Bias in culture-independent assessments of microbial biodiversity in the global ocean

Ben Temperton[1], Anna Oliver[2], Dawn Field[1], Bela Tiwari[3], Martin Mühling[1], Ian Joint[1] and Jack A. Gilbert[1]

[1] Plymouth Marine Laboratory, Prospect Place, Plymouth, UK

[2] NERC Centre for Ecology and Hydrology, CEH Oxford, Oxford, United Kingdom

[3]NEBC Centre for Ecology and Hydrology, CEH Oxford, Oxford, United Kingdom

On the basis of 16S rRNA gene sequencing, the SAR11 clade of marine bacteria has almost universal distribution, being detected as abundant sequences in all marine provinces. Yet SAR11 sequences are rarely detected in fosmid libraries, suggesting that the widespread abundance may be an artefact of PCR cloning and that SAR 11 has a relatively low abundance. Here the relative abundance of SAR11 is explored in both a fosmid library and a metagenomic sequence data set from the same biological community taken from fjord surface water from Bergen, Norway. Pyrosequenced data and 16S clone data confirmed an 11-15% relative abundance of SAR11 within the community. In contrast not a single SAR11 fosmid was identified in a pooled shotgun sequenced data set of 100 fosmid clones. This under-representation was evidenced by comparative abundances of SAR11 sequences assessed by taxonomic annotation; functional metabolic profiling and fragment recruitment. Analysis revealed a similar under-representation of low-GC *Flavobacteriaceae*. We speculate that the fosmid bias may be due to DNA fragmentation during preparation due to the low GC content of SAR11 sequences and other underrepresented taxa. This study suggests that while fosmid libraries

can be extremely useful, caution must be used when directly inferring community composition from metagenomic fosmid libraries.

The majority of microbes are uncultivated but culture-independent methods have provided clear insights into in natural assemblages. The use of fosmids to propagate large (~40 Kb) genomic inserts with a high degree of fidelity[1]has enabled contextual analysis of genes and their genomic neighbourhood; this has led to identification of particular metabolic pathways, such RNA helicase in *Archaea*[2], and the discovery of bacteriorhodopsin in marine bacteria[3]. Fosmid libraries have also given insight into the genomic variation within and between mixed marine microbial assemblages in relation to water depth[4,5] and symbiosis in marine invertebrates[6-8]. However, several studies have indicated that fosmid clone DNA libraries are not fully consistent with other assessments natural microbial communities, with poor representation of key members such as SAR11[4,5,9] and high dominance of *Roseobacter spp.*[10,11]. In contrast to an estimated 25% abundance of SAR11 in marine microbial communities from plasmid libraries[12], Gilbert and colleagues[9] isolated only a single clone containing a 16S rRNA gene with homology to the SAR11 clade from a marine surface water fosmid library of 10,000 clones. Interestingly, this clone consisted largely of the 48-Kb hypervariable region neighbouring the 16S rRNA gene, with a richer GC content than the GC-poor SAR11 core genome. *Roseobacter spp.* and other high-GC taxa seemed to suffer no such underrepresentation.

Here we use GC analysis and taxonomic profiling to compare the representation of key taxa with differing average GC content within a fosmid library of 100 clones to 326,310 metagenomic sequences from 454 pyrosequencing of the same water sample. Sequences were compared to fully sequenced genomes of nineteen key taxa using the BLASTN algorithm, with hits with an *E*-value < $1x10^{-5}$ considered significant. As each fosmid clone contained a fragment from a single bacterial cell, the fosmid library contained a maximum of 100 different taxa, whereas the pyrosequenced sample had no such limit. To account for this, 100 sequences were randomly selected from the pyrosequenced data and compared against the key taxa. This was performed 1000 times and an average abundance for each taxa was measured. To avoid bias of repetition of possible GC-rich or GC-poor sequences within the fosmid data, all fosmid analyses were performed on assembled contiguous DNA fragments (contigs). 10.7% of sequences in pyrosequenced data were significantly similar to "*Candidatus* Pelagibacter ubique" HTCC1062, compared to no hits in the fosmid data. All 1000 replicates of pyrosequenced data contained at least two hits to "*Cand.* P. ubique", thus confirming the ubiquity of this SAR11 strain and, as in other studies, that it is underrepresentation in fosmid libraries. Similarly, sequences similar to *Flavobacteriaceae sp.* were also absent from the fosmid data but comprised ~1.5% of pyrosequenced data. Conversely, only 1.0% of sequences in pyrosequenced data were significantly similar to *Roseobacter denitrificans* OCh114, compared to 48.6% of sequences in fosmid data.

Relative abundances of each of the key taxa were plotted with their average GC content (Figure 1). The average GC content for each dataset was also calculated alongside sequences derived from plasmid libraries from the Global Ocean Sampling Expedition (GOS) from the Bay of Fundy (45°6'42" N, 64°58'48" W), and Brown Bank, Gulf of Maine (42°51'10" N, 66°13'2" W)[13], sites from similar habitats to the Bergen sampling site in the current study. As a comparison, the average GC content for fosmid libraries from the oligotrophic Pacific Ocean station ALOHA (22°45' N, 158°W), Hawaii at depths from 10m-4000m[5] were also calculated as to date, the Bergen dataset is the only large randomly sequenced fosmid library for coastal waters. Average GC content for pyrosequenced data (37.8%) was similar to that of GOS Bay of Fundy (35.4%) and Brown Bank (37.4%) but significantly different to average GC content of both our fosmid library (51.6%) and HOT station ALOHA fosmid libraries (48.4-

54.4%), which had high average GC values despite variable community composition with depth (Figure 2).

To confirm the abundance of key taxa within the samples, fragment recruitment plots were constructed against fully sequenced genomes of “*Cand.* P. ubique” HTCC1062, *Roseobacter denitrificans* OCh114, *Synechococcus sp.* and *Shewanella sp.* available from the National Centre for Biotechnology Information (NCBI). A recently sequenced open-ocean strain of SAR11, “*Cand.* P. ubique” HTCC7211 was also included. Plots were constructed using BLASTN parameters of “-F F -r 5 –q 4 –e 1e-4” to detect distant similarities as low as 65% identity. At 65% identity, ~15% of pyrosequenced sequences recruited to “*Cand.* P. ubique” HTCC1062, compared to <3% of sequences from the fosmid library, confirming the dominance of this species in coastal waters and its poor representation in fosmid datasets. The high degree of genetic conservation in HTCC1062, brought about through genomic streamlining[12], can be seen as a band of high recruitment at >90% identity across the whole genome, interspersed with gaps of little or no recruitment which correspond to regions of hypervariability found in previous studies[9,13,14] (Figure 3).

Similar recruitment (12.6% vs. 2.8%) was observed against “*Cand. P*. ubique” HTCC7211 but at lower percentage identity than against HTCC1062 indicating that these two coastal and open-ocean SAR11 strains are similar but also have niche specific genetic differences. Interestingly, HTCC7211 fragment recruitment plots revealed several putative regions of hypervariability within the genome (~4 kb at 487k; ~16.5 kb at 740k; ~29.5 kb at 801.5k; ~15.6k at 969.4k~10 kb at 1073.4k; ~14 kb at 1337k) suggesting that such regions may be common within the SAR11 clade. ~21.5% of fosmid sequences recruited to *R. denitrificans* OCh114, compared to 4.3% of hits for pyrosequenced data, but at lower percent identity than SAR11, suggesting a lower genetic conservation within this taxon. *Synechococcus sp.* recruited fewer sequences than expected in the pyrosequenced data (0.8%) in light of its global ubiquitous distribution[15] and its recruitment in the fosmid data (3.6%), most likely due to the low coverage of the pyrosequenced metagenome, estimated at ~0.0002% using a calculated effective genome size[16] of 2.33Mb. Even at such low recruitment, recruitment plots yield the locations of highly conserved regions such as rRNA and housekeeping genes as ‘peaks’ of syntenic fragments.

This study has confirmed unequivocally that fosmid libraries do not accurately represent bacterial taxonomic diversity within a given community in comparison to other sequence-based datasets. Fosmid libraries both from this study and from a depth profile study[5] showed significant over- or under-representation of important microbial taxa when compared to pyrosequenced data. Specifically, the low-GC SAR11 clade and Flavobacteria are significantly under-represented, resulting in an elevated GC content of fosmid data in comparison to direct pyrosequencing. One possible explanation is that fragmentation of DNA occurs more readily in DNA with fewer G-C linkages due to a decreased number of hydrogen bonds, weakening the strand against non-perpendicular shear forces and reducing the number of 40 Kb fragments required for fosmid vector insertion, thus lowering representation in a fosmid library. Other factors affecting strand stability such as nearest-neighbour effects[17] may also be important.

Recent work by Pham and colleagues[18] suggested that low numbers of SAR11 sequences detected in surface water fosmid libraries was an accurate representation of the bacterial community, and that over-representation of SAR 11 in earlier studies was due to PCR cloning bias. However, no PCR cloning was used in this study and the percentage hits to SAR11 from the sequencing of random fragments confirms the abundance of SAR11 in surface water. It is clear that large fragment-length inserts of fosmid libraries confers a huge advantage when investigating genetic neighbourhoods. However, exclusion of sequences, including those from the most ubiquitous bacterial clade, suggests that the use of fosmid libraries for metagenomic diversity studies must be performed with caution to avoid the possibility of misrepresenting key components of microbial diversity.

1. Kim, U. J., Shizuya, H., de Jong, P. J., Birren, B., Simon, M.I. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Research* **20**, 1083-1085 (1992).

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

3. Beja, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P. et al. Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science* **(298)** 1902-1906 (2000).

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

5. DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N. U *et al*. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science* **(311)** 496-503 (2006).

6. Schleper, C., DeLong, E. F., Preston, C. M., Feldman, R. A., Wu, K. Y., Swanson, R. V. Genomic Analysis Reveals Chromosomal Variation in Natural Populations of the Uncultured Psychrophilic Archaeon Cenarchaeum symbiosum. *Journal of Bacteriology* **(180)** 5003-5009 (1998).

7. Hughes, D. S., Felbeck, H., Stein, J. L. A histidine protein kinase homolog from the endosymbiont of the hydrothermal vent tubeworm *Riftia pachyptila*. *Applied and Environmental Microbiology* **(63)** 3494-3498 (1997).

8. Campbell, B. J., Stein, J. L., Cary, S. C. Evidence of Chemolithoautotrophy in the Bacterial Community Associated with *Alvinella pompejana*, a Hydrothermal Vent Polychaete. *Applied and Environmental Microbiology* **(69)** 5070-5078 (2003).

9. Gilbert, J. A., Mühling, M., Joint, I. A rare SAR11 fosmid clone confirming genetic variability in the '*Candidatus Pelagibacter ubique*' genome. ISME Journal: 1-4. (2008).

10. Suzuki, M. T., Rappe, M. S., Haimberger Z. W., Winfield, H., Adair, N., Strobel, J. *et al*. Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample. *Applied and Environmental Microbiology* **(63)** 983-989 (1997).

11. Buchan, A., Gonzalez, J. M., Moran, M. A. Overview of the Marine *Roseobacter* Lineage. *Applied and Environmental Microbiology* **(71)** 5665-5677 (2005).

12. Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D. *et al*. Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science* **(309)** 1242-1245. (2005)

13. Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S. *et al*. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **(5)** e77 doi:10.1371/journal.pbio.0050077 (2007)

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

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

16. Raes, J., Korbel, J. O., Lercher, M. J, von Mering, C., Bork, P. Prediction of effective genome size in metagenomic samples. *Genome Biology* **(8)** doi:10.1186/gb-2007-8-1-r10 (2007).

17. Allawi, H. T., SantaLucia Jr, J. Nearest-neighbor thermodynamic parameters for Internal GA Mismatches in DNA. *Biochemistry* **(37)** 2170-2179 (1998).

18. Pham, V. D., Konstantinidis, K. T., Palden, T., DeLong, E. F. Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000m vertical profile in the North Pacific Subtropica Gyre. *Environmental Microbiology* **(10)** 2313-2330 (2008)

Supplementary Information accompanies the paper on **www.nature.com/nature**.

This work was supported by a grant from the Natural Environmental Research Council (NERC) as part of the 2007 Molecular Genetics Facility funding initiative. Additional support came from a NERC grant (NE/C507902/1).

Correspondence and requests for materials should be addressed to Jack Gilbert. (jagi@pml.ac.uk).

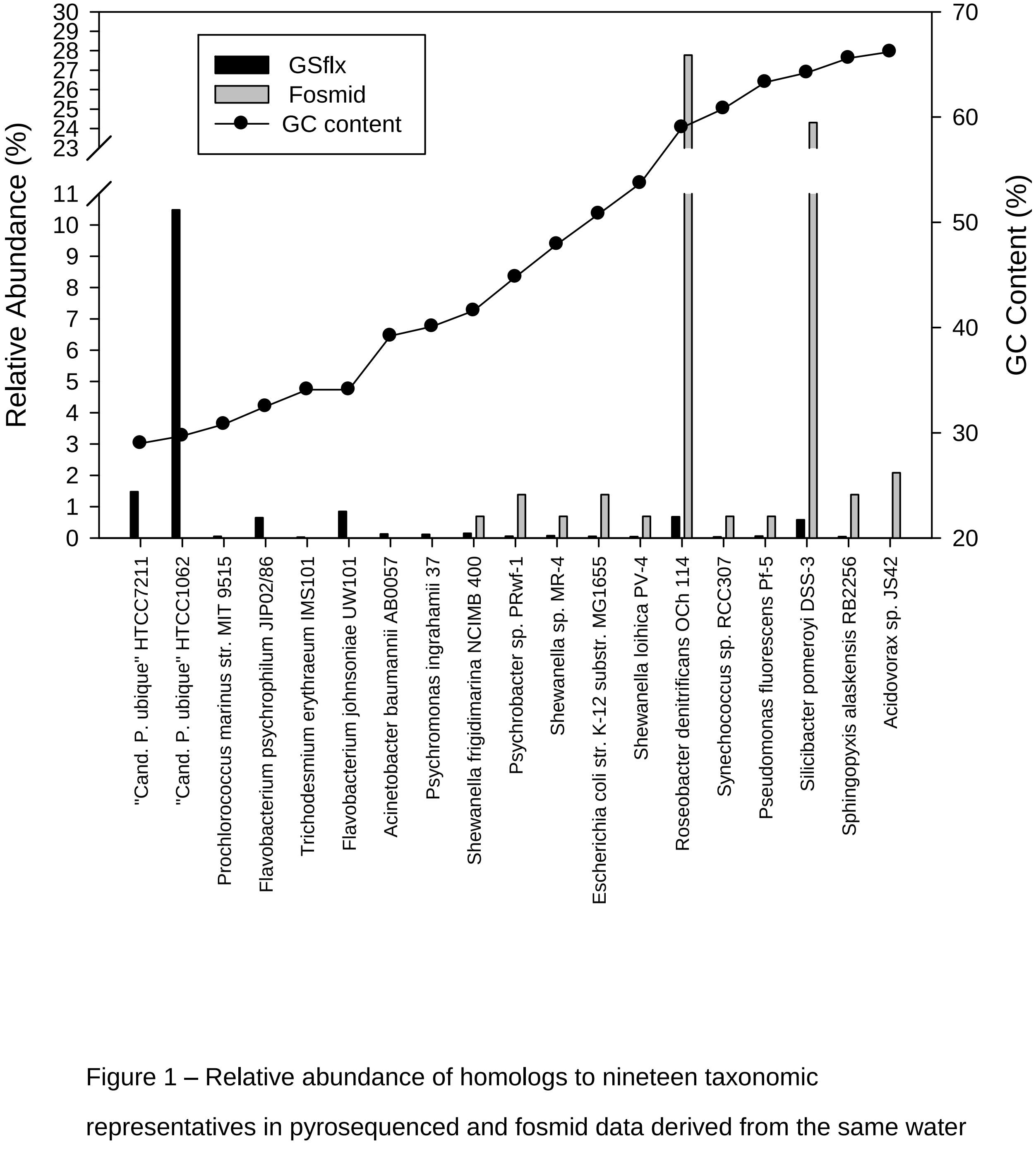

Figure 1 – Relative abundance of homologs to nineteen taxonomic representatives in pyrosequenced and fosmid data derived from the same water sample. Average GC content of each species is also plotted.

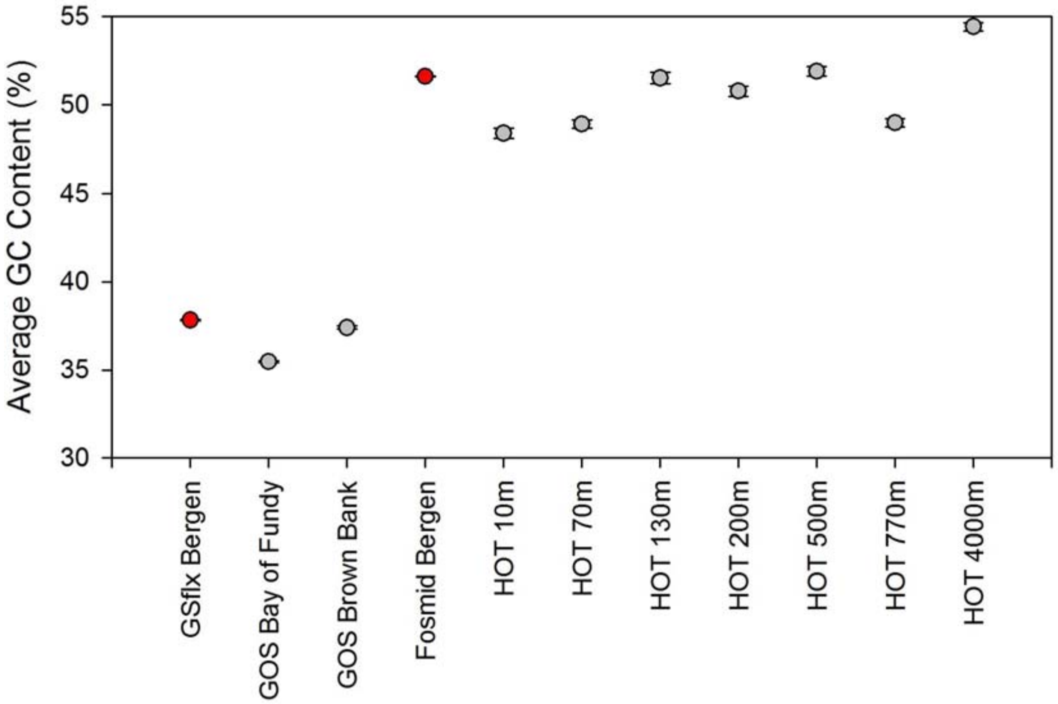

Figure 2 – Mean GC content for the following samples: pyrosequenced GSflx data from Bergen; surface water from Bay of Fundy prepared using a plasmid library[13]; surface water from Brown Bank, Gulf of Maine prepared using a plasmid library[13]; fosmid library from Bergen; samples from HOT station ALOHA prepared using a fosmid library taken at different depths[5]; Bars represent 99% confidence intervals. Data prepared for this paper are highlighted in red.

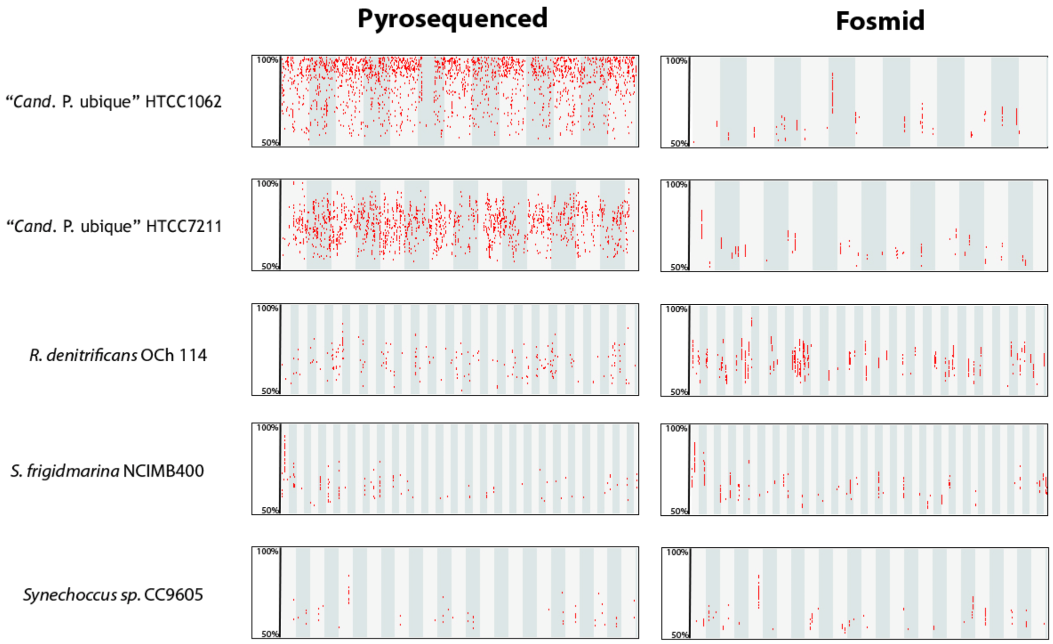

Figure 3 -Fragment recruitment plots for key bacterioplankton species. Sequences from pyrosequenced and fosmid data were aligned against the reference genomes using BLASTN with parameters designed to detect distant homologs as low as 65% identity. "*Cand.* P. ubique" HTCC7211 is included as an open-ocean strain of the SAR11 clade.