (and therefore increased coincident phage adhesion), does not explain our results (Fig 5).

Second, ratios of free phage particles to bacterioplankton average about 10:1 in marine plankton, and are relatively constant with depth (*S9, S10*). Hence, variation in free phage with depth (and hence variable depth recoveries), also does not explain our results well. Given the above considerations, it appears likely that a large proportion of recovered phage in our libraries was derived from virus-infected cells, and not free phage particles that adhered to particulate material on the filters.

The percentage of cyanophage-infected cells in our samples was approximated as described below, assuming the cyanophage sequences in our samples reflected phage in the process of infecting host cells. The average T7/T4 like cyanophage genome is about 4% of that of a typical (~2 Mbp) cyanobacterium. This translates to viral genome:host genome ratios ranging from 0.5:1 to 2.5:1 in the photic zone libraries. Since the average burst size is about 20-80 viruses/cell, we can estimate from virus sequence recovery at each depth that the percentage of infected cyanobacteria in the samples ranged from 1 to 12%, with the maximum occurring at 70 m where the virus:host ratio was maximal

Sequence characterization within and between depths

For taxonomic binning (Fig. 1, main text), BLASTX was used to compare the set of all predicted protein sequences against the NCBI nonredundant protein database, using an expectation value cut-off of $< 10^{-60}$. Top BLAST HSPs in this bin were tabulated according to the NCBI taxonomic identifier for each sequence.

Sequences were compared to the KEGG database using BLASTX (*S7*). Blast results were tabulated and the percentage of sequences within each KEGG pathway was calculated for each depth interval. Cluster analysis and "heat maps" were generated using Cluster 3.0 (*S11*) using the C Clustering Library version 1.30 (*S12*) and Java TreeView (http://jtreeview.sourceforge.net).

For COG assignments at each depth interval, open reading frames (orfs)

were identified using automated genome annotation software fgenesb (Softberry, Mount Kisco, NY). Identified orfs from each sequence were then compared to the COG database using blastp *(S7)* searches with an expectation value cut-off of $\leq e^{-5}$. Results were tabulated, and used to determine the percentage of sequences contained in each COG category at each of the seven depth intervals. The following threshold criteria were used in determining which COGS were displayed groups in the cross-depth "heatmaps": COGs comprising >0.2% of the total COG counts at the given depth interval, and > 3-fold change difference between at least one other depth interval, across all depths compared.

References

S1.   M. T. Suzuki *et al.*, *Microb. Ecol.* **48**, 473 (2004).

S2.   J. Sambrook, I. Russell, in *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY., 2001), pp. 7.27-27.45.

S3.   J. R. Thompson, L. A. Marcelino, *Nucleic Acids Res.* **30**, 2083 (2002).

S4.   W. Ludwig *et al.*, *Nucleic Acids Res.* **32**, 1363 (2004).

S5.   T. Huber, G. Faulkner, P. Hugenholtz, *Bioinformatics* **20**, 2317 (2004).

S6.   J. Felsenstein, Distributed by the author. Department of Genome Sciences, University of Washington, Seattle (2004).

S7.   S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J Mol Biol* **215**,403 (1990).

S8.   R. Overbeek *et al.*, *Nucleic Acids Res.* **33**, 5691 (2005).

S9.   M. G. Weinbauer, I. Brettar, M. G. Hofle, *Limnol Oceanogr* **48**, 1457 (2003).

S10. J. H. Paul, M. B. Sullivan, A. M. Segall, F. Rohwer, *Comp Biochem Physiol B Biochem Mol Biol* **133**, 463 (2002).

S11. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863–14868 (1998).

S12. M. J. L. de Hoon, S. Imoto, J. Nolanand, S. Miyano, *Bioinformatics* **12**, 1453(2004).

S13. J. C. Venter *et al.*, *Science* **304**, 66 (2004).

CTD Temperature [Deg C]

Fluorometric Chlorophyll a [µg/l]

Bottle Oxygen [µmol/kg]
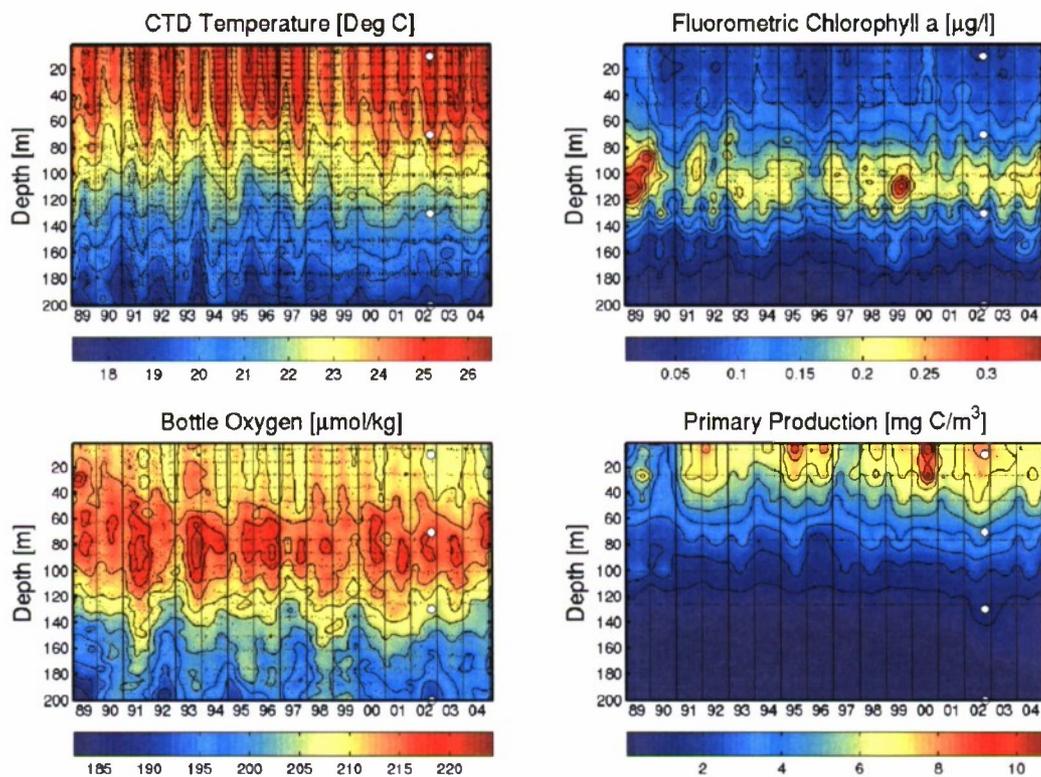
Primary Production [mg C/m³]

fig S1. Contour plots of upper water column temperature, chlorophyll and dissolved oxygen concentrations and rates of primary photoautotrophic production at Station ALOHA (22?45′N, 158?W) for the period 1989-2004. White dots represent the positions of the four upper ocean (10, 70, 130 and 200 m) water samples, collected in October 2002, that were analyzed in this study.
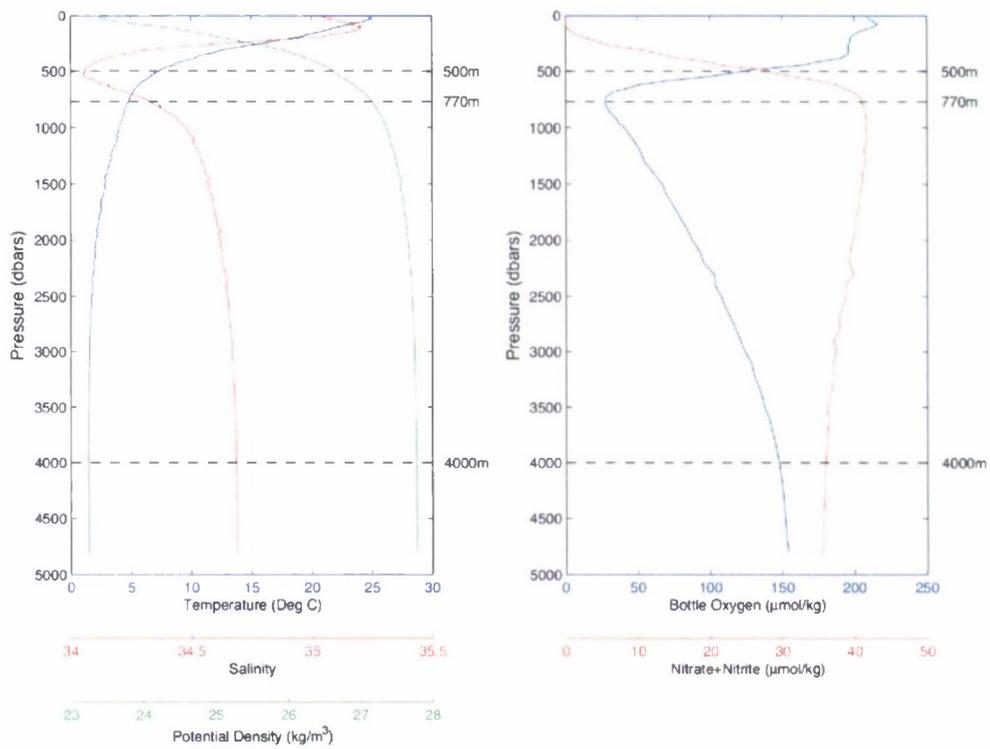
fig S2. Vertical profiles of relevant physical and chemical properties at Station ALOHA (22 deg 45'N, 158 deg W). Shown on the left are temperature, salinity and water density and shown on right are dissolved oxygen and nitrate + nitrite, all from the sea surface to the sea bed at 4750 m. The dashed lines show the positions of the three deep water samples (S00, 770 and 4000 m) analyzed in this study.
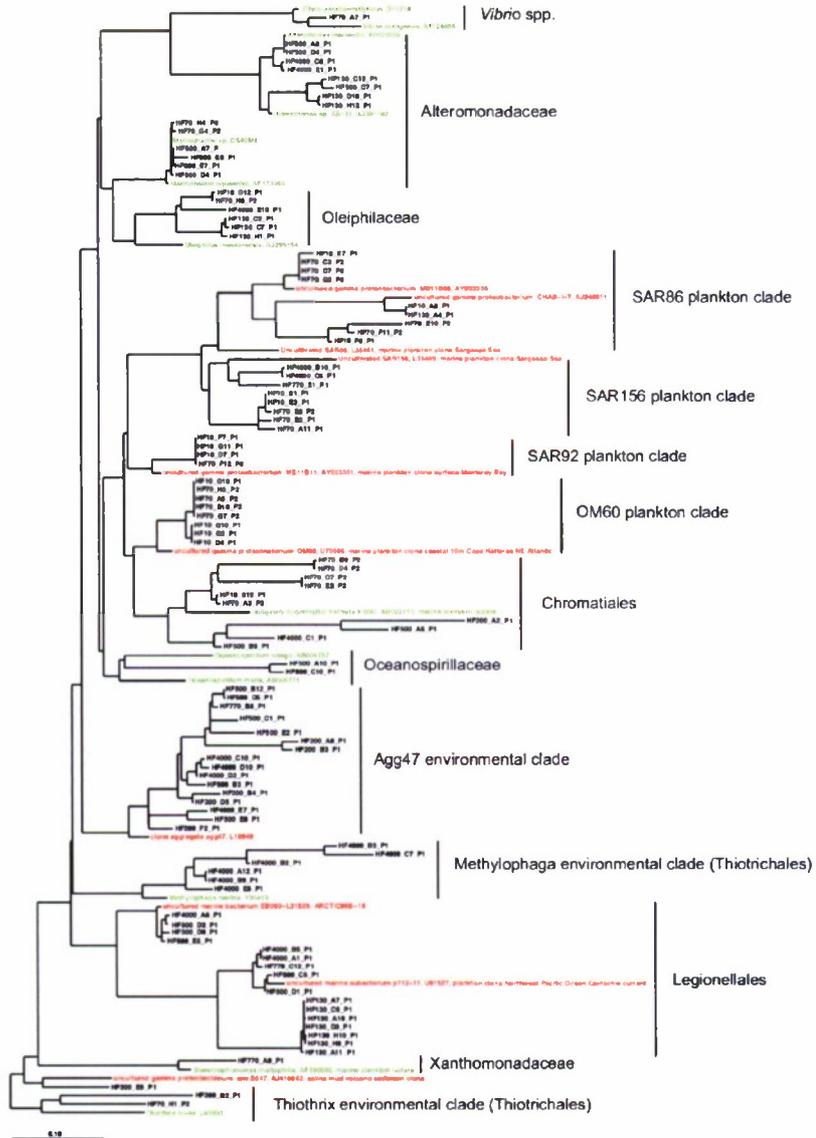
fig. S3. Bacterial SSU rRNA sequences recovered from each depth. Phylogenetic placement of rRNA sequences contained on fosmid clones recovered from each depth. Sequences (600–1300 bp) were aligned using ARB, and phylogenetic placement estimated by parsimony analyses (4). fig. S3A, Phylogenetic positions of Gammaproteobacteria-like fosmids based on rRNA sequence.
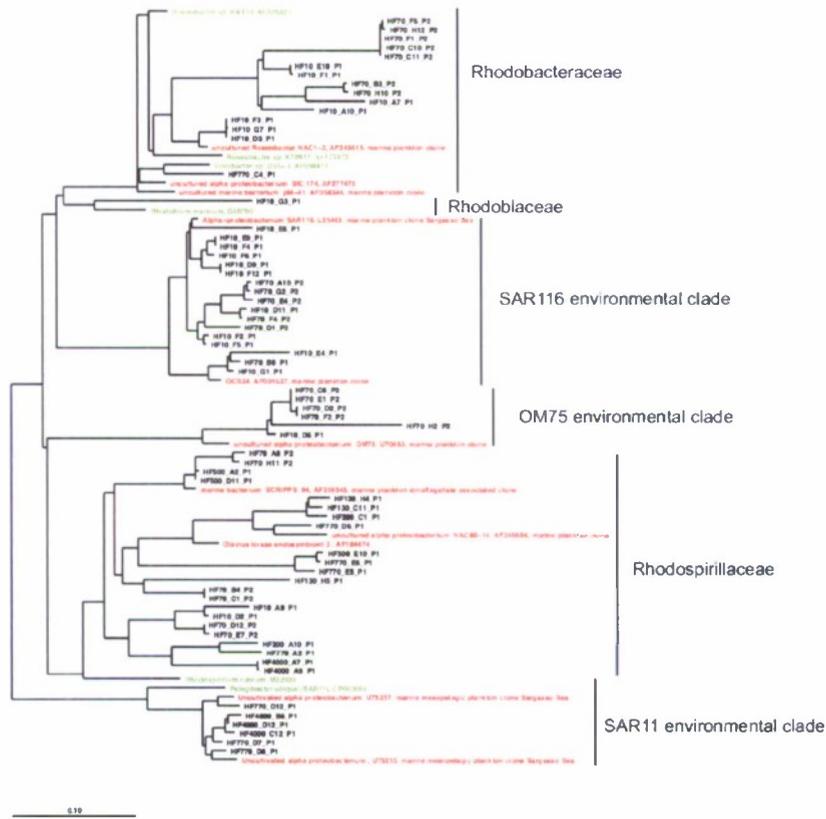
fig. S3B,   Phylogenetic positions of Alphaproteobacteria-like fosmids based on rRNA sequence.

SAR324 environmental clade (Deltaproteobacteria)

OM27 environmental clade (Deltaproteobacteria)

Nitrospina (Deltaproteobacteria)

Myxococcaceae (Deltaproteobacteria)
Desulfurellaceae (Deltaproteobacteria)
Desulfuromonadales (Deltaproteobacteria)
Desulfovibrionaceae (Deltaproteobacteria)
Flavobacteraceae (Bacteroidetes)

Flexibacteraceae (Bacteroidetes)

Crenotrichaceae (Bacteroidetes)

SAR406 environmental clade

Planctomycetaceae

Verrucomicrobiales

Prochlorococcus marinus (Cyanobacteria)

Uncultured Cyanobacteria

Candidatus "Microthrix"

Gemmatimonadaceae

Acidobacteriales

Bacillaceae

Deferribacteriales

Other Chloroflexi

SAR202 environmental clade (Chloroflexi)

fig. S3C, Phylogenetic positions of fosmids from other bacterial groups based on rRNA sequences.
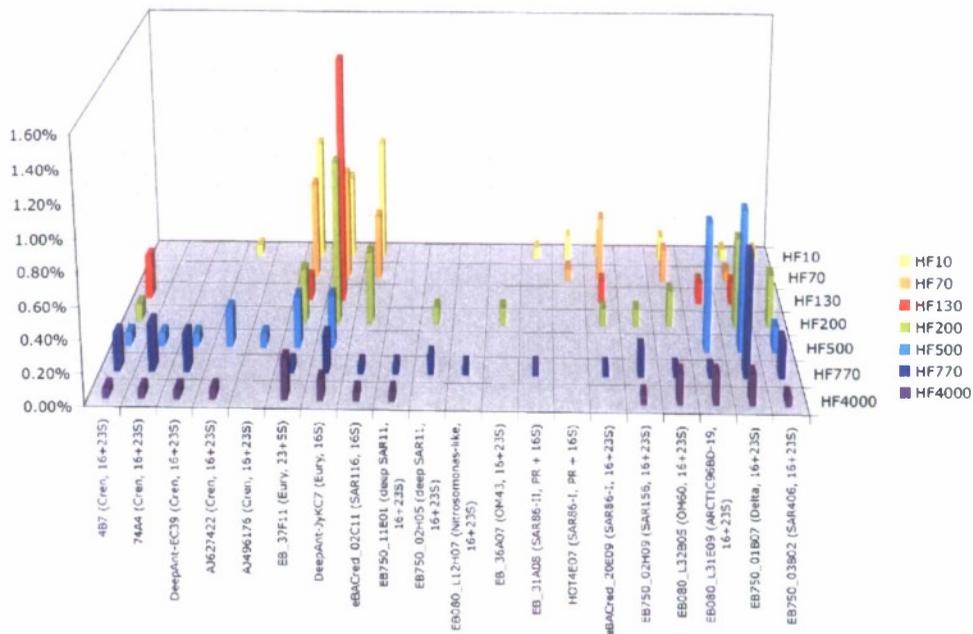
fig S4. Top HSPs matching phylogenetically identified large genome fragments previously recovered from marine plankton. Sequences from each depth were searched against the NCBI non-redundant protein database, and top HSPs matching large DNA insert plankton clones were identified. BLAST searches were performed using blastx, and an expectation cutoff value of 1e-60.

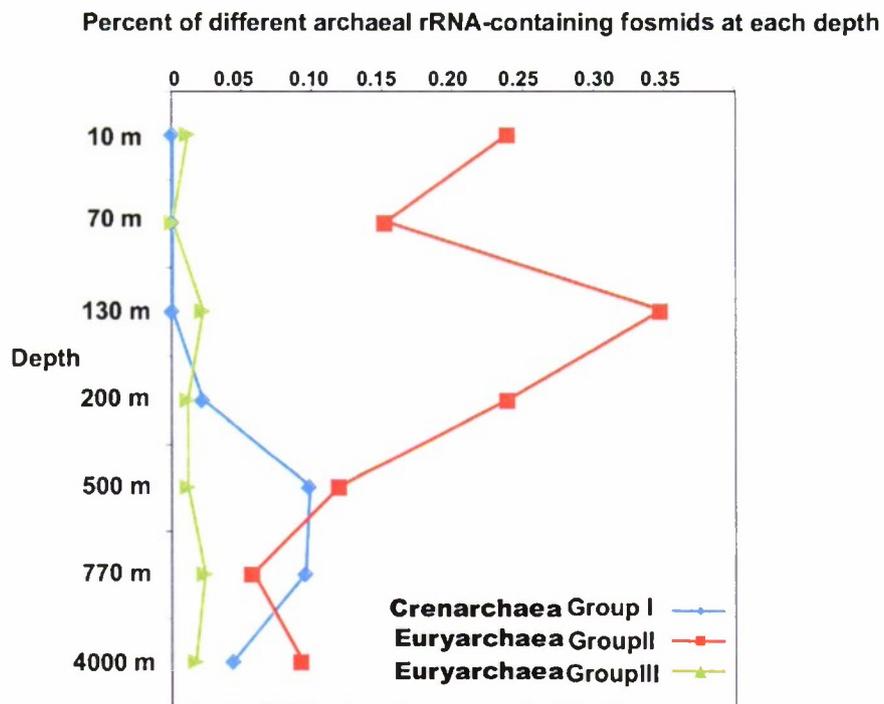**Percent of different archaeal rRNA-containing fosmids at each depth**

fig. SS. Depth distributions of archaeal SSU rRNA-containing fosmids. The percentage of fosmid clones containing an archaeal SSU rRNA gene from each depth. Clones were identified by macroarray colony hybridization, and their SSU rRNAs PCR amplified and sequenced (fig. S6). All fosmid-encoded archaeal rRNAs were associated with one of three groups: Group I Crenarchaea (n= 37), Group II Euryarchaea (n= 128), or Group III Euryarchaea (n= 12).

fig S6. Archaeal SSU rRNA sequences recovered from each depth. Phylogenetic placement of archaeal rRNA sequences contained on fosmid clones recovered from each depth Sequences were aligned using ARB, and phylogenetic placement estimated using neighbor-joining distance calculations with Jukes-Cantor correction (bootstrap support in percentage based on 1000 replications is shown at nodes of > 60%) (4).
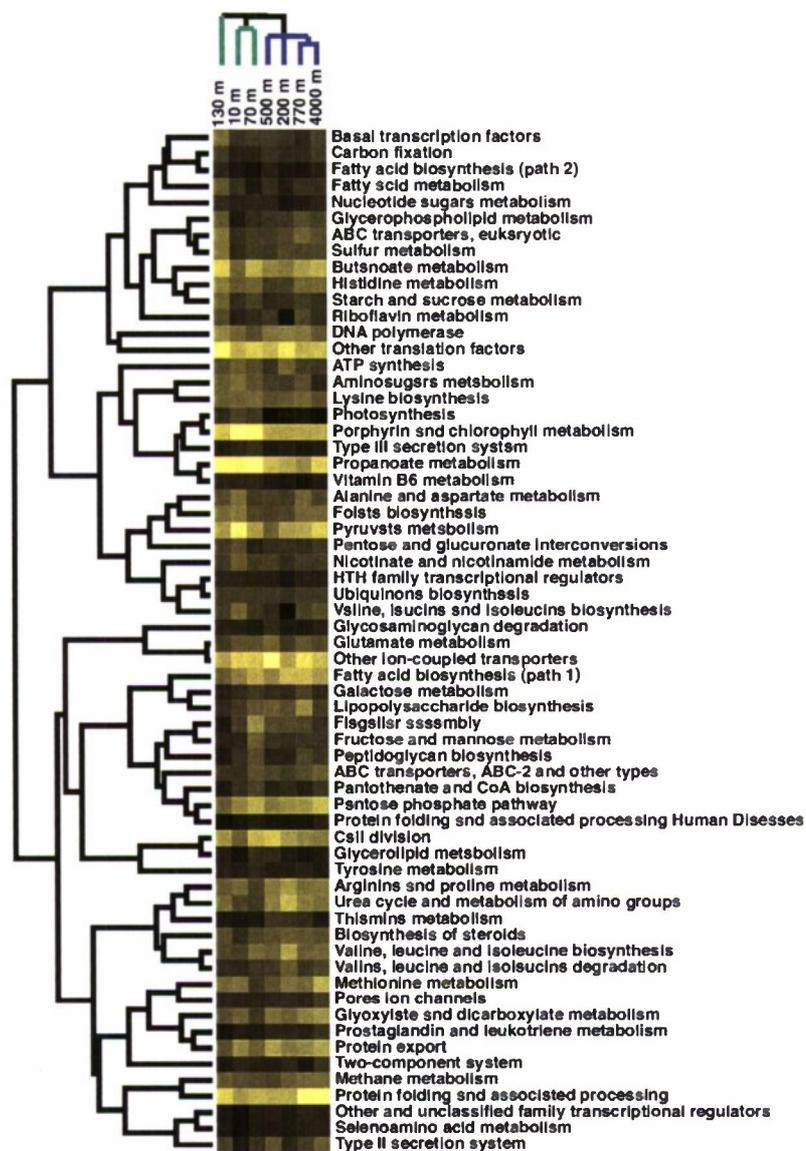
fig S7. Cluster analyses of KEGG annotated gene sequences recovered from different depths. The percentage of KEGG annotated sequences found in each category. The yellow shading is proportional to the percentage of identified sequences falling into each KEGG category at each depth. Dendrograms were constructed as described in Figure 4, main text. The percentage of COG annotated sequences found in each category. Each category shown is represented in at > 0.3 % of the total KEGG-categorized genes at every depth. For display purposes, categories that containing > 2% at > 2depths are not shown.
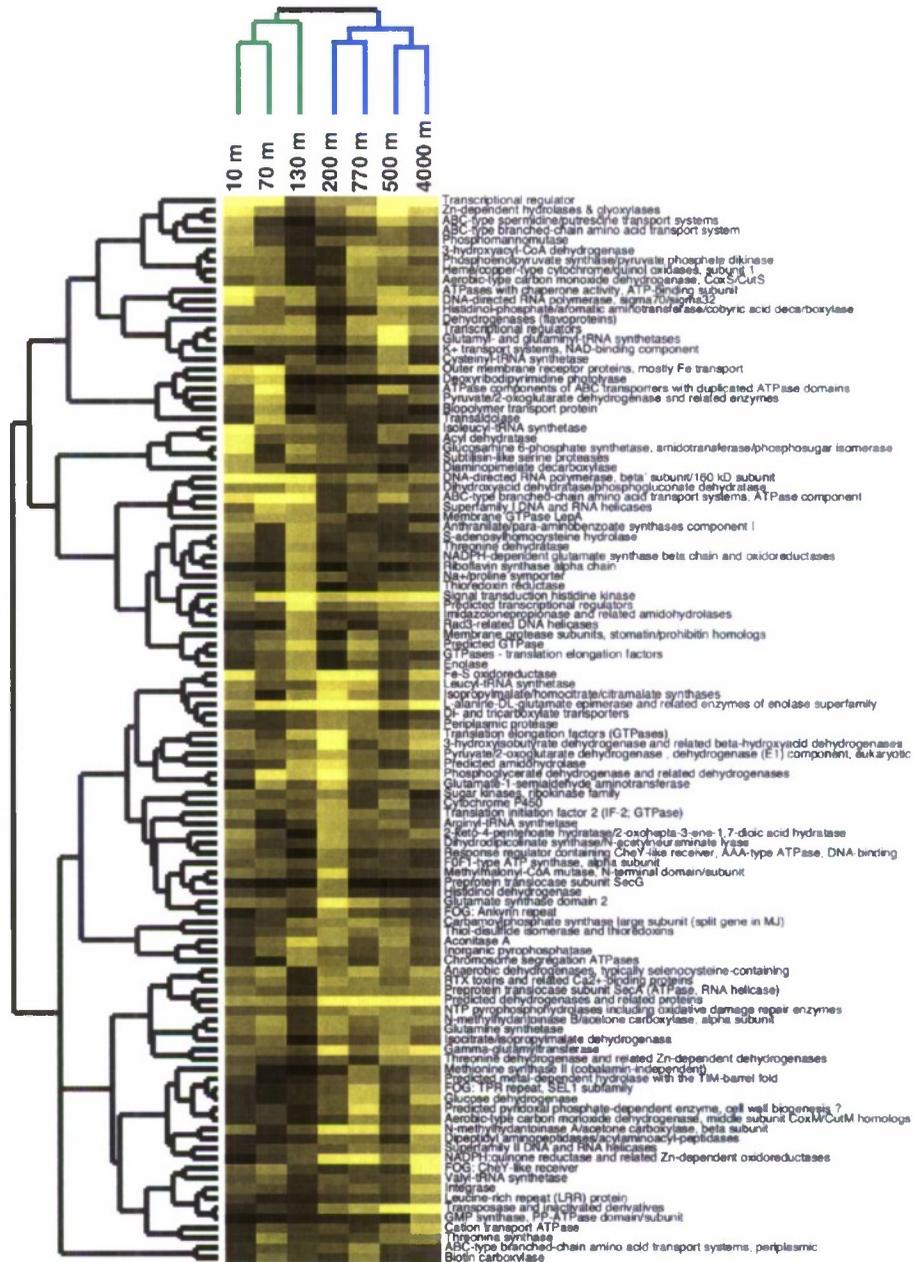
fig S8. Cluster analyses of COG annotated gene seqeunces recovered from different depths. The percentage of COG annotated sequences found in each category. The yellow shade is proportional to the percentage of identified sequences in each COG category at each depth. Dendrograms were constructed as described in Figure 4, main text. Each category shown contains at least two genes > 0.5 % of the total KEGG-categorized genes at each depth.
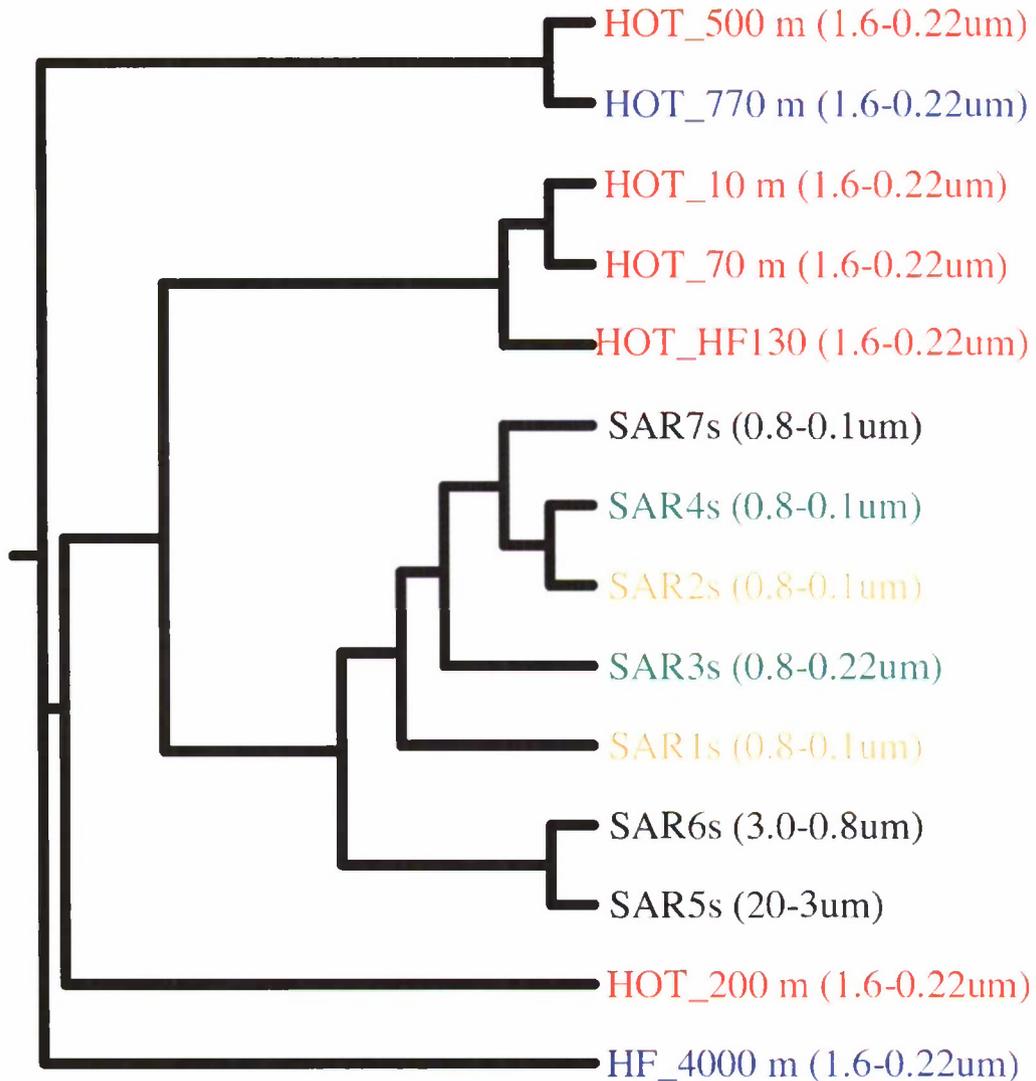
fig S9. Dendrogram of a cumulative TBLASTX bitscore distance matrix comparing Sargasso Sea and North Pacific Subtropical Gyre samples.   Cluster analyses are derived from pairwise cumulative TBLASTX bitscore comparisons, between HOT (this study) and Sargasso Sea microbial community samples (13). The dendrogram was generated as described above in Methods, Supporting Online Materials. Branching patterns, but not branch lengths are not shown. A total of 10,000 sequences were randomly selected from each of the seven Sargasso Sea datasets for comparison to similarly sized HOT sequence datasets.   Values in parenthesis represent the size fraction collected in each sample, in microns. Colors correspond to samples collected on the same date. The Sargasso Sea sample numbers correspond to sequence bins from individually collected seawater samples as previously reported (13).
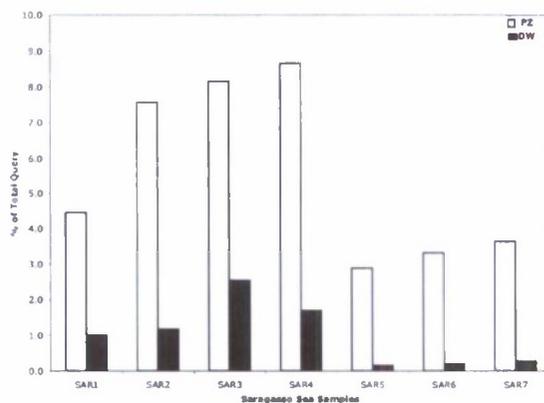
fig. S10A.

fig. S10B.
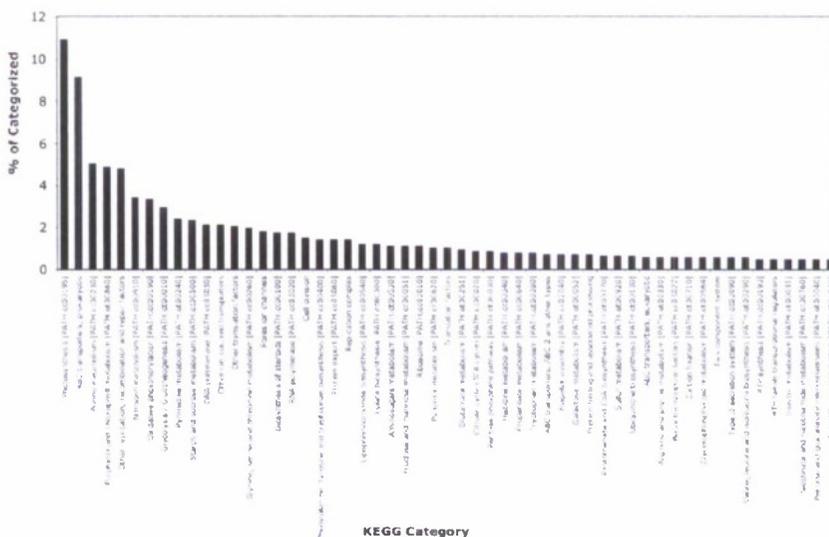
fig S10. Comparison of North Pacific Subtropical Gyre photic zone (PZ) and deep water (DW) sequence sets to Sargasso Sea surface water sequences. fig. S10A. Percentage of PZ or DW sequences relative to their respective totals, that match Sargasso Sea samples at an expectation value of < 1e-20 in blastn searches. The Sargasso Sea sample numbers correspond to sequence bins from individually collected seawater samples, as previously reported (13). fig S10B. KEGG category distributions of PZ-related sequences in combined Sargasso Sea sample bins. Percentage of the total KEGG-categorized sequences in each KEGG category present is shown. Analyses were performed using BLASTX as described in supporting online material methods, with an expectation cutoff threshold of 1xe-S.
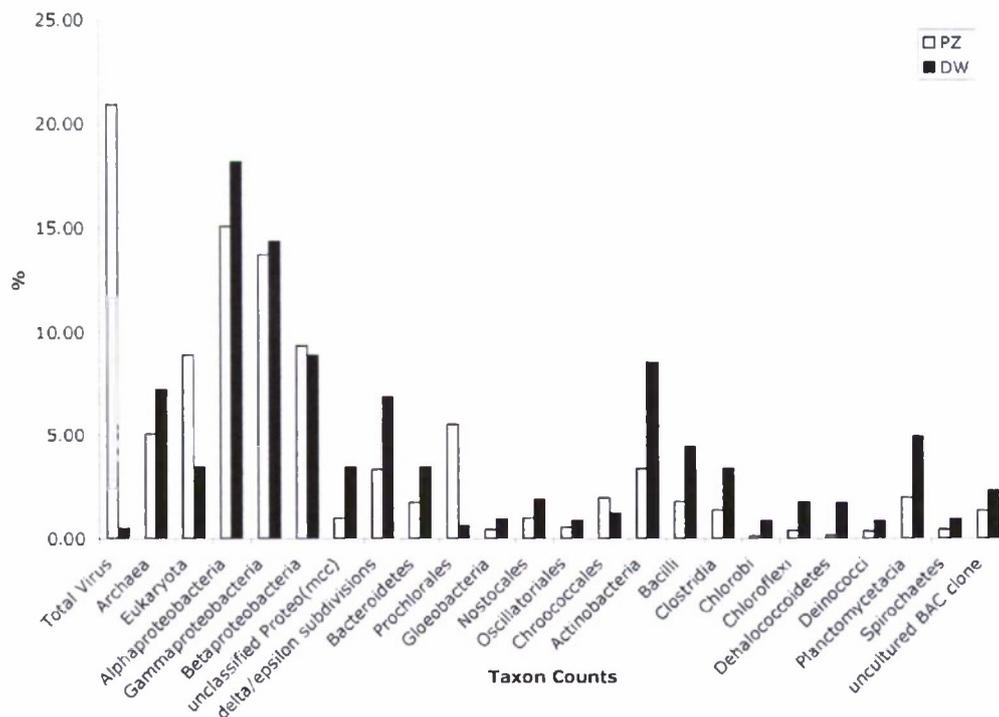
fig. S11. Taxon categorization of top HSPs in photic zone and deepwater sequence bins. The percentage of total sequences from PZ or DW, with top HSPs in each of the taxon categories shown. Sequences were compared in BLASTX searches against the NCBI non-redundant protein database, using an expectation cutoff value 1e-8 for taxon binning.
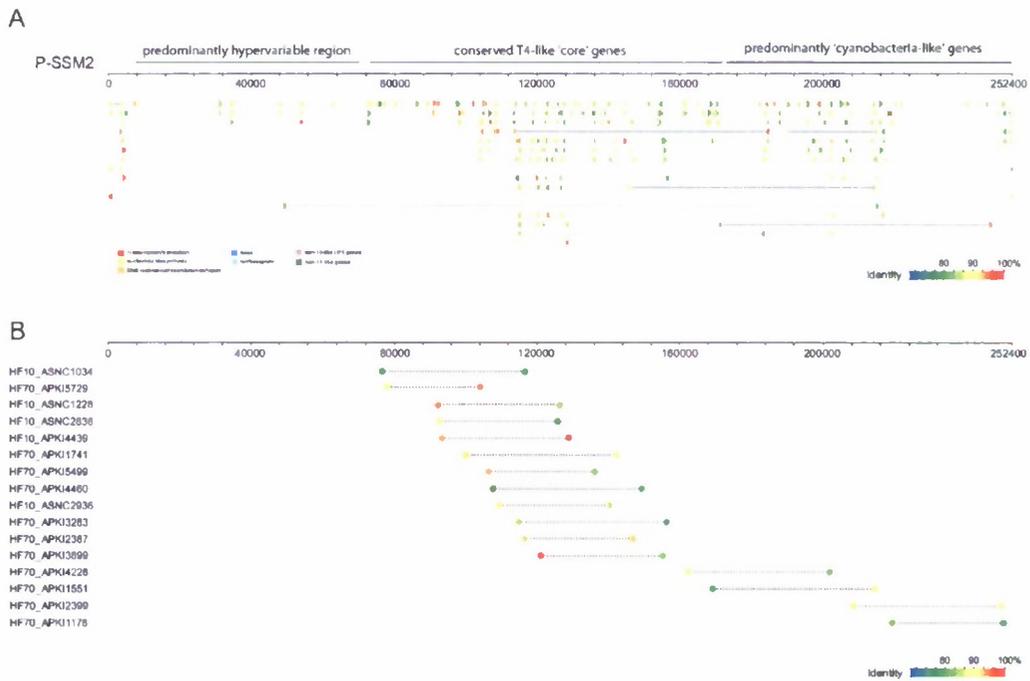
fig S12. Alignment of cyanophage fosmid-mate pairs along a fully sequenced cyanophage genome. fig S12A. Cyanophage-like fosmid end sequences were compared to the complete genome of cyanophage PSSM-2 by blastn analyses at an expectation cutoff of 1x10-3. All sequences matching at similarities > 80% identity are shown. Colors indicate the level of sequence similarity. Lines split single sequences that match the genome at high similarity in different regions across the sequence. fig S12B. Selected fosmids that matched PSSM-2 with high sequence similarity on both fosmid ends were aligned to the whole phage genome sequence. Colors indicate the level of sequence similarity.

DEPTH

Subsystem	10 m	70 m	130 m	200 m	500 m	770 m	4000 m
1,2-Dichloroethane degradation [PATH:ot00631]	1	4	1		4	1	1
1- and 2-Methylnaphthalene degradation [PATH:ot00624]	1	3				2	2
2,4-Dichlorobenzoate degradation [PATH:ot00623]					2		
ABC transporters, ABC-2 and other types		1		6	1	1	1
ABC transporters, eukaryotic			1	3		1	
ABC transporters, prokaryotic		5			5		
ATP synthesis [PATH:ot00193]				1			
ATPases					1		5
Alanine and aspartate metabolism [PATH:ot00252]	4	3	4	1	1		
Aminoacyl-tRNA biosynthesis [PATH:ot00970]	2			3	1	1	2
Aminophosphonate metabolism [PATH:ot00440]						2	
Aminosugars metabolism [PATH:ot00530]	4	2					
Arginine and proline metabolism [PATH:ot00330]		1		1			1
Ascorbate and aldarate metabolism [PATH:ot00053]			1	3			2
Atrazine degradation [PATH:ot00791]	1	2	3	2	1		
Bacterial chemotaxis		1					
Basal transcription factors [PATH:ot03022]					1		
Benzoate degradation via CoA ligation [PATH:ot00632]	2	3	2			1	2
Benzoate degradation via hydroxylation [PATH:ot00362]	2	1					2
Biosynthesis of ansamycins [PATH:ot01051]						2	
Biosynthesis of siderophore group nonribosomal peptides [PATH:ot01053]				4			1
Biosynthesis of steroids [PATH:ot00100]		1		1	1	1	1
Biotin metabolism [PATH:ot00780]			1			3	
Bisphenol A degradation [PATH:ot00363]				2		1	4

Table S1. Fosmid end sequeces with open reading frames orhtologous to genes in specific KEGG and SEED annotated pathway categories, that are statistically more represented at one depth compared to others in pairwise comparisons. Numbers represent the number of parwise comparisons with other depths that show overrepresentation at the depth indicated.

Subsystem	DEPTH						
	10 m	70 m	130 m	200 m	500 m	770 m	4000 m
Butanoate metabolism [PATH:ot00650]	1			3	1	1	
Calcium signaling pathway		1					
Caprolactam degradation [PATH:ot00930]		1		1	1	1	2
Carbon fixation [PATH:ot00710]	1	3	1			1	
Cell division		1			1	1	1
Citrate cycle (TCA cycle) [PATH:ot00020]						5	
Cyanoamino acid metabolism [PATH:ot00460]				4			
Cysteine metabolism [PATH:ot00272]	1	2	1	3	1	1	
Cytokines	3				2	2	
D-Alanine metabolism [PATH:ot00473]	3					3	
D-Glutamine and D-glutamate metabolism [PATH:ot00471]		1		1			
Diterpenoid biosynthesis [PATH:ot00904]			5				
Enzyme		1		1	1	1	1
Fatty acid biosynthesis (path 1) [PATH:ot00061]		1		1	1	1	1
Fatty acid metabolism [PATH:ot00071]			4				
Flagellar assembly [PATH:ot02040]		6					
Fluorene degradation [PATH:ot00628]				2	3		1
Folate biosynthesis [PATH:ot00790]	5		4	1		1	
Fructose and mannose metabolism [PATH:ot00051]		3		3	3	1	3
Galactose metabolism [PATH:ot00052]		1		2		4	2
Ganglioside biosynthesis [PATH:ot00604]			3				
Globoside metabolism [PATH:ot00603]		2				1	1
Glutamate metabolism [PATH:ot00251]	1		1	2		1	1
Glutathione metabolism [PATH:ot00480]	3	1	1	1	1		1
Glycerolipid metabolism [PATH:ot00561]		3			1		1
Glycerophospholipid metabolism [PATH:ot00564]				2		1	
Glycine, serine and threonine metabolism [PATH:ot00260]		1	1			1	3
Glycolysis / Gluconeogenesis [PATH:ot00010]		1	1	1		1	1
Glycosphingolipid metabolism [PATH:ot00600]		1					3

Subsystem	DEPTH						
	10 m	70 m	130 m	200 m	500 m	770 m	4000 m
Glyoxylate and dicarboxylate metabolism [PATH:ot00630]	2			4	3		3
HTH family transcriptional regulators	2				4		2
Histidine metabolism [PATH:ot00340]			2	1			1
Huntington's disease				5			
Inositol metabolism [PATH:ot00031]	1			2	2	2	2
Lipopolysaccharide biosynthesis [PATH:ot00540]				4		5	
Lysine biosynthesis [PATH:ot00300]							3
Lysine degradation [PATH:ot00310]	1	2			2	2	2
Major facilitator superfamily (MFS)			3			2	
Methane metabolism [PATH:ot00680]	1			3	1	3	4
Methionine metabolism [PATH:ot00271]	1		3		1		1
N-Glycan biosynthesis [PATH:ot00510]			3			2	
Nicotinate and nicotinamide metabolism [PATH:ot00760]	1						
Nitrobenzene degradation [PATH:ot00626]		3					2
Nitrogen metabolism [PATH:ot00910]							4
Non-enzyme		2		1	2		
Novobiocin biosynthesis [PATH:ot00401]			3		2	1	2
Nucleotide sugars metabolism [PATH:ot00520]	2				2	2	1
One carbon pool by folate [PATH:ot00670]	1		1			5	2
Other and unclassified family transcriptional regulators				1	1	1	
Other ion-coupled transporters		1			4	4	1
Other replication, recombination and repair factors				4		4	4
Other translation factors	2		2	1	1		1
Other transporters	2		2		5	2	2
Pantothenate and CoA biosynthesis [PATH:ot00770]		2		1	1		
Penicillins and cephalosporins biosynthesis [PATH:ot00311]				6			
Pentose and glucuronate interconversions [PATH:ot00040]	1	2		2	2	2	
Pentose phosphate pathway [PATH:ot00030]		1		3			

Subsystem	DEPTH						
	10 m	70 m	130 m	200 m	500 m	770 m	4000 m
Peptidoglycan biosynthesis [PATH:ot00550]		4			2	1	1
Phenylalanine metabolism [PATH:ot00360]	2	2			2	3	
Phenylalanine, tyrosine and tryptophan biosynthesis [PATH:ot00400]	2		5	1		1	1
Phosphatidylinositol signaling system [PATH:ot04070]	2						
Phosphotransferase system (PTS)	1		2	3		1	
Photosynthesis [PATH:ot00195]	4	4	4	2			
Pores ion channels				2	1	1	2
Porphyrin and chlorophyll metabolism [PATH:ot00860]	1		1	1			
Prion disease		1		1	1		1
Propanoate metabolism [PATH:ot00640]		1		2	1	1	1
Proteasome [PATH:ot03050]		5					
Protein export [PATH:ot03060]				2	1		2
Protein folding and associated processing						3	
Purine metabolism [PATH:ot00230]			4	1			
Pyrimidine metabolism [PATH:ot00240]		4	1		3		
Pyruvate metabolism [PATH:ot00620]			1				
Pyruvate/Oxoglutarate oxidoreductases					2		
RNA polymerase [PATH:ot03020]	5	1	3		2		
Reductive carboxylate cycle (CO2 fixation) [PATH:ot00720]			3	1		1	
Replication complex	1		4			1	
Restriction enzyme							4
Riboflavin metabolism [PATH:ot00740]	1	1	2		3	3	1
Ribosome [PATH:ot03010]	5	1	3	1	1	1	
Selenoamino acid metabolism [PATH:ot00450]			1			1	1
Sporulation	2		2				
Starch and sucrose metabolism [PATH:ot00500]		1	2		1	1	1
Stilbene, coumarine and lignin biosynthesis [PATH:ot00940]						2	
Streptomycin biosynthesis [PATH:ot00521]	1	1	1		2	1	1
Styrene degradation [PATH:ot00643]						2	
Tetracycline biosynthesis [PATH:ot00253]	1	3					

Subsystem	DEPTH						
	10 m	70 m	130 m	200 m	500 m	770 m	4000 m
Thiamine metabolism [PATH:ot00730]	1		1	1	4	1	1
Tight junction	5						
Translation factors	1		2	5		3	1
Tryptophan metabolism [PATH:ot00380]	4	4					
Two-component system	2	1	1		1	1	2
Type II secretion system [PATH:ot03090]	1		1		1	1	1
Type III secretion system [PATH:ot03070]		2					
Tyrosine metabolism [PATH:ot00350]	1	4	1		2	2	1
Ubiquinone biosynthesis [PATH:ot00130]			1		2	2	
Urea cycle and metabolism of amino groups [PATH:ot00220]	1		1	6	1	1	1
Valine, leucine and isoleucine biosynthesis [PATH:ot00290]		1		5		1	1
Valine, leucine and isoleucine degradation [PATH:ot00280]		3	1	5	1		1
Vitamin B6 metabolism [PATH:ot00750]		2					
beta-Alanine metabolism [PATH:ot00410]	1	6				1	
gamma-Hexachlorocyclohexane degradation [PATH:ot00361]					3		

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO. MIT/WHOI 2008-14	2.	3. Recipient's Accession No.

4. Title and Subtitle

Development of a "Genome-Proxy" Microarray for Profiling Marine Microbial Communities, and its Application to a Time Series in Monterey Bay, California

5. Report Date
September 2008

6.

7. Author(s) Virginia Isabel Rich

8. Performing Organization Rept. No.

9. Performing Organization Name and Address

MIT/WHOI Joint Program in Oceanography/Applied Ocean Science & Engineering

10. Project/Task/Work Unit No.
MIT/WHOI 2008-14

11. Contract(C) or Grant(G) No.
(C) MCB-0348001

(G)

12. Sponsoring Organization Name and Address
National Science Foundation
Massachusetts Institute of Technology

13. Type of Report & Period Covered
Ph.D. Thesis

14.

15. Supplementary Notes

This thesis should be cited as: Virginia Isabel Rich, 2008. Development of a "Genome-Proxy" Microarray for Profiling Marine Microbial Communities, and its Application to a Time Series in Monterey Bay, California. Ph.D. Thesis. MIT/WHOI, 2008-14.

16. Abstract (Limit: 200 words)

This thesis describes the development and application of a new tool for profiling marine microbial communities. Chapter 1 places the tool in the context of the range of methods used currently. Chapter 2 describes the development and validation of the "genome proxy" microarray, which targeted marine microbial genomes and genome fragments using sets of 70-mer oligonucleotide probes. In a natural community background, array signal was highly linearly correlated to target cell abundance (R^2 of 1.0), with a dynamic range from 10^2-10^6 cells/ml. Genotypes with $\geq \sim 80\%$ average nucleotide identity to those targeted cross-hybridized to target probesets but produced distinct, diagnostic patterns of hybridization. Chapter 3 describes the development an expanded array, targeting 268 microbial genotypes, and its use in profiling 57 samples from Monterey Bay. Comparison of array and pyrosequence data for three samples showed a strong linear correlation between target abundance using the two methods (R^2=0.85-0.91). Array profiles clustered into shallow versus deep, and the majority of targets showed depth-specific distributions consistent with previous observations. Although no correlation was observed to oceanographic season, bloom signatures were evident. Array-based insights into population structure suggested the existence of ecotypes among uncultured clades. Chapter 4 summarizes the work and discusses future directions.

17. Document Analysis   **a. Descriptors**
microarray
Monterey

b. Identifiers/Open-Ended Terms

c. COSATI Field/Group

18. Availability Statement

Approved for publication; distribution unlimited.

19. Security Class (This Report)
UNCLASSIFIED

20. Security Class (This Page)

21. No. of Pages
282

22. Price

(See ANSI-Z39.18)   *See Instructions on Reverse*   OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce