Working title: Time-series analyses of Monterey Bay coastal microbial picoplankton using a "genome proxy" microarray

Authors: Virginia Rich[1,2], Vinh Pham[2], John Eppley[2], Yanmei Shi[2], and Edward F. DeLong[2,*]

[1] current address: Department of Ecology and Evolutionary Biology, University of Arizona, 1041 East Lowell Street, Tucson, AZ 85721

[2] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 48-427, 15 Vassar Street, Cambridge, MA 02139

* To whom correspondence may be addressed: delong@mit.edu

Running title: Monterey Bay community dynamics by "genome proxy" array

Abstract

To gain improved temporal, spatial and phylogenetic resolution of marine microbial communities, in this study we expanded the original prototype genome proxy array (an oligonucleotide microarray targeting marine microbial genome fragments and genomes), evaluated it against metagenomic sequencing, and applied it to time series samples from the Monterey Bay long term ecological research site. The expanded array targeted 268 microbial genotypes (vs. 14 in the original prototype) across much of the known diversity of cultured and uncultured marine microbes. The target abundances measured by the genome proxy array were highly correlated to pyrosequence-based abundances (linear regression $R^2 = 0.85\text{-}0.91$, $p<0.0001$). Fifty-seven samples from ~4-years in Monterey Bay were examined with the array, spanning the photic zone (0m), the base of the surface mixed layer (30m), and the subphotic zone (200m). A significant portion of the expanded genome proxy array's targets showed signal (95 out of 268 targets present in ≥ 1 sample). The multi-year community survey showed the consistent presence of a core group of common and abundant targeted taxa at each depth in Monterey Bay, higher variability among shallow than deep samples, and episodic occurrences of more transient marine genotypes. The abundance of the most dominant genotypes peaked after strong episodic upwelling events. The genome-proxy array's ability to track populations of closely-related genotypes indicated population shifts within several abundant target taxa, with specific populations in some cases clustering by depth or oceanographic season. Although 51 cultivated organisms were targeted (representing 19% of the array) the majority of targets detected and of total target signal (85% and ~92%, respectively) were from uncultivated lineages, often those derived from Monterey Bay. The array provided relatively cost-effective approach (~$15 per array) for surveying the natural history of uncultivated lineages in the wild.

Introduction

Marine microbial communities are major drivers in global biogeochemical cycling (Arrigo, 2005; Howard et al., 2006; Karl, 2007), sources of metabolic discoveries (e.g. (Béjà et al., 2000; Kolber et al., 2000; Dalsgaard et al., 2003; Kuypers et al., 2003), and the focus of metagenomic surveys beyond the scale of those yet undertaken in other habitats (Venter et al., 2004; Tringe et al., 2005; DeLong et al., 2006; Kennedy et al., 2007; Rusch et al., 2007; Wegley et al., 2007; Wilhelm et al., 2007; Yooseph et al., 2007; Dinsdale et al., 2008; Marhaver et al., 2008; Mou et al., 2008; Neufeld et al., 2008). However, microbial community dynamics remain poorly understood due to technical limitations and the analytical challenges of high-resolution spatial and temporal studies. Most studies capture spatiotemporal snapshots or focus on one or a few groups over space and time. While the value of improved resolution is clear, lower resolution (e.g., in time, space, or diversity of target organisms) studies have provided much insight into microbial community variability over space and time. For example, such studies reveal changing community structure that correlates to environmental parameters, and even climate change responses (e.g., Hawaii Ocean Time Series (Karl, 1999; Karner et al., 2001), Bermuda Atlantic Time Series (Morris et al., 2005), and San Pedro Ocean Time-Series (Fuhrman et al., 2006).

To gain a higher resolution picture of microbial community variability, we developed the “genome proxy” array (Rich et al., 2008) which uses sets of multiple, distributed 70-mer probes to target genotypes (genome fragments and genomes) as a cost-effective high-throughput survey tool to track microbial community variability. The array cross-hybridizes to related genotypes that approach ≥ ~80% average nucleotide identity (ANI, as in (Konstantinidis and Tiedje, 2005), with the stringency and specificity adjustable *in silico* to ≥ ~90% ANI. Related cross-hybridizing strains produced distinct hybridization patterns across their target probe set, and the array can thereby reveal shifts in population structure across samples (Rich et al., 2008). The limit of detection is approximately 0.1% of the community for targeted genotypes, and approximately 1% of the community for related, cross-hybridizing genotypes (Rich et al., 2008).

We report here on an expanded genome proxy array that targets 268 genotypes (from 14 in the original). We ground-truthed the array signal using pyrosequenced community DNA, and applied the optimized array to investigate the time series microbial dynamics over a four-year period at Monterey Bay Station M1 (36.747˚ N, 122.022˚ W). This microbially and oceanographically well-studied coastal environment (e.g. Pennington, 2000; Suzuki et al., 2001a,b; Suzuki et al., 2004; O'Mullan and Ward, 2005; Ward, 2005; Mincer et al., 2007; Pennington et al., 2007) is characterized by strong seasonal upwelling, providing a contextually-rich first real-world application of this tool. In all, we hybridized 57 archived DNA samples collected over 4 years from oceanographic water column features (photic, base of the mixed layer, and subphotic) to identify patterns in and drivers of microbial community structure.

Results and discussion

Development and ground-truthing of the Expanded Genome Proxy Array

The expanded genome proxy array targets 268 microbial genotypes, through suites of probes (~20 per target) dispersed along genomes and genome fragments derived from microbes inhabiting marine habitats. Targeted organisms were selected to span known marine microbial diversity (16S rRNA-containing targets are shown in Fig. 1 and Figs. S1-S5, all targets are listed in Table S1 and summarized in Table S2). For diverse and abundant marine clades, representatives were chosen where possible from each known lineage and from multiple geographic origins.

We compared the results from the expanded array to those obtained using pyrosequencing of the same microbial community DNA for three different Monterey Bay surface samples (Julian Day (JD) 298 in 2000, and JD115 and JD135 in 2001). A full GS-FLX pyrosequencing run (~400,000 reads) was performed per sample, trimmed to remove poor quality sequence (~5.5% of reads), and "hybridized" *in silico* using BLAST (Altschul et al., 1990) to the 268 genotypes targeted by the array. To simulate the amount of sequence divergence tolerated by the array, BLAST parameters were calibrated using array results for genomes of related *Prochlorococcus* strains whose relative cross-hybridization to the array had been experimentally determined (Rich et al., 2008). Using this approach (see Methods), 1.9%-2.5% of the total pyrosequencing reads in these three samples were assigned to array targets (7636/395767 for 0m_2000_298, 8743/345650 for 0m_2001_115, and 9252/39197 for 0m_2001_135), of which ~66-75% were assigned to only 12 targets in all three samples. Eleven of these 12 targets were environmental genomic clones (predominantly from the SAR86 and Roseobacter clades) while the tenth was the genome of a cultured NAC11-7 clade *Roseobacter*.

The normalized pyrosequencing read recruitment was strongly correlated to the normalized unfiltered mean array intensity (linear regression with R^2 of 0.85-0.91 across three samples, p-values

<0.0001; Fig. 2). Such strong correlation between the relatively unbiased (no cloning biases, etc.) direct pyrosequencing method and the high-throughput genome proxy array provided support for the veracity of the array as a tool for profiling studies requiring high sample throughput.

Exploring microbial communities using the genome proxy array

We hybridized community DNA from 57 Monterey Bay samples at station M1 over 4 years (sample overview in Fig. 3) to the expanded genome proxy microarray. Approximately one-third of the array's diverse targets (95 of 268 targets) were present in one or more of the samples at this site. To be considered present, a target was required to show signal in >40% of its probes, to avoid single-probe high-identity cross-hybridizations from unrelated taxa (as empirically determined in Rich et al., 2008, see Methods). The majority of targets detected by array were uncultivated marine lineages, many of which originated from Monterey Bay (Fig. S6a).

i. Shallow versus deep profiles: Hierarchical clustering (Fig. 4) and canonical discriminant analyses (CDA, Fig. 5) revealed clear community structure throughout the oceanographic depth profiles sampled, with greater variability among shallow samples than deep ones (see branch lengths of hierarchical clustering and intensity of array signals in Fig. 4). For example, the Monterey Bay surface photic zone samples (0 and 30m) were less similar to each other (as indicated by branch distances) than the subphotic zone samples were to one another (200m, Fig. 4, Fig. 5). Depth-structuring in microbial populations and communities is well-described in marine systems at the level of rRNA profiling (e.g. Fuhrman et al., 1992; Field et al., 1997; Karner et al., 2001; Bano and Hollibaugh, 2002; Morris et al., 2004; Suzuki et al., 2004; Treusch et al., 2009) and fosmid end-sequencing (DeLong et al., 2006), so it is not surprising that our genome proxy array reveals similar structure with respect to the targeted community genotypes examined here. These differential depth distributions extended to the majority of observed taxa, with 4 notable depth-specific groups of targets (dashed boxes in Fig. 4 and detailed in Table 1). Eight targets were present in >90% of shallow samples ("shallow-consistent"), 10 were present in 50-90% of shallow samples ("shallow-frequent"), 10 were present in >90% of deep samples ("deep-consistent"), and 3 were present in 50-90% of deep samples ("deep-frequent") (Table 1). Notably, the differential presence and distribution of 3-5 targeted genotypes in each depth drove the three depth's separation of array profiles

(Canonical Discrimination Analysis, Fig. 5a).

While there was clear photic vs subphotic depth structure, the 0m and 30m array profiles were intermingled despite their generally different chemical and physical environments (Fig. 3). While we selected 30m as the base of the mixed layer to attempt to capture the nitricline, it is clear that the mixed layer depth (MLD) at this site usually lacks a discrete thermocline and moves dramatically over short time periods (see calculated MLD across sampling dates, Fig. S7). Therefore, our sampling strategy might have been improved by varying sampling depths based on calculated single time-point MLDs for each cruise; however removing 30m samples that were clearly above the MLD and reclustering the array profiles did not resolve samples into 0m and 30m clusters (Fig. S8), emphasizing the highly dynamic nature of these photic-zone waters.

ii. Profile correlations to ocean chemistry: Array-based sample profiles compared between depths were strongly correlated to each tested nutrient as follows: phosphate, nitrate and silicate drove the differentiation of the shallow from the deep samples, while nitrite drove the separation of 30m from 0m (Fig. 5b). Samples from each depth were separately subjected to PCA (Fig. 6), indicating that nutrients did not separate the 0m samples (Fig. 6a), but were important at both 30m and 200m. Specifically, at 30m (Fig. 6b), nutrient variability was correlated to the principal component axes, with a strong upwelling signal of phosphate, nitrate and silicate and a slightly weaker and inverse signal for nitrite (likely from remineralization). Finally, at 200m (Fig. 6c), nitrate and nitrite showed no and weak correlations, respectively, while silicate and phosphate gave strong but non-overlapping correlations. Overall, these correlations to nutrient concentrations recapitulate the oceanographic differences in nutrients with depth at this location (Fig. 3).

iii. Tracking abundant taxa: Not surprisingly, one of the most commonly detected bacterial groups was the *Roseobacter* clade (Fig. 4). This metabolically diverse group commonly comprises up to 20% of cells in coastal waters (reviewed in Buchan et al., 2005), including high abundances (20-40% of rRNA clone libraries) in the mid-Monterey Bay region during upwelling (Suzuki et al., 2001b). More specifically, in fosmid clone libraries from Monterey Bay the *Roseobacter* NAC11-7 and CHAB-I-5 clades comprised nearly 30% of the 16S-containing clones (27 and 29% at 0 and 80m, respectively) and ~80% of the total

Roseobacter signal at 0 and 80m, while at 100m NAC11-7 disappeared and CHAB-1-5 persisted at low abundance (Suzuki et al., 2004) (see Table S3 for clade-by-clade comparison of array results with previous Monterey Bay community surveys). In agreement with these previous single time-point observations, the array profiles indicate high *Roseobacter* abundances over time (Figs. 4 and S9a). Twenty-eight percent of the commonly-occurring targeted taxa in surface waters were NAC11-7 clones (4 of 8 targets in the *shallow-consistent* group, and 1 of 10 *shallow-frequent* group; listed in Table 1), and 1 of the10 *deep-consistent* taxa was a CHAB-I-5 clone (Table 1). In addition, another CHAB-I-5 clone (EB080_L58F04) was present in 35% of shallow samples. Further, differential NAC11-7 distributions drove the differentiation of 30m from 0m samples (3 of 5 driving taxa, Fig. 5a).

A second abundant shallow water bacterial group was the uncultivated gamma-proteobacterial SAR86 clade, which is commonly reported in marine samples (Eilers et al., 2000; Rappe et al., 2000; Suzuki et al., 2001b; Venter et al., 2004; Morris et al., 2006), 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; Eilers et al., 2000; Morris et al., 2006). In Monterey Bay, it is abundant in rRNA clone libraries during upwelling (3-6% of total bacterial SSU DNAs; Suzuki et al., 2001b), and in large-insert clone libraries (5.6%, 5.5%, and 1.6% respectively of the SSU operon-containing clones 0m, 80m and 100m; Suzuki et al., 2004; Table S3). Array-based profiling reflected also this high SAR86 abundance (Figs. 4 and S9b); 22% of common shallow water targets (2 *shallow-consistent* and 2 *shallow-frequent*) were SAR86 clones. The distribution of one particular SAR86 target (a Monterey-derived environmental clone) helped drive the differentiation of 30m samples from those at 0m (Fig. 5a).

A remaining *shallow-frequent* target of note was an alphaproteobacterial SAR116-I clone. Of 12 SAR116 targets, two originated in Monterey Bay, and these were the only phylotypes detected (Fig. 4). The SAR116-II target was present only twice, in 0m samples, while the SAR116-I clone was present in 62% of shallow samples. In large-insert environmental libraries from this site, the *Rhodospirillales* clade 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; Table S3). The SAR116 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 in culture (Stingl et al., 2007). Due to the phylogenetic diversity of this clade (at least 10% divergent 16S rRNA, Stingl et al., 2007), it is likely that the relative specificity of the array platform prohibited it from tracking other native but divergent SAR116 strains. The comparative array vs. fosmid libraries results suggest the need for additional sequencing of environmental SAR116 genotypes.

Another common marine bacterial clade detected by the array was the alphaproteobacterial SAR11 clade, which is one of the most abundant heterotrophs in the global oceans (Morris et al., 2002). Seven of the 10 targeted SAR11 genotypes were present in ≥ 1 Monterey Bay sample, and each showed depth-specific distribution (Figs. 4 and S9c). *Pelagibacter* HTCC1062 and HTCC1002, cultivated strains within the SAR11 subgroup 1a, were present only in shallow samples and occurred in ~30% of samples (29% and 35%, respectively). Several other SAR11 environmental clone genotypes were present only in deep samples, and occurred frequently or sporadically. This is consistent with the known depth distributions of the two major SAR11 clades (Field et al., 1997). 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). The lower frequency of SAR11 genotypes than other clades, combined with the clade's consistently high abundance measures by other methods, suggests the presence of many other SAR11 genotypes in these samples.

Targeted cyanobacteria did not show strong or consistent array signal in Monterey Bay. *Synechococcus* would be expected to be abundant in such nutrient-rich coastal waters (Waterbury, 1986; Partensky et al., 1999), and the array targeted eight marine *Synechococcus* across the group's known genomic diversity. The absence of strong cyanobacterial signal is therefore may be explained by the use of a 1.6µm pre-filter during sample collection, which may have excluded larger *Synechococcus* cells (average uncultured cell size 0.8-2.2 um, Waterbury et al., 1979). Both *Synechococcus* and *Prochlorococcus* were sporadically detected in surface waters (Fig. 4), and the differential distribution of *Prochlorococcus* MED4 helped differentiate 0m from 30m samples (Fig. 5a).

The array captured information about *deep-consistent* genotypes (Fig. 4, Table 1) including four gammaproteobacterial targets (EB080_L31E09, EB750_10B11, EB750_10A10, and HF4000_23L14) related to chemoautotrophic deep-sea invertebrate symbionts and commonly observed in water column 16S rRNA surveys (López-García et al., 2001; Bano and Hollibaugh, 2002; Zubkov et al., 2002; Klepac-Ceraj, 2004; Suzuki et al., 2004; Stevens and Ulloa, 2008; Walsh et al. 2009), one of which (EB080_L31E09, belonging to the ARCTIC96BD-19 clade) was the most abundant 200m genotype. Two were Form II RuBisCO-containing targets (EB750_10B11, EB750_10A10) without phylogenetic markers but whose BLAST homology indicated relatedness to chemoautotrophic symbionts. A pelagic relative (SUP05) of these targets from Sannich Inlet was recently sequenced metagenomically, and appears to be a chemolithoautotroph that may oxidize reduced sulfur compounds, using nitrate as the terminal electron acceptor, as does it close clam-symbiont relatives (Walsh et al., 2009). Although the oxygen minimum zone in Monterey Bay is significantly deeper than 200m (generally ~700-800m), the consistent presence of these chemoautotrophic relatives at 200m as well as in other aerobic pelagic environments, suggests that either they may be facultatively aerobic and can chemlithoautotrophically or chemoheterotrophically thrive under oxic conditions."

In addition, three deltaproteobacterial targets were common in deep samples (with one SAR324 being *consistent* and one being *frequent*), in agreement with the previous depth preference described for this group (e.g. Wright et al., 1997). These targets were also correlated to the differentiation of 200m from 0m and 30m samples. Another notable *deep-consistent* target was a gammaproteobacterial genotype that clusters within a deep-sea environmental clade (that includes clones ZD0417 and DHB-2) commonly observed in 16S rRNA-gene surveys from a variety of locations (López-García et al., 2001). The natural history and biology of this clade remains a mystery. The genome proxy array can in this way be used to investigate the temporal and spatial dynamics of understudied but abundant organisms for which genomic fragments have been sequenced.

In addition to targeted bacteria, 3 of the 15 targeted archaea were common. Previous FISH investigations in Monterey Bay observed deep and abundant crenarchaeal populations (comprising up to 33% of the 200m community), and euryarchaea throughout the water column at low levels (<1%) with an increase in summer surface waters (up to 12% of the community) (Pernthaler et al., 2002; Mincer et al., 2007). The array signal reflected this general trend with euryarchaeal clones present in both shallow and deep samples, and the restriction of crenarchaeal targets to the deepest samples (Fig. 4), with one crenarchaeal genotype present in 57% of 200m samples (Table 1). In addition, however, two *deep-consistent* euryarchaeal clones were among the most abundant taxa at 200m and present in all sampling dates. This apparent inconsistency with previous observations at this site likely reflects methodological constraints of the FISH-based study, which used surface rather than deep euryarchaeal phylotypes to generate probes and thus may have missed deep genotypes. Indeed rRNA clone libraries from diverse locations have observed appreciable euryarchaeal abundances in deep waters (Massana et al., 1997; López-García et al., 2001; DeLong et al., 2006). The array also revealed that crenarchaeal abundances paralleled those of a lower-intensity *Nitrospina* target (clone EB080_L20F04; Fig. 4), as was previously observed in a qPCR study at this site from 1997-99 (Mincer et al., 2007).

iv. Proteorhodopsin-containing taxa: Proteorhodopsin (PR) is a light-driven proton pump abundant in photic zones (Béjà et al., 2000; Sabehi et al., 2004; McCarren and DeLong, 2007; Rusch et al., 2007) and believed to mediate photoheterotrophy in at least some of the diverse microbes that encode it (Sabehi et al., 2005; Gomez-Consarnau et al., 2007; Moran and Miller, 2007; Stingl et al., 2007; Gonzalez et al., 2008). PR-containing targets accounted for 50% of the taxa (11 of 22) abundant in shallow samples (Fig. 4). Specifically, all three abundant SAR86 targets encoded PR, thought in this clade to allow photoheterotrophy (Béjà et al., 2000; Sabehi et al., 2004; Sabehi et al., 2005; Mou et al., 2007; Sabehi et al., 2007). In addition, seven *Proteobacterial* PR-containing targets without phylogenetic markers (designated *Proteobacteria* by BLAST-based identities) were among those abundant in shallow samples. Two of these had sufficiently inverted relative abundances at 0m and 30m to contribute to the differentiation of the two depths (Fig. 5a; EB000_39F01 in 0m, and EB000_39H12 in 30m).

In addition, three PR-containing targets (two without phylogenetic markers, and the NAC11-7 HTCC2255 genome) were among those with strong post-bloom responses. All three were also among the ten most abundant targets in pyrosequence data, in all three sequenced post-bloom samples (circled data points in Fig. 2). This 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. Alternatively, PR might have allowed these taxa to persist longer than other heterotrophs as the bloom waned, as has been hypothesized for the PR-containing *Bacteroidetes* cultivar *Dokdonia* sp. MED134 (Gomez-Consarnau et al., 2007). 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.

v. Dynamics surrounding upwelling and bloom events: Community composition variability did not obviously correlate to Monterey Bay's three typical "oceanographic seasons" (Fig. 4; spring/summer upwelling, fall upwelling, and winter non-upwelling, as defined in e.g. (Pennington, 2000; Pennington et al., 2007). However, there was substantial annual variability in the timing of the seasonal Davenport Upwelling Plume and associated upwelling events, and phytoplankton abundance and growth rates have previously been described as "strikingly pulsed" (Pennington, 2000). Conditions during the period sampled in this study did not follow the average seasonal breakpoints, so it is not surprising that there was little apparent correlation between sample profiles and the site's typical oceanographic seasons. Ordering the samples temporally, instead of clustering them, also did not reveal appreciable seasonal dynamics of most targets (Fig. S10). Profiling of additional years, or at higher temporal resolution, might reveal a stronger cumulative seasonal signal.

Despite the lack of a strong seasonal signal overall, the array profiles showed responses to upwelling . Following some upwelling events (as indicated by nitrate concentrations, Fig. 3), 0m array profiles were notably intense (red starred samples in Figs. 4 and S10, and denoted by blue arrows in Fig. 3), reflecting high target abundances, and these upwelling-influenced profiles are more similar to each other than to most other 0m or 30m samples (as reflected in branch lengths between samples, Fig. 4). When samples are ordered temporally (Fig. S10) the seasonal nature of this response to particular spring and fall upwelling events captured by the 21 sampled dates is clear.

The phytoplankton blooms associated with upwelling are distinct between spring and fall upwelling events in Monterey Bay (Pennington et al., 2007), but this difference is not reflected in the microbes profiled by the array; the post-upwelling profiles do not cluster into two distinct groups based on upwelling season. Thus, for the taxa targeted by the array, there were not recurring post-bloom communities specific to spring or fall blooms.

The post-upwelling signature in the array data was therefore 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 target responses came from *shallow-consistent* or *-frequent* genotypes, including four NAC11-7 targets (EB080_L11F12, EB080_L43F08, EB080_ L27A02, and HTCC2255) and two PR-containing alphaproteobacterial clones lacking phylomarkers (EB000_39F01, EB000_55B11). The NAC11-7 *Roseobacteria* clade is often associated with bloom and post-bloom conditions (West et al., 2007, and reviewed in Buchan et al., 2005), due to their common ability to degrade dimethylsulfoniopropionate, an osmolyte produced by a variety of phytoplankton. The prominent role of NAC11-7 signal at this coastal upwelling site, and their particular intensity after bloom conditions, is therefore consistent with previous observations of this clade. An additional *shallow-frequent* genotype with dramatic increase in post-bloom intensity was a representative (EB000_36A07) of the betaproteobacterial OM43 clade, which has been observed to respond to diatom blooms (Morris et al., 2006), occurring in Monterey Bay during the spring/summer upwelling (Pennington et al., 2007). Given that the OM43 clade appears methylotrophic (Giovannoni et al., 2008), this reinforces the association between phytoplankton blooms and one-carbon compound degraders.

Responses to upwelling were also observed at 200m. The chemical signatures of upwelling and subsequent surface bloom events were observed in patterns in nitrate, phosphate and silicate concentrations at 200m (Fig. 3). 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 sinking and remineralization 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 can be consumed before it is converted to nitrate, and existing nitrate is also taken up by the actively degrading community. 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. It is possible 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.

vi. A window into population heterogeneity: In addition to tracking targeted taxa, the genome proxy array design allows the tracking of close relatives of targeted strains, and through the pattern of probe hybridization can reveal population shifts over time. Population shifts were examined in two ways. First, the relative evenness of the array hybridization signal to each probe-set was examined (see Rich et al., 2008, and Methods) as a measure of the relative identity of the hybridizing genotype to the target genotype. The signal across probe sets from sporadically-distributed taxa was less even than from depth-consistent taxa. It was also less even for common deep taxa compared to common shallow taxa (Fig. S11). Second, for particular targets of interest, the hybridization pattern of signal across the probe set was compared between samples. Specifically, pair-wise correlations (Pearson) of these hybridization patterns were calculated between samples. Clustering of these correlations was then used to identify samples with more or less similar probeset patterns for a given target. This process is shown for a targeted SAR86-II clone in Figure 7, and represents complementary approaches for analyzing probe signal. Averaging the signal across all probes for a given target describes the relative abundance of hybridizing genotypes, while assessing the evenness of that signal across probes (the hybridization pattern) indicates the likely genetic relatedness of hybridizing strains to the target. Then, the similarity of hybridization pattern between different samples indicates potential shifts in hybridizing populations.

As an example, all samples in which SAR86-II clone EB000_45B06 occurred (39 total; 21 samples at 0m, 13 at 30m and 5 at 200m) showed similar hybridization evenness (see Methods). This implied similar overall identities to the targeted strain. Analysis of hybridization patterns, however, suggested the presence of four distinct populations (Fig. 7). Three of these four potential populations had cohesive occurrence patterns (occurring primarily at one depth; Fig. 7), supporting their probable existence and ecological relevance.

These results suggest the power of the genome proxy array platform to dissect fine population structure. This could be further examined by comparing the population structure of array-targeted clones to metagenomic sequence data, and will be explored in follow-up work.

Potential future use of the genome proxy array

The relative value of array versus sequencing approaches for profiling microbial communities cuts across three common research goals : (i) *Overall community profiling* ex situ: It is currently ~100-fold less expensive to repetitively characterize samples using a genome proxy array than by even the most inexpensive metagenomic methods (e.g., Illumina sequencing), and requires a fraction of the computational resources for data processing. While the array provides indirect information (hybridization patterns and intensity) on targeted genotypes and their relatives, metagenomics provides direct information about the entire community where database matches allow such inference. (ii) *Community profiling* in situ: A variety of autonomous sensors exist to perform rapid community profiling by optical (e.g. Sieracki et al., 1998; Olson and Sosik, 2007; Thyssen et al., 2008) or nucleic acid hybridization (e.g. Scholin et al., 2001; Roman 2005) methods. The former discern only those few microbes with distinctive optical features. The latter currently target the 16S rRNA molecule (Preston et al., 2009), although organisms with highly similar 16S sequences can have distinct ecological niches (e.g. Rocap et al., 2003; Konstantinidis and Tiedje 2005). Thus the genome proxy array approach might serve a unique methodological role on such autonomous sensors. (iii) *Population profiling*: The genome proxy array can also discern closely-related populations (see above), effectively assaying both gene content and average nucleotide identity across targeted regions in related genotypes. While metagenomic data can provide population inferences, these have been limited to cases where assemblies are possible (e.g., low-diversity environments, Tyson et al., 2004, or dominant taxa in more complex communities, Venter et al., 2004), or to small sequence reads that represent ~40-fold less of the genome than the genome proxy array. Thus, for now, the genome proxy array retains utility as an *ex situ* community profiling tool, and complements sequencing for applications of *in situ* profiling and population tracking.

Conclusions

Exploration of the array profiles and the underlying causes of their variability allowed a cost-effective understanding of target natural history, and of community dynamics over time. Thus far, we tracked the genotype abundances of 268 target taxa through 57 samples collected over four years in Monterey Bay, at three oceanographically-distinct depths (Fig. 3). While the targets were distributed across known marine microbial diversity and had diverse geographic origins, 95 targeted taxa were present in at least one sample, and 31 were present in >50% of samples. Most taxa showed differential distribution with depth (Fig. 4). 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 chemoautotrophic 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. Finally, the array suggested that some targets were present as multiple distinct populations over time and space; these population dynamics suggest new directions for future research on microbial population dynamics.

Methods

Sampling and DNA Extractions: Samples were collected from Station M1 (36.747˚ N, 122.022˚ W) in Monterey Bay at approximately monthly intervals, with several longer gaps, between JD271 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 1.6µm-pore prefilter (Whatman) and concentrated onto a 25mm Supor-200 0.2µm-pore filter (Pall Corp, Ann Arbor, MI), using a MasterFlex peristaltic pump system (Cole-Parmer Instrument Company, Vernon Hills, IL) at ≤15psi. Filters were stored dry in 2ml screw-cap tubes, immediately placed in a -20°C freezer shipboard, and transferred on ice to a -80°C freezer upon landfall.

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. All MB DNA extractions were performed simultaneously in 96-well format to minimize extraction variability, as in (Rich et al., 2008). Briefly, cell lysis was performed by incubating each filter with 242ml lysis buffer (lysis buffer: 40 mM EDTA, 50 mM Tris pH 8.3, 0.73 M sucrose, 1.15 mg ml^{-1} lysozyme, 200 mg ml^{-1} RNase, 0.2 mm-filter-sterilized) in a microcentrifuge tube at 37°C for 30 min, rotating. Protein degradation was accomplished by adding SDS to 1%, and 13.5ml Proteinase K solution (10 mg ml^{-1} in 40 mM EDTA, 50 mM Tris pH 8.3, 0.73 M sucrose), and incubating overnight at 55°C, rotating. DNA was then extracted with the DNeasy 96 Tissue kit (Qiagen, Valencia, CA), using modifications of the manufacturer's protocol. Each tube was vortexed with 300ml of Buffer AL and incubated at 70°C for 10 min, then vortexed with 300ml of 99% ethanol and pipetted onto a 96-well spin plate. The plate was sealed with an airpore sheet (supplied with kit) and spun at 40°C, 4612 x *g* in a Sorvall Legend RT centrifuge (Kendro Laboratory Products, Newtown, CT). After a 10 min spin 500 ml Buffer AW1 was added to each well, the plate was re-sealed and spun 5 min, then 500 ml Buffer AW2 was added to each well, and the plate was re-sealed and spun 5 min. Columns were then incubated for 15 min at 70°C atop a new rack of elution microtubes RS (supplied with kit). DNA was eluted with 2 x 200 ml Buffer AE preheated to 70°C, incubated 1 min, and spun 2 min. Finally, DNA was concentrated by Excela-Pure 96-well PCR purification kits (Edge BioSystems, Gaithersburg, MD), following the manufacturer's protocol. DNA was rinsed with 100 ml nuclease-free water, resuspended in 20 ml dilute TE (1 mM Tris pH 8, 0.1 mM EDTA pH 8), and transferred to a clean 96-well plate. Extracted DNAs were quantified spectrophotometrically (Nanodrop, Thermo Scientific) and stored at -80°C 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 et al., 2008), and include the 75m DNA sample used in that study. DNA was extracted as described in (Frias-Lopez 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). Nutrient (nitrate, nitrite, silicate and phosphate) data used for correlation analyses are in Supplemental Table S4, and additional plots can be accessed at http://www.mbari.org/bog/.

Arrays Design, Hybridization, and Data Processing: The expanded genome proxy array was designed as in (Rich et al., 2008). Briefly, each genotype was targeted using suites of ~20 70-mer oligonucleotide probes designed using the program ArrayOligoSelector (Zhu et al., 2003). Probes had approximately the same %GC (40%) and were distributed across the target genome or genome fragment, with no more than one probe per gene and avoiding 16S and 23S rRNA genes. The array included positive and negative control probes designed using the same method, to *Halobacterium salinarum* NRC-1 and a random genome sequence, respectively.

The expanded array had a broader scope than the prototype of Rich et al., 2008 (268 target genotypes, as opposed to the prototype's 14) and included a co-spot oligo for spot alignment and gridding purposes (using the "alien" oligo sequence of (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 Hawaii environmental BAC- and fosmid-based genomic libraries. Targeted genotypes are detailed in Table S1, summarized in Table S2, and presented in a schematic phylogenetic overview in Fig. 1. Previously-unpublished sequences used for array design were submitted to Genbank under accession numbers GU474833-GU474949.

Hybridizations were performed as in (Rich et al., 2008), by labeling randomly-amplified sample DNA with a single fluorophore (Cy3) for hybridization. The following modifications were made to the Rich et al., 2008, hybridization method: Round A, B and C amplification 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). For each sample, at least three replicate arrays were hybridized. (As arrays constructed in-house, some did not produce high quality data due to significant surface peeling of the poly-lysine coating during hybridization or excessive background fluorescence; ~20% of arrays were discarded and additional arrays were hybridized.)

Data were pre-processed as in (Rich et al., 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.

Finally, the signal for each targeted genotype was calculated. To be considered present, at least 40% of its 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 in Rich et al., 2008), or the targeted genotype was considered "absent" and its value set to zero. This was done to remove erroneous target abundances due to uninformative single-gene cross-hybridizations. For targets that passed this thresh-holding step, the mean or tukey biweight (TBW) across each probe set was taken, as in (Rich et al., 2008). We did not examine which probes for each organism showed signal, since probes were not designed to distinguish particular genes; i.e., no alignments were used to target conserved or variable parts of given genes, but instead the probe was chosen purely on hybridization characteristics.

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

Data Analyses: Clustering analyses of sample hybridization data were performed in GenePattern (Reich, 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. 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 0m samples were chosen for parallel pyrosequencing and array hybridization, based on their DNA yields. Approximately 3µg each of samples 2000 JD298, 2001 JD115 and 2001 JD135 were sequenced at the Schuster Lab pyrosequencing facility (Pennsylvania State University) on a GS-FLX DNA sequencer (454 Life Sciences, Brandford, CT).

Sequence Clean-Up: To remove poor quality pyrosequences, the length distribution of the raw 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 303bp for 2001JD135). The outlying 5% of the reads were removed. 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 et al., 2008). The genomes were fragmented *in silico* 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 self-self (MED4 to MED4, 100% identity), MIT9515-vs-MED4 (86% average genomic identity, calculated as in (Konstantinidis and Tiedje, 2005), and MIT9312-vs-MED4 results (78.5% average genomic identity). A variety of command-line BLASTN parameters were tested for similarity of results to those of the array: 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. The first parameter set (X150 q-1 r1 W7 FF) yielded the best separation of the distribution of MED4-MED4 hits from MED4-MIT9515 and MED4-

MIT9312 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 et al., 2008), the equation for that comparison was used to adjust the bit score to the read length for each individual read.

Monterey Bay pyrosequencing versus array comparison: Using the BLASTN parameters and parsing criteria optimized above, the reads from each pyrosequenced Monterey Bay sample were BLASTed against all 268 genomes and genome fragments to which the array was targeted. Reads were assigned to (i.e., 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 recruitment abundance of each array target was then normalized to the length of the target query, and to the database size. For each of the three samples, the pyrosequence-based abundances of each genotype were then compared to the array-based abundances. Despite a full plate of sequencing per sample, recruitment of reads to each target was insufficient to screen presence/absence based on the signal evenness across each target, a standard step in the array data analysis pipeline. Therefore, unthresholded array data without the evenness filter (that is, the signal for each organism before requiring at least 40% of its probes to be above the described threshold) were compared to pyrosequencing data for each target genotype.

Acknowledgements

We gratefully acknowledge the captain and crew of the R.V. *Point Lobos* for expert assistance at sea and Drs. Christina Preston and Lynne Christianson for sample collection over four years. We also thank Francisco Chavez and Reiko Michisaki of the MBARI Biological Oceanography Group for the corresponding chemical oceanographic time-series data. Lastly, we thank Matt Sullivan and three anonymous reviewers for helpful comments on the manuscript. This work was supported by grants to EFD from the Gordon and Betty Moore Foundation, a National Science Foundation award EF 0424599 (C-MORE), NSF Microbial Observatory Award MCB-0348001, and the Office of Science (BER) U.S. Department of Energy.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

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. Monterey, CA: Naval Postgraduate School.

Bano, N., and Hollibaugh, J.T. (2002) Phylogenetic composition of bacterioplankton assemblages from the Arctic Ocean. Appl Environ Microbiol 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.

Buchan, A., Gonzalez, J.M., and Moran, M.A. (2005) Overview of the marine roseobacter lineage. Appl Environ Microbiol 71: 5665-5677.

Dalsgaard, T., Canfield, D.E., Petersen, J., Thamdrup, B., and Acuna-Gonzalez, J. (2003) N_2 production by the anammox reaction in the anoxic water column of Golfo Dulce, Costa Rica. Nature 422: 606-608.

DeLong, E. F. (1992) Archaea in coastal marine environments. Proc Natl Acad Sci U S A 89: 5685-5689.

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.

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.

Eilers, H., Pernthaler, J., Glockner, F.O., and Amann, R. (2000) Culturability and *in situ* abundance of pelagic bacteria from the North Sea. Appl Environ Microbiol 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. Proc Natl Acad Sci U S A 95: 14863-14868.

Field, K.G., Gordon, D., Wright, T., Rappé, M., Urback, E., Vergin, K., and Giovannoni, S.J. (1997) Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. Appl Environ Microbiol 63: 63-70.

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. Proc Natl Acad Sci U S A 105: 3805-3810.

Fuhrman, J. A., McCallum, K., and Davis, A. A. (1992) Novel major archaebacterial group from marine plankton. Nature 356: 148-149.

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. Proc Natl Acad Sci U S A 103: 13104-13109.

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). New York, NY: Wiley and Sons, pp. 47-84.

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 methylotrophs. Environ Microbiol 10: 1771-1782.

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

González, J.M., Fernández-Gómez, B., Fernàndez-Guerra, A., Gómez-Consarnau, L., Sánchez, O., Coll-Lladó, M. et al. (2008) Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (Flavobacteria). Proc Natl Acad Sci U S A 105: 8724-8729.

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 ccean. Science 314: 649-652.

Karl, D.M. (1999) A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. Ecosystems 2: 181-214.

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

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.

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

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

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.

Könneke, M., Bernhard, A.E., de la Torre, J.R., Walker, C.B., Waterbury, J.B., and Stahl, D.A. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437: 543-546.

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

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.

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 Microbiol Ecol 36: 193-202.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, *et al.*. (2004) ARB: a software environment for sequence data. Nucleic Acids Res 32:1363-1371.

Marhaver, K.L., Edwards, R.A., and Rohwer, F. (2008) Viral communities associated with healthy and bleaching corals. Environ Microbiol 10: 2277-2286.

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. Appl Environ Microbiol 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. Environ Microbiol 9: 846-858.

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. Environ Microbiol 9: 1162-1175.

Moran, M.A., and Miller, W.L. (2007) Resourceful heterotrophs make the most of light in the coastal ocean. Nat Rev Microbiol 5: 792.

Morris, R.M., Rappé, 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.

Morris, R. M., Rappe, M. S., Urbach, E., Connon, S. A., and Giovannoni, S. J. (2004) Prevalence of the *Chloroflexi*-related SAR202 bacterioplankton cluster throughout the mesopelagic zone and deep ocean. Appl Environ Microbiol 70: 2836-2842

Morris, R., Vergin, K., Cho, J.-C., Rappé, 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. Limnol Oceanogr 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. Environ Microbiol 8: 1361-1370.

Mou, X., Hodson, R.E., and Moran, M.A. (2007) Bacterioplankton assemblages transforming dissolved organic compounds in coastal seawater. Environ Microbiol 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.

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 Paciic bacterioplankton communities. Limnol Oceanogr 40: 148-158.

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. Environ Microbiol 10: 1526-1535.

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. Appl Environ Microbiol 71: 697-705.

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

Partensky, F., Blanchot J. and Vaulot, D. (1999) Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. In Marine cyanobacteria and related organisms. Charpy, L. and Larkum, H. (eds). Monaco: Musée océanographique. Bulletin de l'Institut Océanographique (Monaco) NS**19**: 431-449.

Pennington, J.T., and Chavez, F.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 Res Part 2 Top Stud Oceanogr 47: 947-973.

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.

Preston, C.M., Marin 3rd, R., Jensen, S.D., Feldman, J., Birch, J.M., Massion, E.I. *et al.* (2009). Near real-time autonomous detection of marine bacterioplankton on a coastal mooring in Monterey Bay, California, using rRNA-targeted DNA probes. Environ Microbiol 11: 1168–1180.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B., Ludwig, W., Peplies, and Glöckner. F. O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188-7196.

Rappé, 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 Microbiol Ecol 33: 219-232.

Reich M, L.T., Gould, J., Lerner ,J., Tamayo, P., and Mesirov, J.P. (2006) GenePattern 2.0. Nat Genet 38: 500-501.

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

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.

Roman, B., Scholin, C., Jensen, S., Marin, R., Massion, E., and Feldman, J. (2005) The 2nd generation environmental sample processor: evolution of a robotic underwater biochemical laboratory. In OCEANS 2005 MTS/IEEE Conference. Washington DC: Marine Technology Society.

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.

Rusch, D.B. (2008) *Prochlorococcus* variants from the GOS metagenome. In Prochlorococcus 20th Anniversary Colloquium. Cambridge MA: Massachusetts Institute of Technology.

Rusch, D.B., Martiny, A., Dupont, C.L., Halpern, A.L., and Venter, J.C. (2010) Metagenomic characterization of novel Prochlorococcus clades from iron depleted oceanic regions. In The 2010 Genomic Science Contractor-Grantee and Knowledgebase Workshop. Arlington VA: Department of Energy, Office of Biological Environmental Research.

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

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 Biol 3: e273.

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

Scholin, C.A., Massion, E.I., Wright, D., Cline, D., Mellinger, E., and Brown, M. (2001) Aquatic Autosampler Device. US patent 6187530.

Sieracki, C.K, Sieracki, M.E., and Yentsch, C.S. (1998) An imaging-in-flow system for automated analysis of marine microplankton. Mar Ecol Prog Ser 168: 285–296.

Stevens, H., and Ulloa, O. (2008) Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. Environ Microbiol 10: 1244-1259.

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 J 1: 361-371.

Suzuki, M.T., Béjà, O., Taylor, L.T., and Delong, E.F. (2001a) Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. Environ Microbiol 3: 323-331.

Suzuki, M. T., Preston, C.M., Chavez, F.P., and DeLong, E.F. (2001b). 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.

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. J Plankton Res 30: 333-343.

Treusch, A. H., Vergin, K. L., Finlay, L.A., Donatz, M.G., Burton, R.M., Carlson, C.A., Giovannoni, S.J. (2009). Seasonality and vertical structure of microbial communities in an ocean gyre. ISME J 3(10): 1148-1163.

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.

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 Biol 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.

Walsh, D.A., Zaikova, E., Howes, C.G., Song, Y.C, Wright, J.J., Tringe, S.G. et al. (2009) Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. Science 326: 578-582.

Ward, B.B. (2005) Temporal variability in nitrification rates and related biogeochemical factors in Monterey Bay, California, USA. Mar Ecol Prog Ser 292: 97-109.

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

Waterbury, J.B., Watson, S.W., Valois, F.W., and Franks, D.G. (1986) Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. Can Bull Fish Aquat Sci 214: 71-120.

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*. Environ Microbiol 9: 2707-2719.

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. Biol Direct 2: 27.

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. Appl Environ Microbiol 63: 1441-1448.

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 Biol 5: e16.

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

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 Res Part 2 Top Stud Oceanogr 49: 3017-3038.

Figure Legends and Tables

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 the SILVA database Release 99 (Pruesse et al., 2007) with the ARB software package (Ludwig et al., 2004).

Figure 2. Cross-comparison of array- and pyrosequence-based target abundances for three MB samples; p-values associated with each linear regression were <0.0001. Using BLASTN parameters optimized to mimic array cross-hybridization, all 268 targeted genomes and genome fragments were compared (using BLAST) to the pyrosequence data derived from identical samples. 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. Circled datapoints indicate proteorhodopsin-containing clones abundant by array signal post-upwelling as described in the text: red circles = EB000_55B11, orange circles = EB000_39F01, and pink circles = Rhodobacterales HTCC2255.

Figure 3. Sample origin from Monterey Bay Station M1 over depth (y-axis) and time (x-axis) against the backdrop of oceanographic context. The 57 samples (black diamonds) hybridized to the array derive from three depths (0, 30 and 200m) over ~ 4 years; time (with months indicated by their first-letter designations) is indicated along the X-axis. The 0m samples used for cross-validation pyrosequencing

are indicated by red stars. Panels show temperature, nitrate, nitrite, silicate and phosphate concentrations. 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 phosphate.

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. Targets are color-coded by phylogenetic identity, gene content of particular interest (note column indicating presence/absence of 16S rRNA gene), and origin (see color legend; MB = Monterey Bay, HOT = Hawaii Ocean Time series). Intensity of yellow-to-red color for each genotype and sample date indicates relative target signal; note that relative abundance is quantitative for each genotype between samples but not between genotypes. 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 clusters is indicated by the blue vertical dashed line. Abundant targets referred to in the text are boxed with dashed lines, “shallow-consistent” = red, “shallow-frequent” = green, “deep-consistent” = purple, “deep-frequent” = navy. Red asterisks denote samples with particularly intense 0m profiles; the 30m and 200m samples for the same dates, when available, are indicated by blue asterisks.

Figure 5. Canonical discriminant analysis (c.d.) of Monterey Bay sample (0m **O** , 30m **+** , and 200m △) array data, with parameter correlations to c.d. axes indicated by vector length and direction. Diamonds designate center of each depth’s data cloud. (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. Target taxonomic affiliations (by 16S identity, or by clone BLAST hits for clones with no 16S rRNA gene): EB000_39F01 = putative *Alphaproteobacteria*, ProMED4 = *Cyanobacteria;Prochlorococcus*, EB080_L43F08 = *Alphaproteobacteria;Rhodobacterales;NAC11-7*, HTCC2255 = *Alphaproteobacteria;Rhodobacterales;NAC11-7*, EB080_L27A02 =

Alphaproteobacteria;Rhodobacterales;NAC11-7, EB750_ 01B07 = putative *Deltaproteobacteria*, EB750_ 10B11 = *Gammaproteobacteria;related to S-oxidizing symbionts*, EB080_ L31E09 = *Gammaproteobacteria;ARCTIC96BD-19 clade, S-oxidizing symbiont relative*, EB000_39H12 = putative *Proteobacteria*, EBAC_27G05 = *Gammaproteobacteria;SAR86-III*, EB000_65A11 = *Gammaproteobacteria;EB000_65A11 clade*.

Figure 6. 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 designated by its month and year. (a) 0m samples; the sample variability among 0m samples is not strongly correlated to differing nutrient concentrations. (b) 30m samples; there is a strong correlation to all four nutrients, reflecting the 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 7. Revealing population heterogeneity by the genome proxy array: complementary probeset analyses moving from overall target abundance to 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) Relative evenness of hybridization signal across the SAR86-II target EB000_45B06 target probe set (as Tukey biweight-over-mean value; see Methods). By this index alone, subpopulations are not strongly evident, (c) Pair-wise Pearson correlations of the signal pattern across the EB000_45B06 probeset, between every sample in which it occurred. Samples are clustered based on similarity of probeset 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.

Table 1: Array targets common in shallow or deep samples

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, known to have contaminated the HTCC2255 culture (http://www.roseobase.org/roseo/htcc2255.html). **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).

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, measured as the relative proportion of total array signal across all samples hybridized.

Figure S7. 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.

Figure S8. Clustering of hybridizations by sample and by genotype, per Figure 4, using only the subset of the 30m samples definitively below the mixed layer depth (MLD). MLD is shown in Figure S7 and 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.). Excluding the 30m samples above the MLD does not result in discrete clustering of the 0m and 30m samples.

Figure S9. Array profiles for all targets within three common phylogenetic clades: (a) Roseobacter (b) SAR86 (c) SAR11.

Figure S10. Heatmap of array hybridizations with samples ordered chronologically, without clustering of samples (columns) or genotypes (rows). The break between the 2000-2002 and 2003-2004 sampling periods is indicated by the black vertical dashed line. Intensity of cell color indicates relative target signal for that genotype and sample date; note that relative abundance is quantitative for each genotype between samples but not between genotypes. Samples are named Depth_Year_CollectionDate, and are color-coded by by oceanographic season (see color legend and text). Red asterisks denote samples with particularly intense 0m profiles. Gray columns indicate no samples for that depth and date. (a) 0m samples, (b) 30m samples, (c) 200m samples, with the three depths vertically stacked.

Figure S11. Evaluating the genetic relatedness of community DNA hybridized to the array. On the left are mean organism signals as shown in Figure 4, 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).

Table S1: Array targets

Table S2: Array targets summarized by phylogenetic cluster

Table S3. Comparison of array with other broad taxonomic surveys of Monterey Bay.

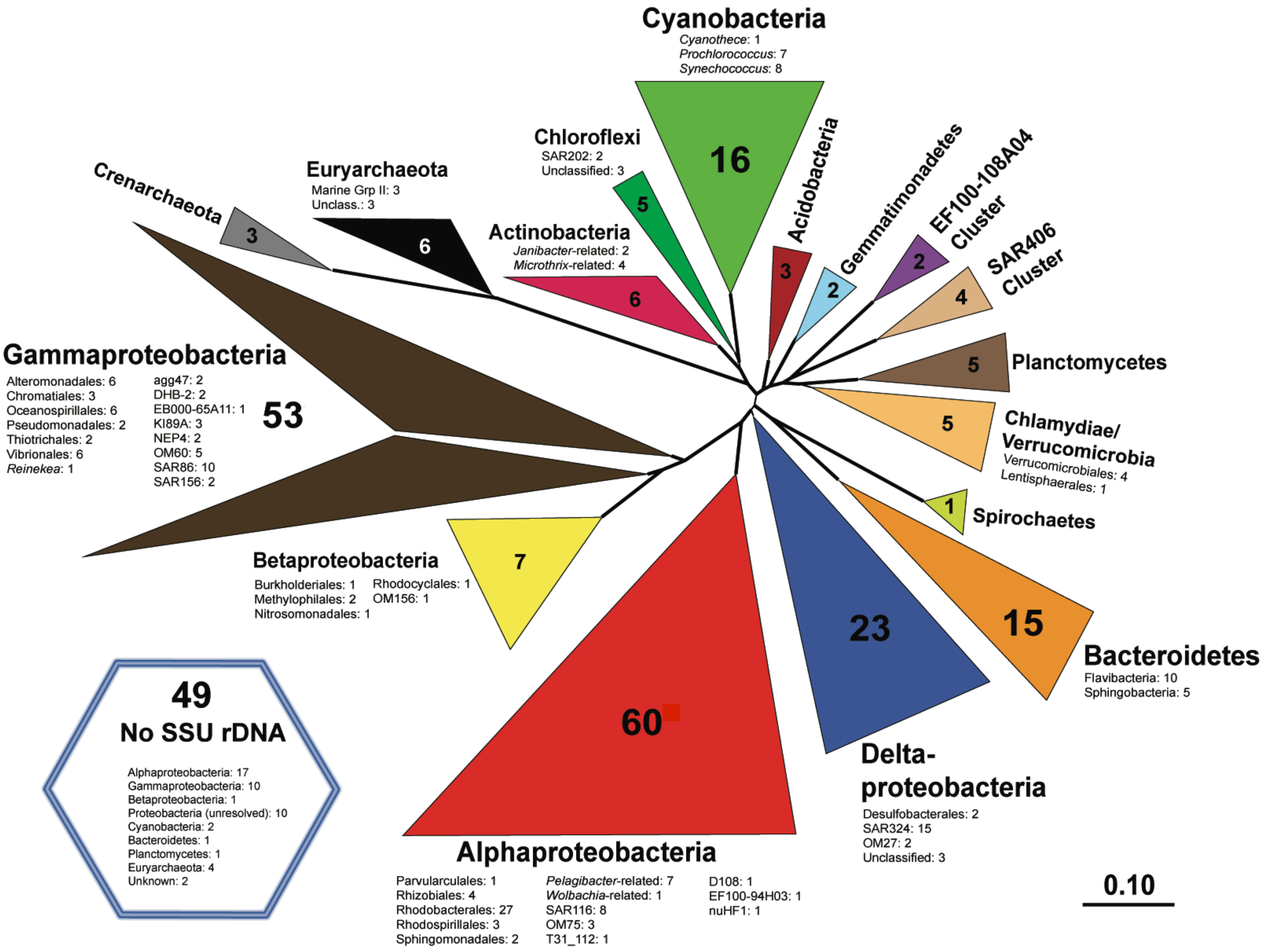
Cyanobacteria
Cyanothece: 1
Prochlorococcus: 7
Synechococcus: 8
16
Chloroflexi
SAR202: 2
Unclassified: 3
5
Euryarchaeota
Marine Grp II: 3
Unclass.: 3
6
Crenarchaeota
3
Actinobacteria
Janibacter-related: 2
Microthrix-related: 4
6
Acidobacteria
3
Gemmatimonadetes
2
EF100-108A04 Cluster
2
SAR406 Cluster
4
Planctomycetes
5
Chlamydiae/ Verrucomicrobia
Verrucomicrobiales: 4
Lentisphaerales: 1
5
Spirochaetes
1
Bacteroidetes
Flavibacteria: 10
Sphingobacteria: 5
15
Delta-proteobacteria
Desulfobacterales: 2
SAR324: 15
OM27: 2
Unclassified: 3
23
Gammaproteobacteria
Alteromonadales: 6
Chromatiales: 3
Oceanospirillales: 6
Pseudomonadales: 2
Thiotrichales: 2
Vibrionales: 6
Reinekea: 1
agg47: 2
DHB-2: 2
EB000-65A11: 1
KI89A: 3
NEP4: 2
OM60: 5
SAR86: 10
SAR156: 2
53
Betaproteobacteria
Burkholderiales: 1
Methylophilales: 2
Nitrosomonadales: 1
Rhodocyclales: 1
OM156: 1
7
Alphaproteobacteria
Parvularculales: 1
Rhizobiales: 4
Rhodobacterales: 27
Rhodospirillales: 3
Sphingomonadales: 2
Pelagibacter-related: 7
Wolbachia-related: 1
SAR116: 8
OM75: 3
T31_112: 1
D108: 1
EF100-94H03: 1
nuHF1: 1
60
49
No SSU rDNA
Alphaproteobacteria: 17
Gammaproteobacteria: 10
Betaproteobacteria: 1
Proteobacteria (unresolved): 10
Cyanobacteria: 2
Bacteroidetes: 1
Planctomycetes: 1
Euryarchaeota: 4
Unknown: 2
0.10

Figure 1

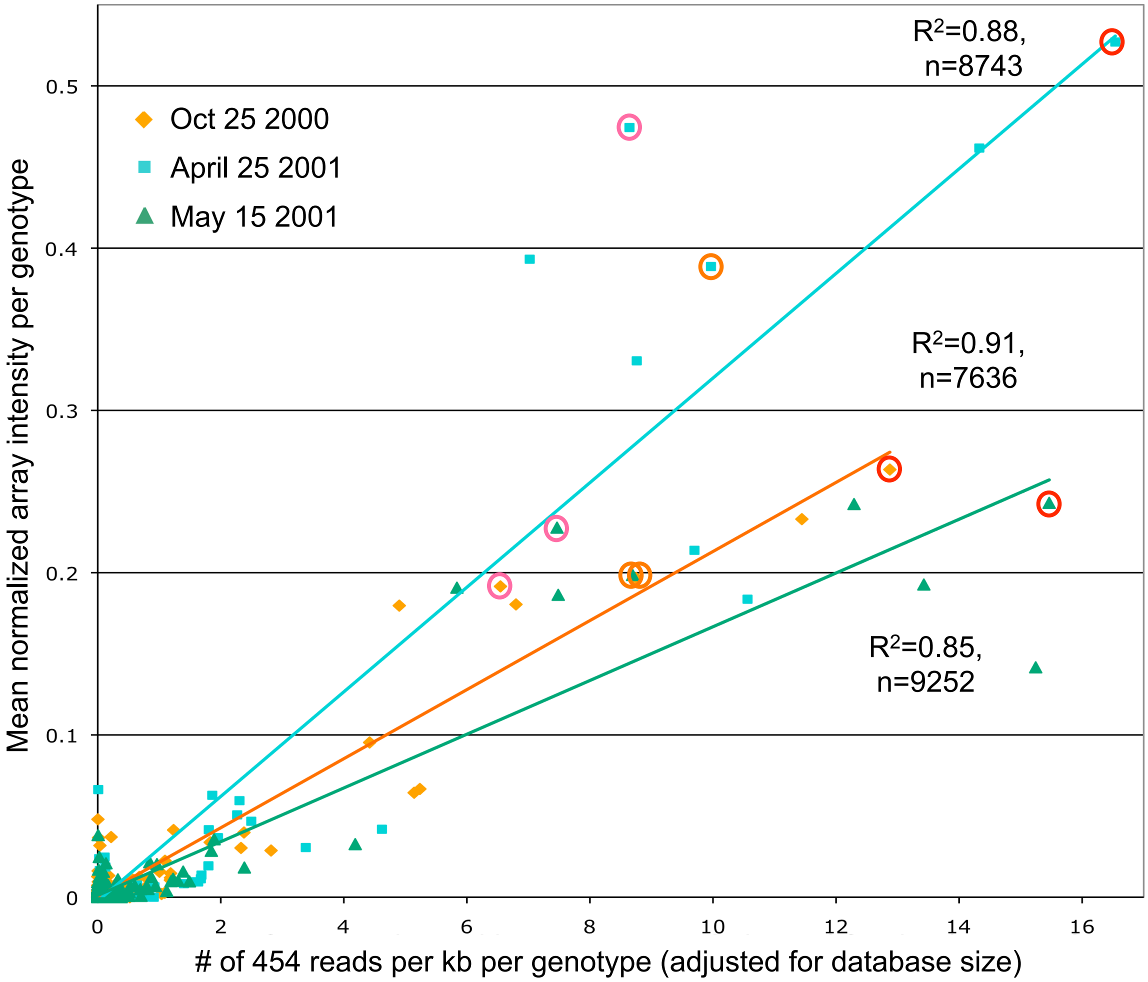
R2=0.88,
n=8743
R2=0.91,
n=7636
R2=0.85,
n=9252
Oct 25 2000
April 25 2001
May 15 2001
Mean normalized array intensity per genotype
0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
12
14
16
# of 454 reads per kb per genotype (adjusted for database size)

Figure 2

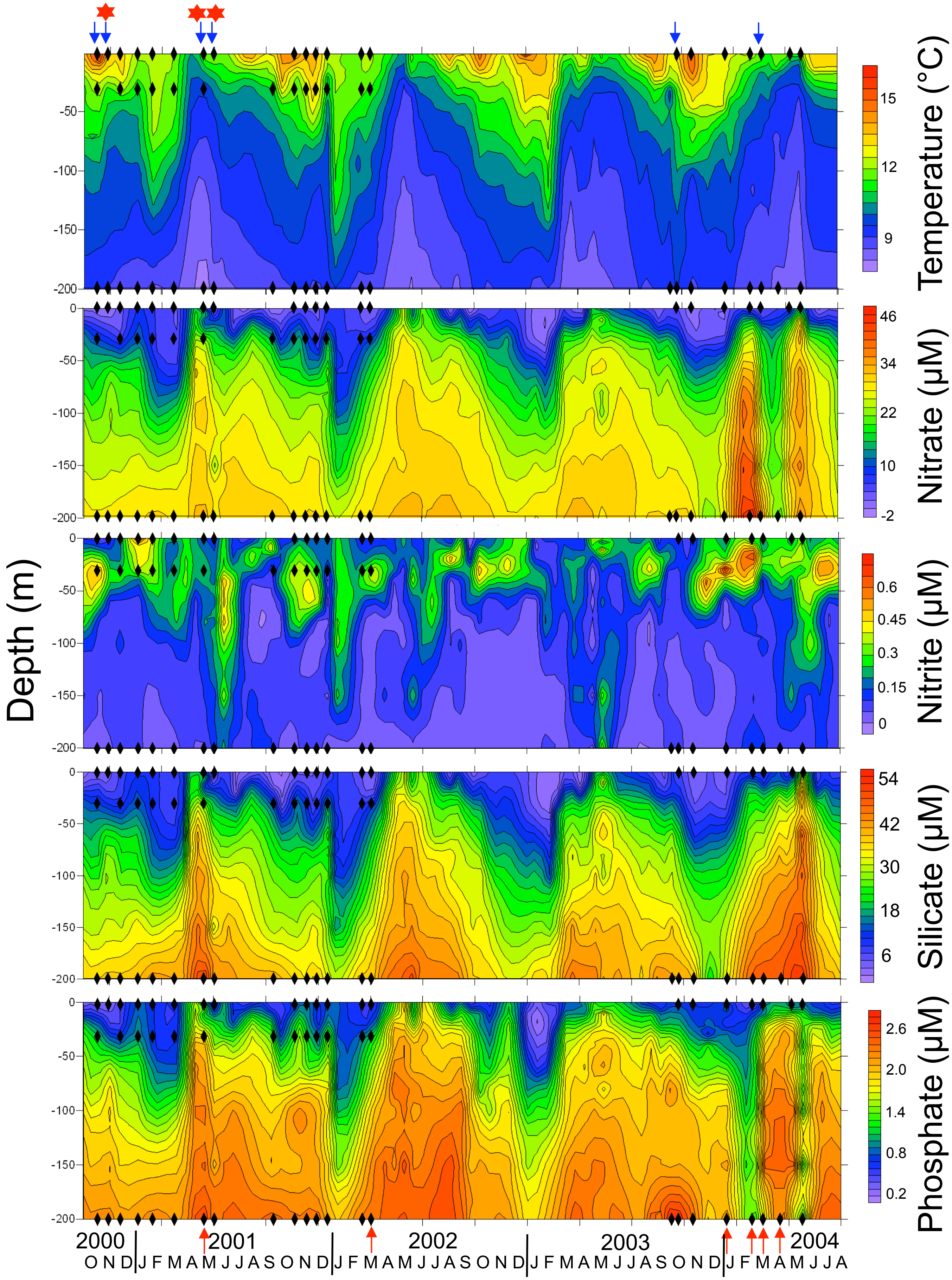
Depth (m)
-50
-100
-150
-200
0
Temperature (°C)
15
12
9
Nitrate (µM)
46
34
22
10
-2
Nitrite (µM)
0.6
0.45
0.3
0.15
0
Silicate (µM)
54
42
30
18
6
Phosphate (µM)
2.6
2.0
1.4
0.8
0.2
2000
2001
2002
2003
2004
O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A

Figure 3

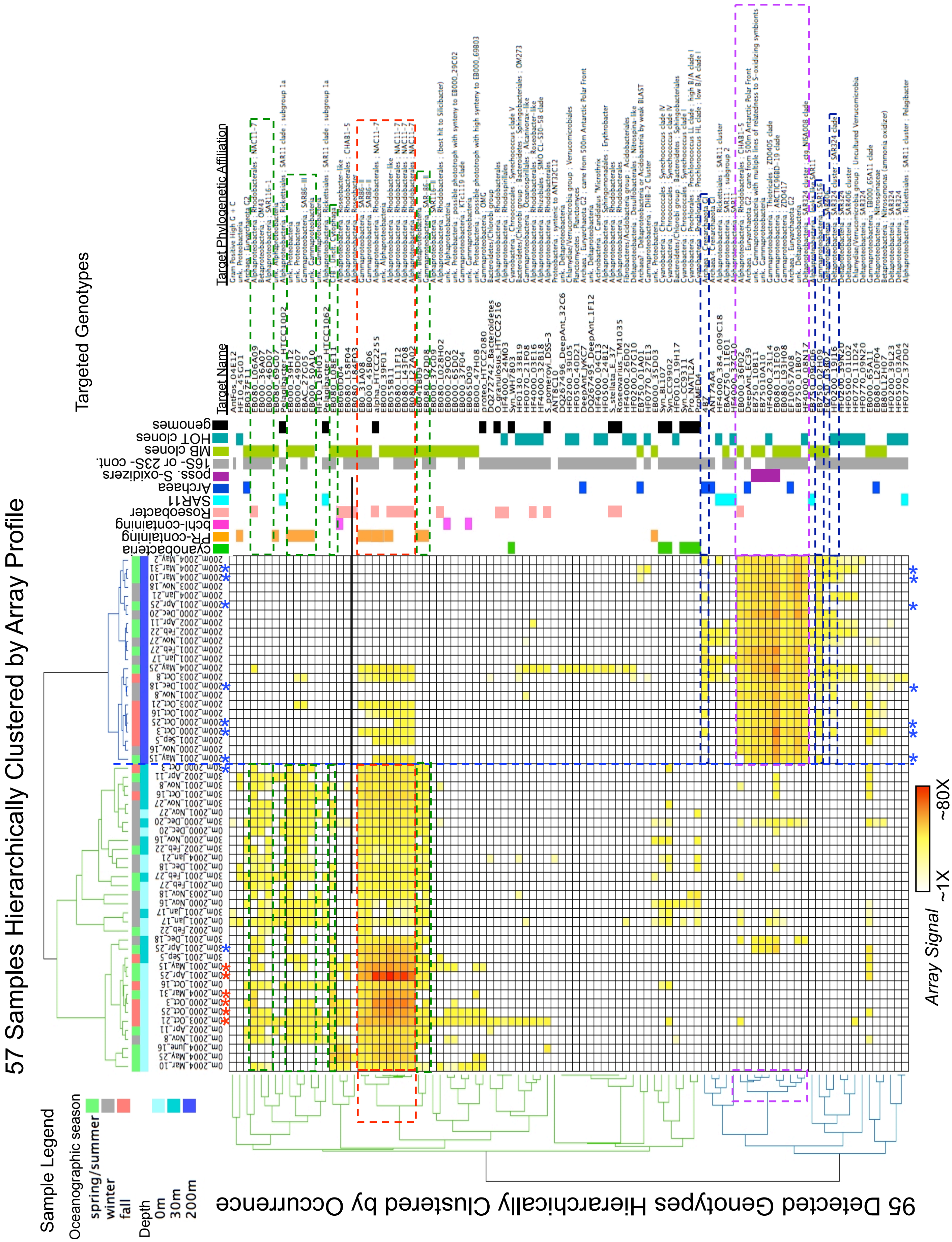
57 Samples Hierarchically Clustered by Array Profile
95 Detected Genotypes Hierarchically Clustered by Occurrence
Targeted Genotypes
Target Name
Target Phylogenetic Affiliation
genomes
HOT clones
MB clones
16S- or 23S- cont.
poss. S-oxidizers
Archaea
SAR11
Roseobacter
bchl-containing
PR-containing
cyanobacteria
Array Signal
~1X
~80X
Sample Legend
Oceanographic season
spring/summer
winter
fall
Depth
0m
30m
200m

Figure 4

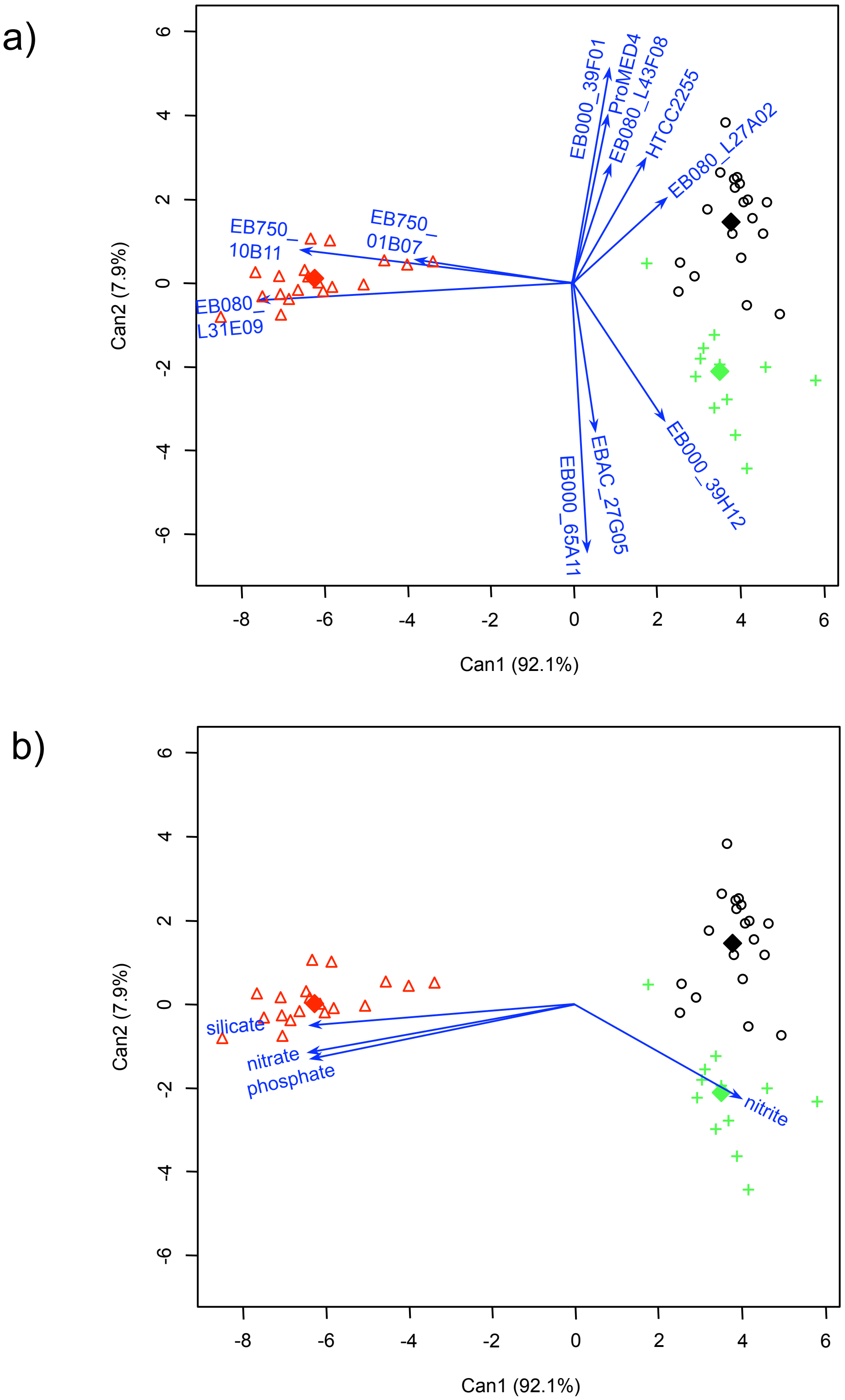
a)
EB000_39F01
ProMED4
EB080_L43F08
HTCC2255
EB080_L27A02
EB750_
10B11
EB750_
01B07
EB080_
L31E09
EBAC_27G05
EB000_65A11
EB000_39H12
Can2 (7.9%)
Can1 (92.1%)
6
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
6
b)
silicate
nitrate
phosphate
nitrite
Can2 (7.9%)
Can1 (92.1%)
6
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
6

Figure 5

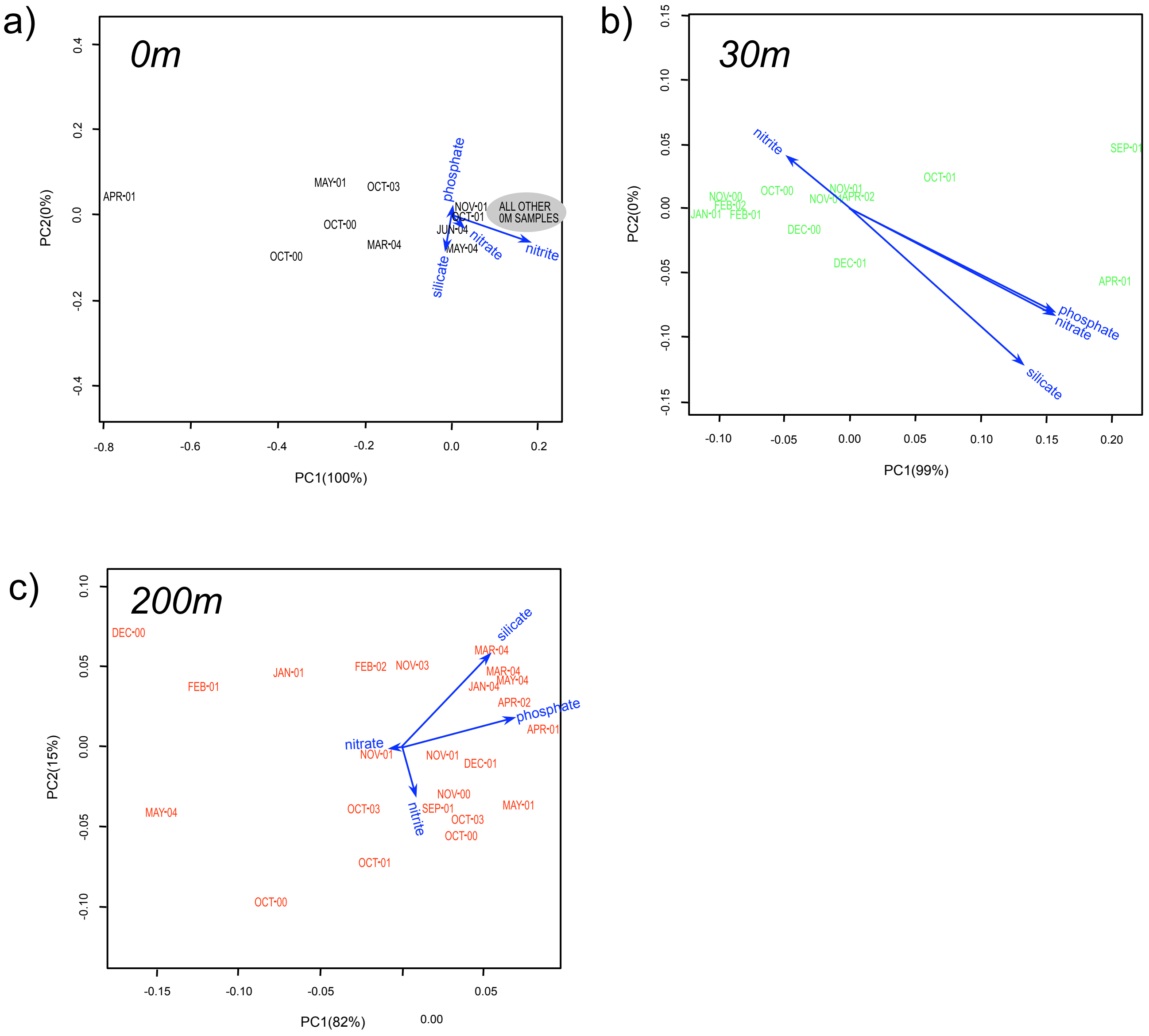
a)
0m
APR-01
MAY-01
OCT-03
OCT-00
MAR-04
OCT-00
phosphate
NOV-01
OCT-01
JUN-04
MAY-04
nitrate
nitrite
silicate
ALL OTHER
0M SAMPLES
PC2(0%)
0.4
0.2
0.0
-0.2
-0.4
-0.8
-0.6
-0.4
-0.2
0.0
0.2
PC1(100%)
b)
30m
nitrite
NOV-00
FEB-02
JAN-01
FEB-01
OCT-00
NOV-01
NOV-01
APR-02
DEC-00
DEC-01
OCT-01
SEP-01
APR-01
phosphate
nitrate
silicate
PC2(0%)
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15
0.20
PC1(99%)
c)
200m
DEC-00
FEB-01
JAN-01
FEB-02
NOV-03
silicate
MAR-04
MAR-04
MAY-04
JAN-04
APR-02
phosphate
APR-01
nitrate
NOV-01
NOV-01
DEC-01
MAY-04
NOV-00
OCT-03
SEP-01
nitrite
OCT-03
MAY-01
OCT-00
OCT-01
OCT-00
PC2(15%)
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.10
-0.05
0.05
PC1(82%)
0.00

Figure 6

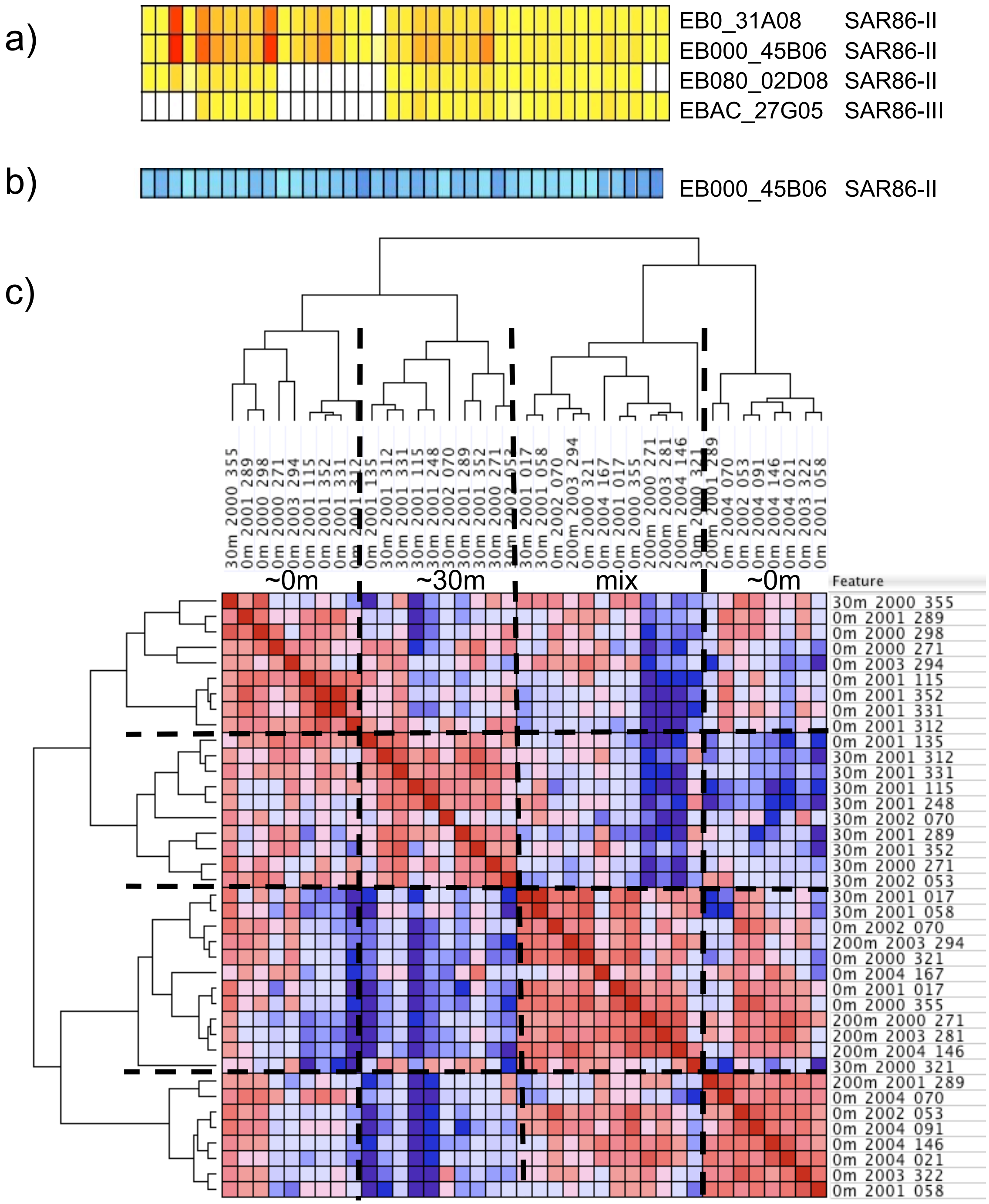
a)
EB0_31A08 SAR86-II
EB000_45B06 SAR86-II
EB080_02D08 SAR86-II
EBAC_27G05 SAR86-III
b)
EB000_45B06 SAR86-II
c)
~0m
~30m
mix
~0m
Feature
30m 2000 355
0m 2001 289
0m 2000 298
0m 2000 271
0m 2003 294
0m 2001 115
0m 2001 352
0m 2001 331
0m 2001 312
0m 2001 135
30m 2001 312
30m 2001 331
30m 2001 115
30m 2001 248
30m 2002 070
30m 2001 289
30m 2001 352
30m 2000 271
30m 2002 053
30m 2001 017
30m 2001 058
0m 2002 070
200m 2003 294
0m 2000 321
0m 2004 167
0m 2001 017
0m 2000 355
200m 2000 271
200m 2003 281
200m 2004 146
30m 2000 321
200m 2001 289
0m 2004 070
0m 2002 053
0m 2004 091
0m 2004 146
0m 2004 021
0m 2003 322
0m 2001 058

Figure 7