The ISME Journal (2009) 3, 792–796

www.nature.com/ismej

ORIGINAL ARTICLE

Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing

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

[1]*Plymouth Marine Laboratory, Plymouth, UK and* [2]*NERC Centre for Ecology and Hydrology, CEH Oxford, Oxford, UK*

On the basis of 16S rRNA gene sequencing, the SAR11 clade of marine bacteria has an 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 SAR11 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 sequence data set of 100 fosmid clones. This underrepresentation was evidenced by comparative abundances of SAR11 sequences assessed by taxonomic annotation and fragment recruitment. Analysis revealed a similar underrepresentation of low-GC *Flavobacteriaceae*. We speculate that a contributing factor towards the fosmid bias may be DNA fragmentation during preparation because of the low GC content of SAR11 sequences and other underrepresented taxa. This study suggests that, although fosmid libraries can be extremely useful, caution must be taken when directly inferring community composition from metagenomic fosmid libraries.

The ISME Journal (2009) **3,** 792–796; doi:10.1038/ismej.2009.32; published online 2 April 2009
Subject Category: microbial population and community ecology
Keywords: fosmid; cloning bias; marine; SAR11

Introduction

The majority of microbes are uncultivated, but culture-independent methods have provided clear insights into natural assemblages. The use of fosmids to propagate large (~40 kb) genomic inserts with a high degree of fidelity (Kim *et al.*, 1992) has enabled contextual analysis of genes and their genomic neighbourhood; this has led to the identification of particular metabolic pathways, such as RNA helicase in *Archaea* (Stein *et al.*, 1996), and the discovery of bacteriorhodopsin in marine bacteria (Béjà *et al.*, 2000). Fosmid libraries have also given an insight into the genomic variation within and between mixed marine microbial assemblages in relation to water depth (Suzuki *et al.*, 2004; DeLong *et al.*, 2006) and symbiosis in marine invertebrates (Hughes *et al.*, 1997; Schleper *et al.*, 1998; Campbell *et al.*, 2003). However, several studies have indicated that fosmid clone DNA libraries are not fully consistent with other assessments of natural microbial communities, with a poor representation of key members such as SAR11 (Suzuki *et al.*, 2004; DeLong *et al.*, 2006; Gilbert *et al.*, 2008) and a high dominance of *Roseobacter* spp. (Suzuki *et al.*, 1997; Buchan *et al.*, 2005). In contrast to an estimated 25% abundance of SAR11 in marine microbial communities from plasmid libraries (Giovannoni *et al.*, 2005) and 12–37% abundance from direct counting of SAR11 using fluorescent *in situ* hybridization (Mary *et al.*, 2006), Gilbert *et al.* (2008) 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. It is interesting that 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.

Correspondence: J Gilbert, Plymouth Marine Laboratory, Prospect Place, Plymouth, Devon PL1 3DH, UK.
E-mail: jagi@pml.ac.uk
Received 2 February 2009; revised 3 March 2009; accepted 3 March 2009; published online 2 April 2009

Here we present the direct comparison of a pyrosequenced metagenomic data set (Gilbert *et al.*, 2008) and pyrosequencing of 100 metagenomic fosmid clones from the same sample. Through this comparison, we describe the differential representation of microbial diversity within each data set.

Materials and methods

Sample site and isolation
Samples were collected from a mesocosm study in coastal waters close to Bergen, Norway (60.27°N: 5.22°E). Water samples were collected as described by Gilbert *et al.* (2008). Briefly, CO_2 was bubbled through mesocosms containing 11 000 l of coastal water to simulate conditions of ocean acidification. Control mesocosms were kept at present-day CO_2 levels. Plankton blooms were induced by the addition of nitrate and phosphate. Immediately after the collapse of the phytoplankton bloom, 15-l water samples were pre-filtered using a GF/A (1.6 μm) filter (Whatman, Sanford, ME, USA), and then microbial cells were collected on a 0.22-μm Sterivex filter (Millipore, Billerica, MA, USA). Sterivex filters were stored at −80 °C until nucleic acid extraction.

DNA/RNA purification and sequencing
DNA and RNA were isolated from each sample and prepared for metagenomic and meta-transcriptomic pyrosequencing as described by Gilbert *et al.* (2008). All data for this study came from the high CO_2 post-bloom mesocosm. All pyrosequencing data were generated using the GS FLX platform at the NERC-funded Advanced Genomics Facility at the University of Liverpool (http://www.liv.ac.uk/agf/).

Fosmid library construction and sequencing
A library of 10 000 fosmid clones was created from the same DNA as that for pyrosequencing using the pCC2FOS CopyControl fosmid library production kit (Epicentre, Madison, WI, USA), following the manufacturer's guidelines. A total of 100 clones were randomly selected from this library and cultures were induced to high copy number using an autoinduction solution (Epicentre). Each fosmid clone was cultured and DNA was extracted as detailed in the manufacturer's guidelines for the FosmidMAX 96-well DNA extraction kit (Epicentre). DNA was treated with Plasmid-Safe DNase (Epicentre). DNA from 24 randomly selected clones was pooled and treated as one sample. This was carried out for all 96 clones. Sequencing of these four pools of fosmids produced 273 065 sequences containing a total of 68 061 629 bp. To remove sequences from the host *E. coli str. K12 substr. MG1655* before analysis, sequences were aligned against the host strain genome from the National Centre for Biotechnology Information (NCBI) using BLASTN. Removal of sequences with 98% identity over 100% length to the host strain excluded 642 sequences, leaving 272 423 sequences containing a total of 67 879 375 bp. Subsequently, the fosmid sequence fragments were assembled using the Newbler assembly program (Roche, 454) to provide access to assembled contigs of each fosmid.

Results and discussion

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 with fully sequenced genomes of 19 key taxa using the BLASTN algorithm, with hits with an *E*-value $<1\times10^{-5}$ 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 carried out a 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 carried out on assembled contiguous DNA fragments (contigs). Sequences (10.7%) in pyrosequenced data were significantly similar to '*Candidatus* Pelagibacter ubique' HTCC1062, compared with 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, the fact that it is underrepresented 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 with 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 data set 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 from the Brown Bank, Gulf of Maine (42°51′10′ N, 66°13′2′ W) (Rusch *et al.*, 2007), sites with habitats similar to the Bergen sampling site in this 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 10 to 4000 m (DeLong *et al.*, 2006) was also calculated, as to date, the Bergen data set is the only large randomly sequenced fosmid library for coastal waters. Average GC content for

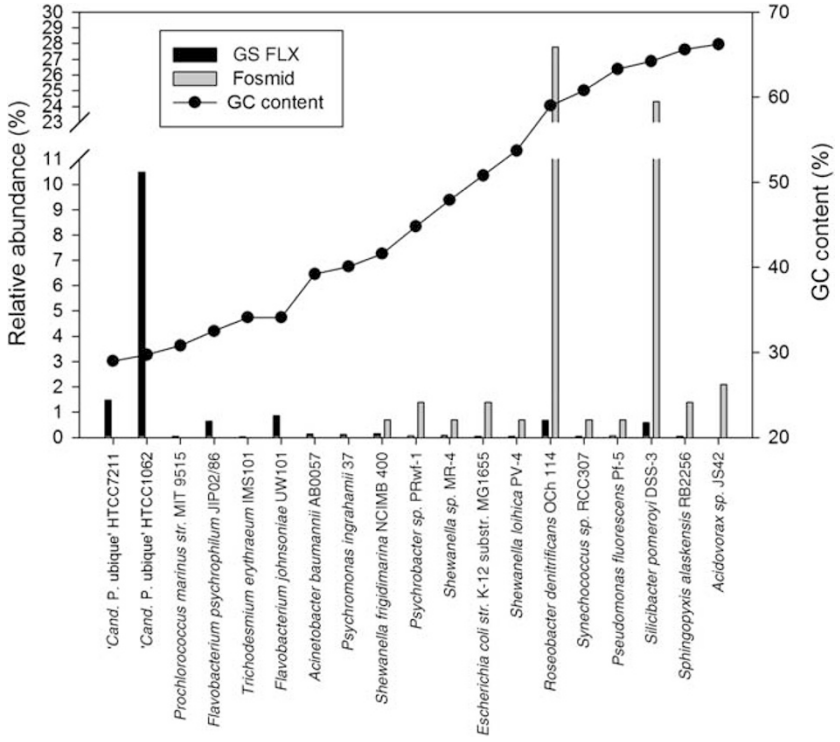

Figure 1 Relative abundance of homologues to 19 taxonomic representatives in pyrosequenced and fosmid data derived from the same water sample. Average GC content of each species is also plotted.

pyrosequenced data (37.8%) was similar to that of GOS Bay of Fundy (35.4%) and Brown Bank (37.4%), but significantly different to the 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, *R. denitrificans* OCh114, *Synechococcus* sp. and *Shewanella* sp. available from the 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 with <3% of sequences from the fosmid library, confirming the dominance of this species in coastal waters and its poor representation in fosmid data sets. The high degree of genetic conservation in HTCC1062, brought about through genomic streamlining (Giovannoni *et al.*, 2005), 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 earlier studies (Wilhelm *et al.*, 2007; Rusch *et al.*, 2007; Gilbert *et al.*, 2008) (Figure 3).

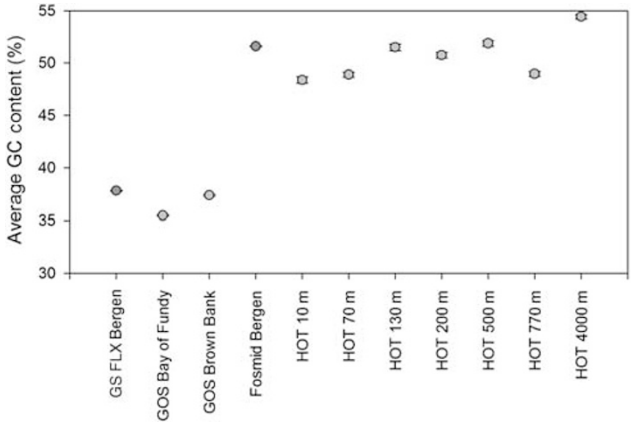

Figure 2 Mean GC content for the following samples: pyrosequenced GS FLX data from Bergen; surface water from the Bay of Fundy, prepared using a plasmid library (Rusch *et al.*, 2007*)*; surface water from Brown Bank, Gulf of Maine, prepared using a plasmid library (Rusch *et al.*, 2007*)*; fosmid library from Bergen; samples from HOT station ALOHA, prepared using a fosmid library taken at different depths (DeLong *et al.*, 2006); Bars represent 99% confidence intervals. Data prepared for this paper are highlighted in red. A full colour version of this figure is available at *The ISME Journal* online.

Similar recruitment (12.6% vs 2.8%) was observed against '*Cand.* P. ubique' HTCC7211, but at a lower percentage identity than that against HTCC1062, indicating that these two coastal and open-ocean SAR11 strains are similar but also have niche-specific genetic differences. It is interesting

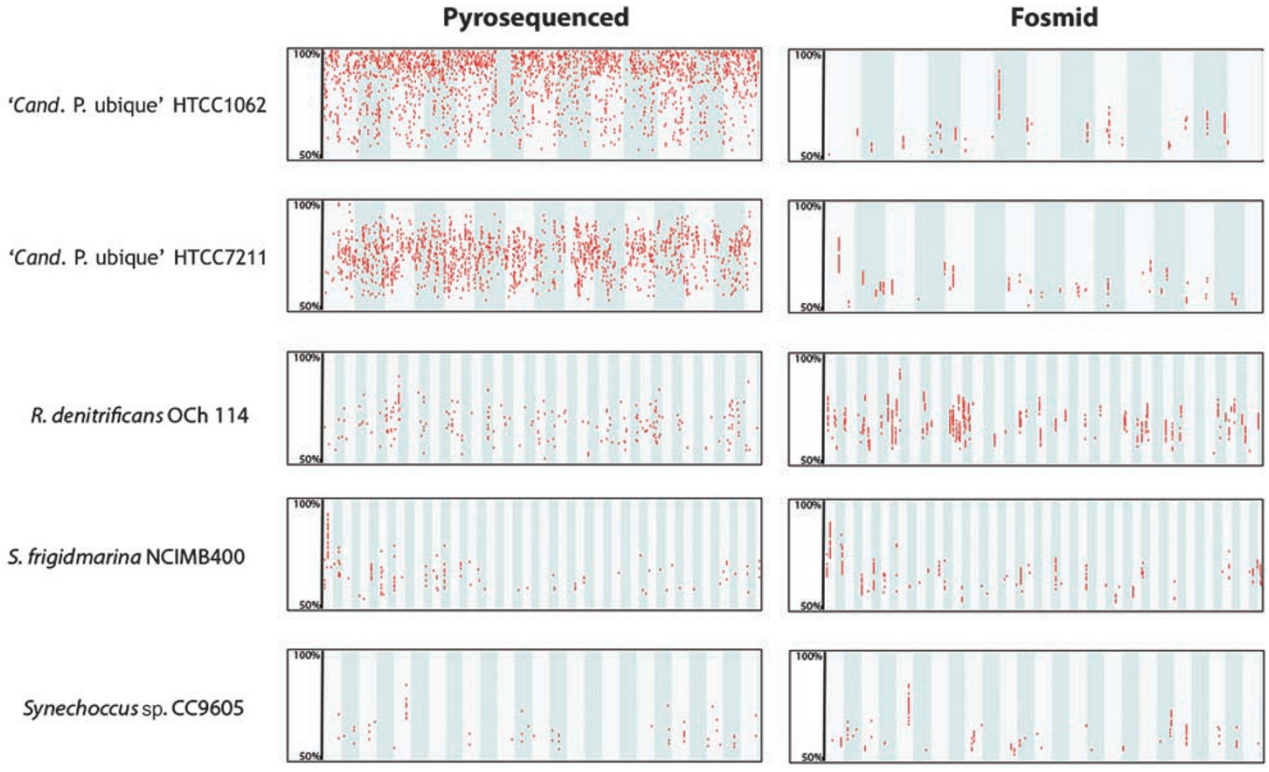

Figure 3 Fragment recruitment plots for key bacterioplankton species. Sequences from pyrosequenced and fosmid data were aligned against the reference genomes using BLASTN, with parameters designed to detect distant homologues as low as 65% identity. '*Cand.* P. ubique' HTCC7211 is included as an open-ocean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

that 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. Approximately 21.5% of fosmid sequences recruited to *R. denitrificans* OCh114, compared with 4.3% of hits for pyrosequenced data, but at a 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 the light of its global ubiquitous distribution (Scanlan and West, 2002) and its recruitment in the fosmid data (3.6%), most likely because of the low coverage of the pyrosequenced metagenome, estimated at ~0.0002%, using a calculated effective genome size (Raes *et al.*, 2007) of 2.33 Mb. Even at such a 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 shown that fosmid libraries do not accurately represent bacterial taxonomic diversity within a given community in comparison with pyrosequenced data sets. Fosmid libraries, both from this study and from a depth profile study, (DeLong *et al.*, 2006) showed significant over- or underrepresentation of important microbial taxa when compared with pyrosequenced data. Specifically, the low-GC SAR11 clade and Flavobacteria are significantly underrepresented, resulting in an elevated GC content of fosmid data in comparison with direct pyrosequencing. One possible contributing factor is that fragmentation of DNA occurs more readily in DNA with fewer GC linkages because of 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. This may also explain why the small insert libraries of the GOS data set do not appear to suffer from raised GC content. Other factors affecting strand stability, such as nearest-neighbour effects (Allawi and SantaLucia, 1998), may also be important.

Recent work by Pham *et al.* (2008) suggested that low numbers of SAR11 sequences detected in surface water fosmid libraries might be an accurate representation of the bacterial community, and that overrepresentation of SAR11 in earlier studies could be because of PCR cloning bias. However, no PCR cloning was used in this study, and the percentage hits to SAR11 from the sequencing of random fragments were similar to those enumerated using fluorescent *in situ* hybridization in surface water in other studies (Mary *et al.*, 2006). It is clear that large fragment-length inserts of fosmid libraries confer a

huge advantage when investigating genetic neighbourhoods. However, the exclusion of sequences, including those from the most ubiquitous bacterial clade, suggests that the use of fosmid libraries for metagenomic diversity studies must be carried out with caution to avoid the possibility of misrepresenting key components of microbial diversity.

Acknowledgements

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 an NERC grant (NE/C507902/1).

References

Allawi HT, SantaLucia J. (1998). Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry* **37**: 2170–2179.

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

Buchan A, González JM, Moran MA. (2005). Overview of the marine roseobacter lineage. *Appl Environ Microbiol* **71**: 5665–5677.

Campbell BJ, Stein JL, Cary SC. (2003). Evidence of chemolithoautotrophy in the bacterial community associated with *Alvinella pompejana*, a hydrothermal vent polychaete. *Appl Environ Microbiol* **69**: 5070–5078.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

Gilbert JA, Mühling M, Joint I. (2008). A rare SAR11 fosmid clone confirming genetic variability in the '*Candidatus* Pelagibacter ubique' genome. *ISME J* **2**: 790–793.

Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL *et al.* (2005). Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**: 82–85.

Hughes DS, Felbeck H, Stein JL. (1997). A histidine protein kinase homolog from the endosymbiont of the hydrothermal vent tubeworm *Riftia pachyptila*. *Appl Environ Microbiol* **63**: 3494–3498.

Kim UJ, Shizuya H, de Jong PJ, Birren B, Simon MI. (1992). Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res* **20**: 1083–1085.

Mary I, Heywood JL, Fuchs BM, Amann R, Tarran GA, Burkhill PH *et al.* (2006). SAR11 dominance among metabolically active low nucleic acid bacterioplankton in surface waters along an Atlantic Meridional Transect. *Aquatic Microb Ecol* **45**: 107–113.

Pham VD, Konstantinidis KT, Palden T, DeLong EF. (2008). Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical profile in the North Pacific Subtropical Gyre. *Environ Microbiol* **10**: 2313–2330.

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**: R10.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The sorcerer II global ocean sampling expedition: Northwest Atlantic through eastern tropical pacific. *PLoS Biol* **5**: e77.

Scanlan DJ, West NJ. (2002). Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiol Ecol* **40**: 1–12.

Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV. (1998). Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. *J Bacteriol* **180**: 5003–5009.

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol* **178**: 591.

Suzuki MT, Preston CM, Béjà O, de la Torre JR, Steward GF, DeLong EF. (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.

Suzuki MT, Rappé MS, Haimberger ZW, Winfield H, Adair N, Ströbel J *et al.* (1997). Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample. *Appl Environ Microbiol* **63**: 983–989.

Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ. (2007). Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* **2**: 27.