The seasonal structure of microbial communities in the Western English Channel

Jack A. Gilbert[1*], Dawn Field[2], Paul Swift[2], Lindsay Newbold[2], Anna Oliver[2], Tim Smyth[1], Paul J. Somerfield[1], Sue Huse[3] and Ian Joint[1]

[1] Plymouth Marine Laboratory, Prospect Place, Plymouth, PL1 3DH, UK. [2]NERC Centre for Ecology and Hydrology, CEH Oxford, Mansfield Road, Oxford, OX1 3SR, UK. [3]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA.

*Corresponding Author: Tel: +44 1752 633416, Fax: +44 1752 633101,

Email: jagi@pml.ac.uk

Running title: Marine bacterial seasonal succession

Summary

Very few marine microbial communities are well characterized even with the weight of research effort presently devoted to it. Only a small proportion of this effort has been aimed at investigating temporal community structure. Here we present the first report of the application of high-throughput pyrosequencing to investigate intra-annual bacterial community structure. Microbial diversity was determined for 12 time points at the surface of the L4 sampling site in the Western English Channel. This was performed over 11 months during 2007. A total of 182,560 sequences from the V6 hyper-variable region of the small-subunit ribosomal RNA gene (16S rRNA) were obtained; there were between 11,327 and 17,339 reads per sample. Approximately 7000 genera were identified, with one in every 25 reads being attributed to a new genus; yet this level of sampling far from exhausted the total diversity present at any one time point. The total data set contained 17,673 unique sequences. Only 93 (0.5%) were found at all time-points, yet these few lineages comprised 50% of the total reads sequenced. The most abundant phylum was *Proteobacteria* (50% of all sequenced reads), while the SAR11 clade comprised 21% of the ubiquitous reads and ~12 % of the total sequenced reads. In contrast, 78% of all OTUs were only found at one time-point and 67% were only found once, evidence of a large and transient rare assemblage. This time-series shows evidence of seasonally structured community diversity. There is also evidence for seasonal succession, primarily reflecting changes among dominant taxa. These changes in structure were significantly correlated to a combination of temperature, phosphate and silicate concentrations.

Introduction

Microbial communities are complex and highly diverse but are extremely important, being responsible for the vast majority of global biogeochemical cycling. The ocean is thought to contain approximately 1 x 10^{29} bacterial cells (Whitman et al., 1998). Given this large number, and that bacteria have been evolving for >3.5 billion years, it is probable that there is a vast array of different genotypes and phenotypes (Sogin et al., 2006); yet, because of this diversity, marine microbial communities are only well characterized in a few point locations in the vast ocean even with the weight of research effort presently devoted to it (Sogin et al., 2006).

Using the latest DNA sequencing technology, Sogin et al. (2006) recently highlighted the presence of a "rare biosphere" of marine bacteria. They demonstrated that a 1L sample of the deep-sea contained an estimated 20,000 different bacterial taxa. This estimate was derived from 118,000 sequences of a variable region of the small subunit ribosomal RNA gene (16S rRNA).

Studies such as the Global Ocean Survey (Rusch et al., 2007) have demonstrated how diverse bacterial communities can be in different marine provinces. Of the 7.7 million sequences obtained in the GOS study, ~57% were unique at an identity of 98%, suggesting a high degree of heterogeneity in the oceans (Rusch et al., 2007). Given the variable environmental conditions in different oceanic regions, it is reasonable to expect that individual sites will harbor diverse bacterial communities. Indeed, this has been demonstrated in many studies using rRNA as a taxonomic marker (e.g. Giovannoni et al, 1990; Delong, 1992; Armann et al., 1995; Pace, 1997; Huber et al, 2002; Rappé & Giovannoni, 2003; Hewson & Fuhrman, 2004). However, most PCR-cloning-based studies may have inherent bias. Sogin et al. (2006) used a simple method of high-throughput sequencing that reduces the biases associated with

cloning and normal-sequencing effort. They focused on a hyper-variable region of the 16S rRNA, to dramatically increase the number of amplicons sequenced to approximately 118,000, hence allowing less-dominant taxa to be identified.

There are two primary challenges, given the vast diversity of taxa in the seas: creation of a more complete inventory of this diversity and improving understanding of the possible controls of its structure. Here we address both questions by studying microbial communities over time at the Western Channel Observatory (WCO), located off the southern coast of the United Kingdom in the English Channel (http://www.westernchannelobservatory.org.uk/). The area is a transition zone, being at the northern boundary of many southern planktonic species and the southern boundary in range of northern species (Southward et al., 2005). The area is also one of the longest and best time-series in the world, having been samples since the early years of the 20th century (Southward et al., 2005). WCO is sampled weekly for a range of physicochemical and biological measures. During 2007, water samples were collected monthly for the molecular characterization of microbial communities. A primary aim of this work is to begin to integrate long-term environmental monitoring at this station with in-depth surveys of microbial communities, to better understand the mechanisms that shape bacterial diversity in this ecosystem.

Results and Discussion

Application of massively-parallel pyrosequencing of the 16S rRNA V6 region at a well characterized coastal observatory. Changes in microbial community structure over time were investigated in samples collected from 12 time-points (Table S1) at the WCO, using the massively parallel DNA sequencing methodology outlined by Sogin et al. (2006). After initial sequencing, sequences of low quality were

removed, including any sequence containing an ambiguous base ("N"), that was less than 50 nt, or that did not have a recognizable primer at each end (Huse et al., 2008), and the primers were trimmed from the sequences. The resulting dataset comprised a total of 182,560 sequences of the V6 hyper-variable regions of the small-subunit ribosomal RNA gene (16S rRNA). The number of sequence reads varied from 11,327 to 17,339 per sample. When these data were compared to terminal restriction length polymorphism (T-RFLP) analysis of the same samples, it was clear that there was a significant increase in resolution using the 16S-tag methodology, primarily because a larger number of ribotypes were identified (**Supplementary Information)**. Nevertheless, T-RFLPs is still useful as a general indicator of change, but this suggests that it may require significant methodological experience to provide realistic profiles for a particular environment.

High total diversity and a few ubiquitous taxa account for a high proportion of sequences. Overall, 17,673 unique operational taxonomic units (OTUs) were identified from the total data set of 182,560 sequence reads. Strikingly, only 0.5% of the OTUs were found at every time point. These were also highly abundant, contributing 54% of the total sequence reads. Only 22% of the OTUs were found on more than one occasion, but they contributed 90% of the read abundance. Rarefaction curves of individual and pooled samples show no evidence of a plateau being reached, even at 90% sequence identity (Fig. S1 A-D).

Clustering of unique tags confirms the presence of a very large number of genera. Rarefaction OTU abundance prediction following clustering of sequences at 95% nucleotide identity provides a rule-of-thumb estimate of the number of genera in

a sample (Roesch et al., 2007). We acknowledge the problems in determining bacterial diversity using only a single marker gene, given the high degree of horizontal gene transfer that complicates the definition of a prokaryotic species. Nevertheless, such analyses are helpful in scaling the likely diversity of marine bacteria. Clustering at the 5 % level of similarity suggests that ~7000 genera were present at WCO ranging from a maximum of 1383 in March to a minimum of 617 in June. This high diversity contrasts to other communities, such as the gut flora, where a limited number of genera are found (Dethlefsen et al., 2008), but also demonstrates the complexity of soil-based ecosystems which demonstrate much greater diversity (Roesch et al., 2007). But high diversity is expected in marine communities (Sogin et al., 2006).

A large proportion of lineages are rare and divergent. It is to be expected that the number of unique sequences (and hence OTUs) observed in a sample will increase as a function of the total number of reads. In order to compare diversity across samples they were adjusted to a common total (11,327 reads, the number in the smallest data set of March 26th) by random re-sampling. Since samples were still in the steeply sloping region of the rarefaction curve (and therefore under-sampled), this resulted in the loss of 5280 (30%) OTUs and the retention of 12,393.

Surprisingly, following adjustment, 78% of the OTUs were still specific to one time-point (i.e. found in only one sample, but >=1 times) and 67% of the OTUs were singletons (occurred only once; Table 1), confirming that the large proportion of rare OTUs was not just an artefact of more deeply sequencing particular samples. Furthermore, the majority of these singletons diverged from other tags in the dataset

by >2 base pairs, which is outside the range of the average sequencing error after quality filtering of 0.25% (Huse et al., 2007). 55% of tags remained unclustered at a nucleotide identity of 90%, confirming the high diversity of the "rare biosphere" in the WCO.

Diversity varies through the year. Temporal changes in assemblages of known bacterial phyla were investigated by annotating the OTUs through the GAST pipeline (Sogin et al., 2006; Huse et al., 2007). 83.5% of sequences (11,356 OTUs, 113,474 sequences) were identified as bacterial, 0.04% as archaeal, 13.8% as eukaryotic organelles (from either chloroplasts or mitochondria), and 2.7% could not be identified. The archaeal sequences are likely the result of bacterial primer homology in these few archaeal groups, while the unknowns could represent unidentified lineages (although it is possible that these could also be non-16S rDNA). Only sequences identified as bacterial were used for subsequent analysis. A total of 35 phyla were identified. *Proteobacteria* accounted for 50.3% of the reads (Table S2), while the ten most abundant phyla contributed 99.4% of the read abundance. *Alphaproteobacteria* was by far the most abundant lineage with 31.7% of sequences. Importantly, SAR11 was largely responsible for the numerical dominance of the *Alphaproteobacteria*, contributing 12.5% of all sequences.

Richness was highest in February and March, with minor peaks in May, June-July and August-September **(**Table 2**)**. Dominance was high (0.940-0.982) throughout the year, with peaks in February-March and June-July (Fig. 1, Table 2), reflecting dominance by a relatively small number of taxa.

Alphaproteobacteria generally dominated throughout the year except on August 20th and Sept 29th when *Bacteroidetes* was of equal or greater abundance (Fig. 2). *Bacteroidetes* lineages were most abundant in August-December (Fig. 2). *Gammaproteobacteria* was the second most abundant taxon during in February-March and May-June, and was least abundant towards the end of the year. *Cyanobacteria*, the fourth most common phylum, was far less common with a peak abundance during May, immediately following the spring bloom, with a chlorophyll maximum in April (Table 3, Fig. 2). Interestingly, the peaks in bacterial cell densities observed in June and August were due to the presence of different lineages (Table 3; Fig 2; Table S2).

The relative diversity of each major lineage was compared (in the re-sampled data set), based on the number of unique OTUs annotated to each phylum (Fig. 3). Again *Alphaproteobacteria* was generally the most diverse group comprising 27% (3340 OTUs). *Bacteroidetes* and *Alphaproteobacteria* showed a near identical trend in diversity throughout the year and similar trends, to a greater or lesser extent, occurred for all annotated bacterial taxa (Fig. 3).

The seasonal pattern in community structure was investigated using non-parametric multivariate analysis, based on occurrences of the 11,356 unique bacterial OTUs in the re-sampled data, and showed evidence for a cyclical pattern (RELATE test, R=0.5, p<0.01). Multi-dimensional scaling (MDS) analysis (Fig. 4) showed potential for seasonal succession, with a relatively large shift in community structure between late March and late April, and development back towards winter conditions from late July. Thus observed changes in overall diversity are determined by underlying changes in

community structure, with different lineages increasing and declining as the assemblage cycles through time. Interestingly, the seasonal structure of the community appeared to be driven by the most abundant organisms; the 14 re-sampled bacterial OTUs that have >1000 sequences also show evidence of a cyclical pattern, albeit not as strong as when all bacterial OTUs are included (RELATE test, R=0.36, p<0.01). The co-performed T-RFLP analysis showed no evidence of cyclicity or seasonality. This suggests that the T-RFLP results do not reflect the most abundant taxa as might be expected, and actually misrepresent the changes in community structure. However, it is conceded that the resolving power of T-RFLP can be greatly affected by variation in the methodology.

Community structure correlates with changes in temperature and nutrients. Temporal changes in the total assemblage, as revealed by 16S-tag sequencing analysis, were significantly correlated with a combination of temperature, phosphate and silicate (Table 4). Temperature has been previously identified as an important influence on marine bacterial community structure (Fuhrman et al., 2008), and phosphate has long been considered to limit planktonic productivity (Atkins, 1923). However, silicate concentration has not previously been shown to have a significant correlation with overall bacterial diversity; however, it may correlate with other environmental factors that do directly influence bacteria.

Separate analyses were conducted on the 10 most abundant phyla. A combination of temperature and SRP provided the closest match with changes in community structure among *Alphaproteobacteria*, *Actinobacteria*, *Bacteroidetes* and *Gammaproteobacteria*, and all of these relationships were highly significant (Table

4). A combination of temperature and SRP and silicate provided the closest match to changes in community structure among *Cyanobacteria*, which was the fourth most abundant bacterial phylum (Table 4). Silicate has previously been linked to cyanobacterial abundance as a proxy for nutrient concentrations (Repka et al., 2004). Temperature, in combination with the abundance of nanoeukaryotes and coccolithophores, provided significant matches with changes in the community structure of *Betaproteobacteria* (Table 4), which may indicate potential trophic interaction. For other phyla (*Firmicutes*, *Deferribacteres*, *Planctomycetes* and *Verrucomicrobia*) the matches were not significant at p=0.05. Strikingly, while temperature is undoubtedly a highly important factor, no single key variable could be identified as solely responsible for the seasonal variation in diversity.

Conclusions

Despite the large number of 16S-tag sequences obtained in the current study, we still have a long way to go to characterize the 'rare biosphere' present in our oceans (Sogin et al., 2006; Huse et al., 2008; Huber et al., 2007; Kim et al., 2008). However, this investigation of temporal changes in diversity provides a succinct introduction to seasonal structure and the factors that control it. We have confirmed that there were high levels of seasonally-structured microbial diversity at WCO. There is also evidence for seasonal succession, primarily reflecting changes among dominant taxa; however, to fully explain the observed seasonal structure it is necessary to include all OTUs. Microbial assemblages at this coastal site were highly dominated by taxa from the SAR11 clade, further demonstrating the ubiquity and persistence of this important marine group. Total diversity was vastly under-sampled, with 1 in 25 reads representing a new genus; this reflects the need for continued investigation into

community diversity and structure using these high-throughput techniques. The community structure of individual phyla showed correlative relationships to combinations of environmental parameters, primarily driven by temperature and nutrient concentration, with inorganic phosphorus being highlighted as an important resource for the dominant microbial lineages.

This study has demonstrated an ideal for monitoring microbial diversity in marine ecosystems. Investigating inter-annual variability as part of a monitoring program is the next logical step, this will help to inform modeling of environmental impacts on microbial diversity and ultimately function. This is currently being performed retrospectively and is being planned for future years.

Experimental Procedures

Sampling, physico-chemical and biological variable analyses, nucleic acid extraction and sequencing. Seawater samples were collected on 12 occasions during 2007 from the L4 sampling site (50° 15.00' N, 4° 13.02') of the Western Channel Observatory (http://www.westernchannelobservatory.org.uk). The sampling dates were Feb 16th, Mar 7th, Mar 25th, Apr 23rd, May 8th, June 6th, June 25th, Jul 30th, Aug 20th, Sept 29th, Oct 25th, Dec 12th. Sixteen environmental variables were obtained from the WCO database, available on the website; these included surface temperature, salinity, water density, silicate concentration, nitrate + nitrite concentration, chlorophyll concentration, total organic nitrogen (TON) concentration, ammonium concentration, soluble reactive phosphorus (SRP) concentration, and total organic carbon (TOC) concentration. Other variables measured were flow cytometric determination of bacterial abundance, microscope counts of cryptophyte, picoeukaryote, nanoeukaryote, coccolithophore and small dinoflagellate abundance

(Table 3). All methods for determining these variables are available on the WCO website (http://www.westernchannelobservatory.org.uk/all_parameters.html).

Nucleic acid was extracted from 5 L seawater, collected from the surface and filtered immediately through a 0.22 µm Sterivex cartridge (Millipore), which was stored at -80 °C. DNA was isolated from each sample (Neufeld et al., 2007)) and then stored at -20 °C. Samples were sent to the ICoMM Initiative laboratory at Woods Hole in January 2008 for V6-region 16S rRNA amplification using the method of Huber *et al.* (2007) and 454-pyrosequencing using the GS-flx platform and LR70 kit.

Adjustment of the total number of sequence reads to smallest dataset size by resampling. Re-sampling of the 12 samples to identical sequencing depth was done by randomly selecting reads in fasta format using Daisy_chopper v1.0 (http://www.genomics.ceh.ac.uk/GeneSwytch/Tools.html). Comparison of the full and adjusted data sets confirmed the need to reduce the data to equal sequencing effort for correct biological interpretation. For example, following adjustment, the February 16th sampled changed from being estimated as the most diverse time-point to being less than that of March 7th (Fig. S2).

Estimates of OTU abundance and richness. All data generated from the sequencing project was handled through the Visualization and Analysis of Microbial Population Structure project (VAMPS) website (http://vamps.mbl.edu/index.php). The VAMPS workflow was used to create a profile of the unique sequences, their taxonomic assignment and their abundance in each sample. To create OTUs we aligned tags with MUSCLE, and created a pairwise distance matrix with quickdist (Sogin et al., 2006).

The output from quickdist was used as input to DOTUR (Distance Based OTU and Richness Determination; Schloss & Handelsman, 2005) to obtain Chao1 and ACE richness indices, operational taxonomic unit (OTU) abundance and rarefaction data. ACE and Chao indices are shown in Table S3 to allow comparison between this and other published studies (Table S1). Simpson's 1-□', an index of dominance, was calculated in Primer v6 (Clarke & Gorley, 2006).

Sequence annotation. OTUs in the normalized dataset were annotated with the GAST process (Sogin et al., 2006) and is freely available through the VAMPS website (http://vamps.mbl.edu/index.php). This project is also registered with the GOLD database (Gm00104).

Terminal Restriction Fragment Length Polymorphism (T-RFLP) analysis of bacterial 16S rRNA. A 50 µl PCR reaction containing: 50ng DNA, 0.5 µM of each primer, 2 U Taq polymerase (Sigma), 0.5 µl Taq buffer (Sigma), 0.5 mM dNTPs (Bioline), and 0.5 µl 10 mg/ml Bovine Serum Albumin (New England Biolabs). A ~520 bp product was amplified using primers 6FAM- 27F and 536R (Suzuki et al., 1998). PCR products were purified using QIAquick gel purification kit (QIAGEN) and analyzed using T-RFLP. Relative abundance was calculated for each by dividing individual peak height by total sample peak height value.

Statistical Analysis. Statistical analyses were performed using PRIMER v6 (Clarke & Gorley, 2006). BIO-ENV (Clarke & Ainsworth, 1993) was used to explore relationships between changes in community structure and measured environmental variables. In addition to physico-chemical measurements (Table 4), abundances of a range of planktonic eukaryotes (Table 3) were included. Briefly, the method searches

for subsets of environmental variables, which in combination maximize the rank correlation between resemblance matrices (Clarke et al., 2006) derived from the subsets of environmental variables and a fixed matrix derived from the OTUs. For the analyses, environmental variables were normalized (by subtracting the mean and dividing through by the standard deviation) to convert them to a common scale. Euclidean distances between samples were used to quantify resemblance. For the 16S OTUs abundances were log(N+1) transformed, to downweight the most abundant OTUs, and the Bray-Curtis similarity coefficient was used to quantify resemblances. A permutation test (Clarke et al., 2008) was used to assess whether the observed match could have arisen by chance.

Acknowledgements

This project was funded in part through the International Census of Marine Microbes 2007 funding initiative (http://vamps.mbl.edu/index.php) and by the UK Natural Environment Research Council through its Oceans 2025 programme. We thank all involved in the collection and processing of samples for the Western Channel Observatory (www.westernchannelobservatory.org.uk), especially Denise Cummings.

References

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev.* 59:143-69.

Atkins, W.R.G. (1923) The phosphate content of fresh and salt waters in its relationship to the growth of the algal plankton. *J. mar. biol. Ass. U.K.* 13: 119-150.

Clarke, K.R., and Ainsworth, M. (1993) A method of linking multivariate community structure to environmental variables. *Mar Ecol Prog Ser* 92: 205-219.

Clarke, K.R., and Gorley, R.N. (2006) PRIMER v6: User Manual/Tutorial. PRIMER-E, Plymouth.

Clarke, K.R., Somerfield, P.J., Chapman, M.G., and Needham, H.R. (2006) Dispersion-based weighting of species counts in assemblage analysis. *Mar Ecol Prog Ser* 320: 11-27.

Clarke, K.R., Somerfield, P.J., and Gorley, R.N. (2008) Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage. *J Exp Mar Biol Ecol* 366: 56-69.

Dethlefsen, L., Huse, S., Sogin, M.L., and Relman, D.A. (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol*. 6:e280.

DeLong, E.F. (1992) Archaea in coastal marine environments. *Proc Natl Acad Sci USA* 89: 685–5689.

Fuhrman, J.A., Steele, J.A., Hewson, I., Schwalbach, M.S., Brown, M.V. *et al*. (2008) A latitudinal diversity gradient in planktonic marine bacteria**.** *Proc Natl Acad Sci USA* 105: 7774-8. 19.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and Field, K.G. (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345: 60–63.

Hewson, I. and Fuhrman, J.A. (2004) Richness and diversity of bacterioplankton species along an estuarine gradient in Moreton Bay, Australia. *Appl Environ Microbiol*. 70: 3425-33.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2002) Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subsea floor habitat. *Appl Environ Microbiol*. 68: 1585-94.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R. et al. (2007) Microbial population structures in the deep marine biosphere. *Science.* 5: 97-100.

Huse, S.M., Dethlefsen, L., Huber, J.A., Welch, D.M., Relman, D.A. *et al*. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet*. 4: e1000255.

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

Kim, B.S., Kim, B.K., Lee, J.H., Kim, M., Lim, Y.W. *et al*. (2008) Rapid phylogenetic dissection of prokaryotic community structure in tidal flat using pyrosequencing. *J Microbiol.* 46: 357-63.

Neufeld, J.D., Schäfer, H., Cox, M.J., Boden, R., McDonald, I.R. *et al*. (2007) Stable isotope probing implicates Methylophaga spp and novel Gammaproteobacteria in marine methanol and methylamine metabolism. *ISME J* 1: 480–491.

Pace, N. (1997) A molecular view of microbial diversity and the biosphere. *Science* 276: 734–740.

Rappé, M.S. and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu Rev Microbiol*. 57: 369-94.

Repka, S., Koivula, M., Harjunpä, V., Rouhiainen, L., and Sivonen, K. (2004) Effects of phosphate and light on growth of and bioactive peptide production by the Cyanobacterium anabaena strain 90 and its anabaenopeptilide mutant. *Appl Environ Microbiol*. 70:4551-60.

Roesch, L.F., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K. *et al*. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity**.** *ISME J*. 1: 283-90.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S. et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5: 398-431.

Schloss, P.D. and Handelsman, J. (2005). Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Appl Environ Microbiol* 71: 1501-1506.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M. et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA* 103: 12115-20.

Southward, A.J., Langmead, O., Hardman-Mountford, N.J., Aiken, J., Boalch, G.T. *et al*. (2005) Long-term oceanographic and ecological research in the Western English Channel. *Adv Mar Biol.* 47: 1-105.

Suzuki, M., Rappé, 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. *Appl Environ Microbiol* 64: 4522-4529.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* 95:6578-83.

Figure Legends

Figure 1. Dominance (Simpsons 1-□') calculated using re-sampled data and bacterial-only OTUs.

Figure 2. Relative abundance of the ten most frequently annotated bacterial taxa for each sample during 2007. Abundances are based on those sequences which could be annotated using the GAST pipeline against the Ribosomal Database Project 16S rRNA dataset.

Figure 3 – The diversity of the ten most diverse annotated bacterial taxa for each sample during 2007. Diversity is based on the number of unique operational taxonomic units found in each phylum at each time point.

Figure 4. Nonmetric Multi Dimensional Scaling ordination of a Bray-Curtis resemblance matrix among samples calculated from log(n+1) transformed abundances of 11,356 bacterial OTUs. 2D stress = 0.11.

Table 1 – The abundance of rare sequences, defined as OTUs with only one representative sequence or OTUs only found at one time-point for 100% identical (unique) OTUs, clustered at 97% (3 %), 94% (6%) and 90% (10 %) identity for both the original data and the re-sampled to smallest dataset data. Percentage of total OTU abundance is given in parentheses. Singletons refer to reads found only once in the entire experiment.

		OTU number	**Singletons**	**Date specific (≥ 1) OTU**
Original data	Unique	17673	11806 (67%)	13714 (78%)
	3% dissimilar	9229	4227 (46%)	5456 (59%)
	6% dissimilar	6288	2612 (42%)	3481 (55%)
	10% dissimilar	4375	1742 (40%)	2370 (54%)
Re-sampled data	Unique	12393	8244 (67%)	9727 (78%)
	3% dissimilar	7179	3374 (47%)	4443 (62%)
	6% dissimilar	5072	2137 (42%)	2926 (58%)
	10% dissimilar	3544	1395 (39%)	1944 (55%)

Table 2 – The number of sequences, unique OTU abundance and Simpson's 1-λ' for all, and bacterial-only, re-sampled OTUs.

Sample	Re-sampled by smallest sample size (11327)					
	Only bacterial Sequences			Primary dataset – bacteria, archaea, eukaryote organelles and unknowns		
	OTUs	Sequences	Dominance	OTUs	Sequences	Dominance
Feb 16th	1937	9867	0.978	2169	11327	0.983
Mar 7th	2075	9245	0.982	2343	11327	0.985
Mar 26th	1887	9017	0.979	2098	11327	0.983
Apr 23rd	1177	8870	0.975	1318	11327	0.977
May 8th	1239	8897	0.959	1401	11327	0.965
Jun 4th	1138	9179	0.957	1284	11327	0.970
Jun 25th	1427	8689	0.977	1579	11327	0.982
Jul 30th	1337	8844	0.976	1510	11327	0.979
Aug 20th	1112	10204	0.959	1301	11327	0.966
Sep 29th	1337	10393	0.956	1494	11327	0.963
Oct 25th	1446	9814	0.940	1628	11327	0.954
Dec 12th	1118	10455	0.960	1282	11327	0.966

Table 3 – Values environmental variables for each sampling occasion.

	Feb 16th	Mar 7th	Mar 26th	Apr 23rd	May 8th	June 4th	June 25th	July 30th	Aug 20th	Sept 29th	Oct. 25th	Dec. 12th
Temperature (ºC)	10.1	10.3	10.3	11.8	12.4	13.9	15.2	15.5	15.5	15.9	15.4	13.9
Salinity (PSU)	35.1	35.1	34.9	35.2	35.1	35.2	35.0	35.1	34.9	35.2	35.1	35.3
$NO_2 + NO_3$ (µmol L^{-1})	5.3	7.4	6.0	1.5	0.1	0.6	0.1	0.1	2.0	1.6	3.5	8.0
Ammonium (µmol L^{-1})	0.1	0.4	0.2	0.0	0.2	0.3	0.1	0.8	0.3	0.1	0.2	0.3
SRP (µmol L^{-1})	0.5	0.5	0.4	0.1	0.1	0.0	0.0	0.0	0.1	0.2	0.3	0.4
Silicate (µmol L^{-1})	4.1	5.5	4.2	1.7	1.3	1.0	0.3	0.9	2.9	2.3	2.7	4.6
TON (µmol L^{-1})	2.4	3.4	3.1	3.5	3.6	4.8	5.2	4.4	2.3	1.9	1.7	1.4
TOC (µmol L^{-1})	23.6	29.1	25.4	23.2	27.5	47.3	47.6	39.2	18.9	17.2	15.8	12.4
Bacteria (10^6 mL^{-1})	0.60	0.37	0.64	0.70	1.81	2.19	1.27	0.73	1.51	1.16	0.42	0.62
Chlorophyll (µg L^{-1})	0.44	1	0.64	2.53	1.24	1.39	1.58	2.22	1.68	1.03	0.48	0.46

Table 4 Subsets of environmental variables (X if $P<0.05$, n.s if $P>0.05$) from BIO-ENV which best explain variation in community structure derived from all OTUs and the 10 most abundant phyla. Blanks indicate no correlation. SRP – soluble reactive phosphorus. TOC – total organic carbon. *Firmicutes*, *Deferribacteres*, *Planctomycetes* and *Verrucomicrobia* all showed no-significance in their correlations and so were not included.

	Rho	*p*	Variables					
			Temperature (ºC)	Silicate (μmol L^{-1})	SRP (μmol L^{-1})	Density (kg m^{-2})	TOC (μmol L^{-1})	Nanoeukaryotes (cells mL^{-1})
All OTUs	0.73	0.002	X	X	X			
Gammaproteobacteria	0.75	0.002	X		X			
Alphaproteobacteria	0.73	0.001	X		X			
Bacteroidetes	0.72	0.003	X		X			
Actinobacteria	0.63	0.014	X		X			
Cyanobacteria	0.57	0.016	X	X	X	X	X	X
Betaproteobacteria	0.52	0.040	X					

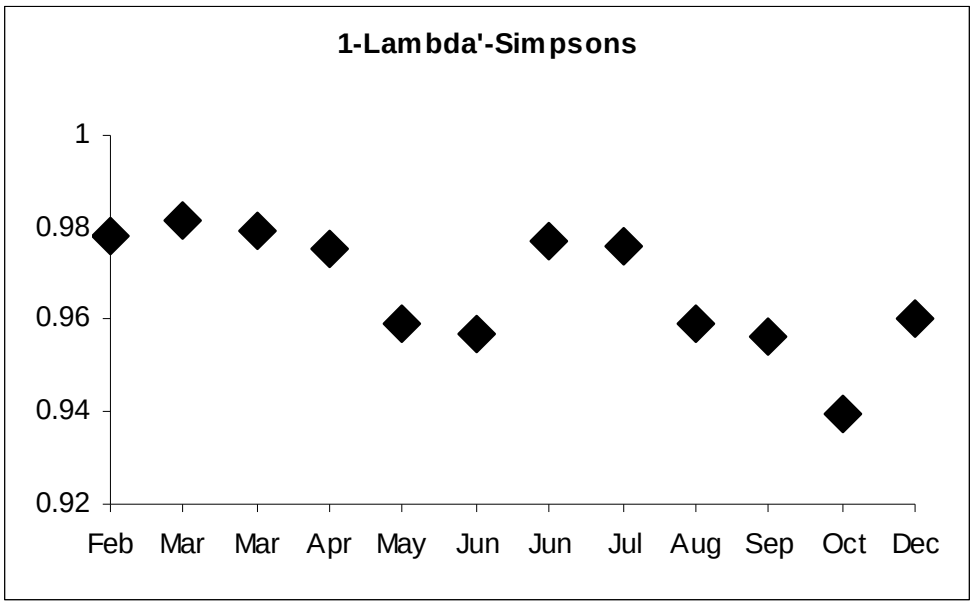

Figure 1. Dominance (Simpsons 1-□') calculated using re-sampled data and bacterial-only OTUs.

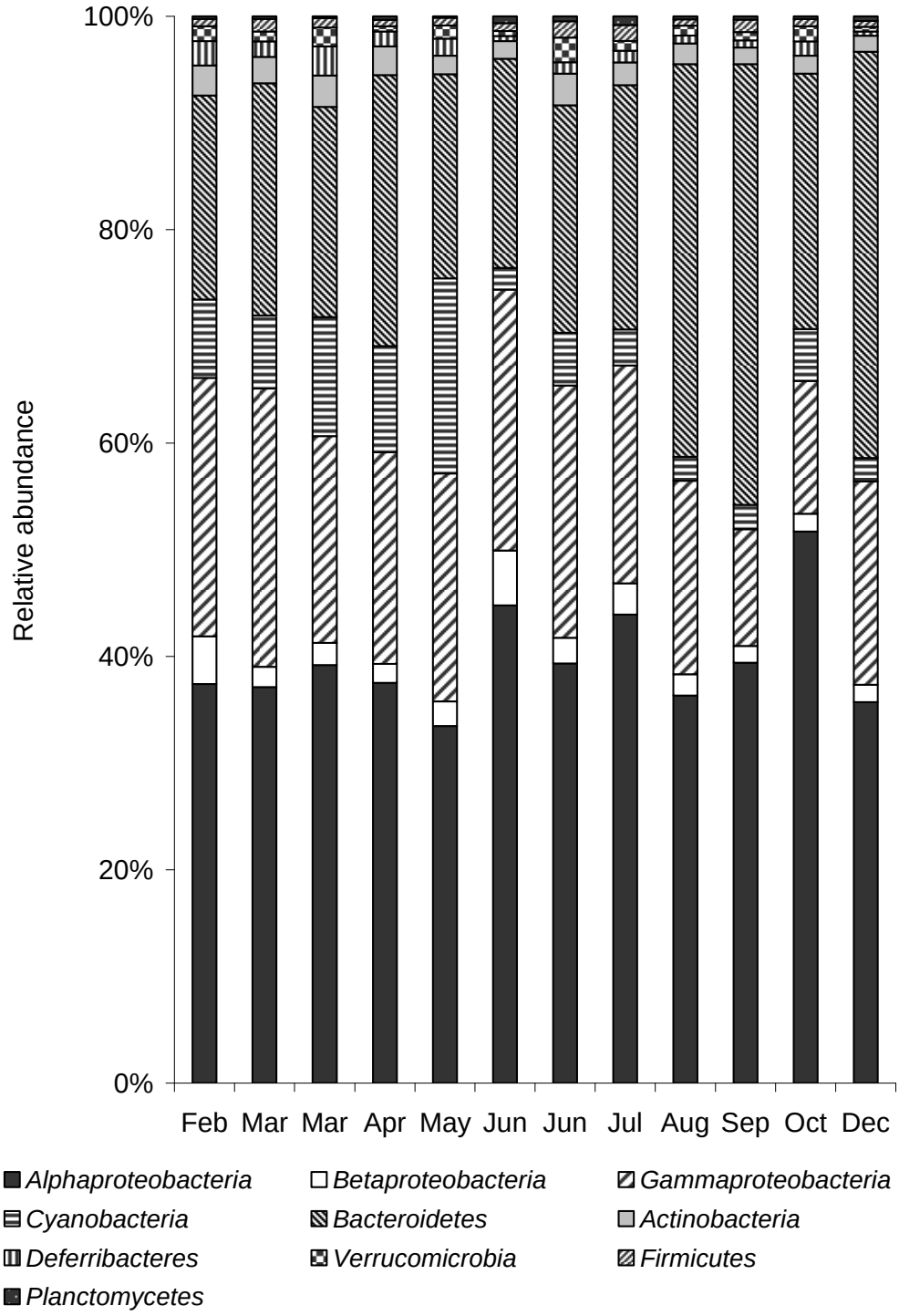

Figure 2. Relative abundance of the ten most frequently annotated bacterial taxa for each sample during 2007. Abundances are based on those sequences which could be annotated using the GAST pipeline against the Ribosomal Database Project 16S rRNA dataset.

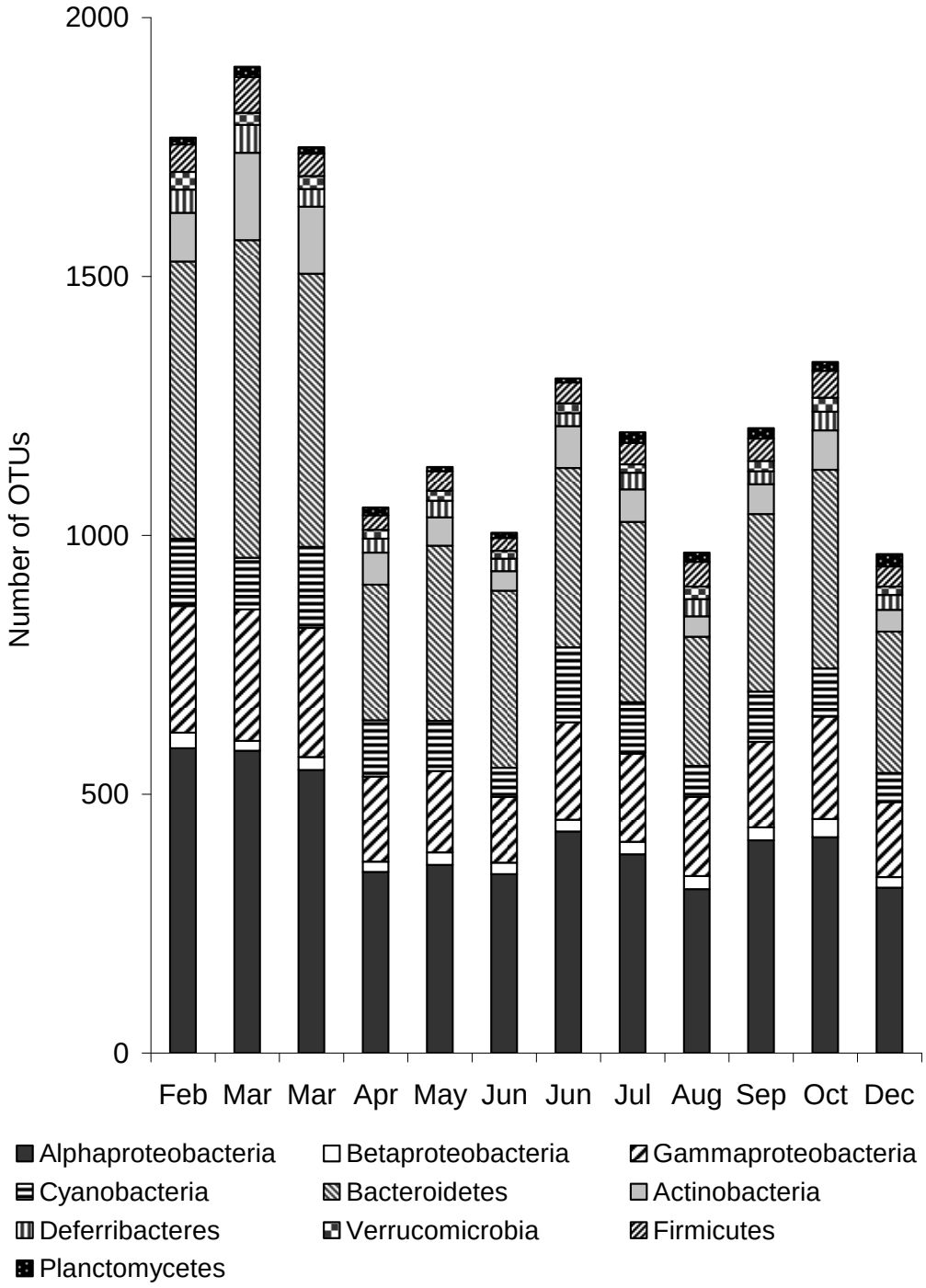

Figure 3 – The diversity of the ten most diverse annotated bacterial taxa for each sample during 2007. Diversity is based on the number of unique operational taxonomic units found in each phylum at each time point.

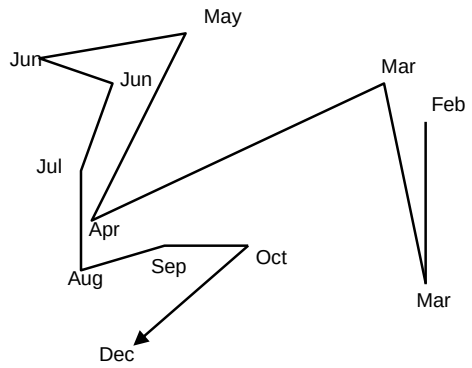

Figure 4. Nonmetric Multi Dimensional Scaling ordination of a Bray-Curtis resemblance matrix among samples calculated from log(n+1) transformed abundances of 11,356 bacterial OTUs. 2D stress = 0.11.