The ISME Journal (2009) 3, 1117–1119
© 2009 International Society for Microbial Ecology All rights reserved 1751-7362/09 $32.00
www.nature.com/ismej

COMMENTARY

Bias in assessments of marine SAR11 biodiversity in environmental fosmid and BAC libraries?

Roi Feingersch and Oded Béjà

The ISME Journal (2009) **3,** 1117–1119; doi:10.1038/ismej.2009.80; published online 2 July 2009

Introduction

With present estimates that suggest that more than 99% of microorganisms in different environments are resistant to our collective efforts to cultivation, environmental genomics seems to be the way to explore those hidden treasures (Béjà, 2004). The use of fosmid (F1 origin-based cosmid vector) and bacterial artificial chromosome (BAC) cloning vectors to clone environmental HMW DNA has facilitated the way to reach DNA fragments from unknown bacterial and archaeal groups (Rondon *et al*., 2000; Béjà *et al*., 2000a). These approaches have led to the identification of different metabolic pathways, among which is the finding of proteorhodopsins in marine bacteria (Béjà *et al*., 2000b). Recently, these libraries have also been used to estimate the diversity and metabolic potential of given environments (DeLong *et al*., 2006; Legault *et al*., 2006; Martín-Cuadrado *et al*., 2007).

Different studies have indicated that these libraries do contain a certain bias against the abundant cosmopolitan SAR11 group (Rappé and Giovannoni, 2003), with very few hits to SAR11 rRNA-containing clones (Béjà *et al*., 2000a; Suzuki *et al*., 2004; DeLong *et al*., 2006; Gilbert *et al*., 2008; Pham *et al*., 2008). This under-representation was originally suggested (Béjà *et al*., 2000a) to be attributed to the possible presence of 'toxic genes' in the region surrounding the SAR11 rRNA and the effect of 'clonability' (Sorek *et al*., 2007) to the *Escherichia coli* host cells. Indeed, previous reports have shown that the expression of genes cloned on BAC vectors can lead to an altered phenotype in the recombinant clones (Rondon *et al*., 1999). The possibility will always exist that, in any given population, some microbial DNA fragments may harbor genes toxic to *E. coli*, which will be under-represented in BAC libraries. Alternatively, this bias was recently attributed (Temperton *et al*., 2009) to the very low GC content of the currently available SAR11 core genomes (29.1–29.7% G + C; (Giovannoni *et al*., 2005; Wilhelm *et al*., 2007)), which may cause a fragmentation problem during the cloning procedure.

We here analyzed ∼10 000 random BAC ends from a recently analyzed Mediterranean BAC library (Sabehi *et al*., 2005; Feingersch *et al*., in preparation), this time with respect to SAR11 clone representation and possible bias.

Results and discussion

According to top BLASTx analysis of more than 10 000 random BAC ends with expectation cutoff values of $\leqslant 1 \times 10^{-50}$, the Mediterranean BAC library was composed of 18% SAR11-containing BACs (40% of ends in the Mediterranean Sea library were assigned to *Alphaproteobacteria,* of which 45% were SAR11) (Feingersch *et al*., in preparation). These estimates fit previous abundance measures of the SAR11 clade in Mediterranean coastal waters during the summer (∼20% of 4′,6-diamidino-2-phenylindole counts; Alonso-Sáez *et al*., 2007).

BAC ends were recruited on different '*Candidatus* Pelagibacter ubique' genomes (HTCC1062, HTCC1002 and HTCC7211) and, as could be seen in the example in Figure 1a, fragments were recruited almost evenly across the entire '*Cand*. P. ubique' HTCC1062 genome except to the already known regions of hypervariability (Rusch *et al*., 2007; Wilhelm *et al*., 2007; Gilbert *et al*., 2008). To check for possible bias against low-GC fragments, GC% content plots of the different ends were constructed. As could be seen in Figure 1b, the library was mainly composed of two different BAC-end populations, one population with a GC content above 50% and the other with a GC content of about 32%. Ends recruited on the '*Cand*. P. ubique' HTCC1062, HTCC1002 and HTCC7211 genomes were clearly assigned to the low-GC BAC-end population (gray area in Figure 1b). When the SAR11-recruited ends were BLASTed again, the proportion of SAR11 in alphaproteobacterial hits rose from 45% (18% of total hits) to 90% (78% of total hits) (Figure 1c). This

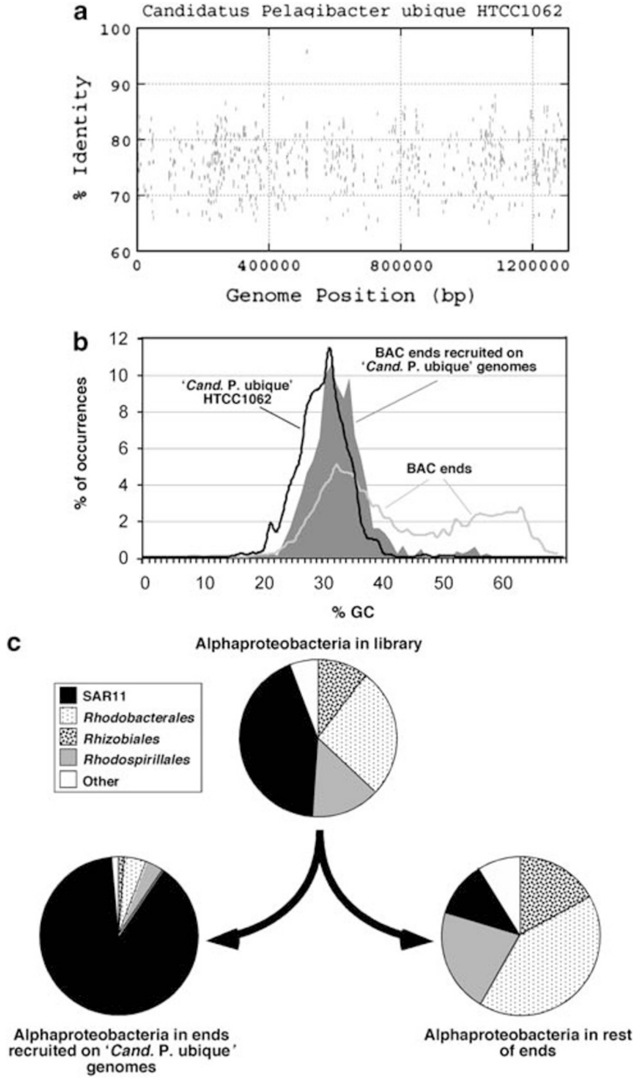

Figure 1 (**a**) Fragment recruitment plot of BAC ends to '*Cand.* P. ubique' HTCC1062 genome. BAC ends were aligned using the NCBI BLASTn. The recruitment was made with fragments of more than 200 bp, expectation values $<1\times10^{-5}$ (-p blastn -e 1e-5 -F F -r 2) and BLASTn High Score Pair (HSP) covers more than 80% of the hit. Data are shown using the Gnuplot program (http://www.gnuplot.info). Percent identity between the aligned sequence and the genomic sequence is shown on the *y* axis, whereas the position on which the sequence was aligned is shown on the *x* axis. (**b**) GC% distribution of the entire BAC library ends, GC% distribution of BAC ends recruited on '*Cand.* P. ubique' HTCC1062, HTCC1002 and HTCC7211 genomes, and GC% distribution of the '*Cand.* P. ubique' HTCC1062 genome (added as a reference). End sequences from the Mediterranean BAC library with recruitment results on any of the three SAR11 genomes (as in (**a**) for '*Cand.* P. ubique' HTCC1062) were GC% calculated and their distribution is shown as a gray shade. (**c**) *Alphaproteobacteria* distribution in Mediterranean BAC library ends or in SAR11-recruited ends based on BLASTx searches (-p blastx -e 1e-5 -F F) against the NCBI non-redundant (nr) database. The top HSP determined according to the NCBI taxonomic identifier was used to classify a query only if the expectation value was $\leqslant 1\times10^{-50}$. Sequences from the Mediterranean BAC library that were classified as *Alphaproteobacteria* were further grouped into 'recruited on SAR11 genome sequences' or 'not recruited on SAR11 genome sequences' and BLASTx querying was performed again. The results are shown in the two lower pies.

indicates that the population of low-GC ends in our BAC library is mostly composed of SAR11-like clones.

It is interesting to note that only one BAC end (out of 53 rRNA-containing ends) gave an SAR11 affiliation (Feingersch *et al.*, in preparation). This discrepancy between low proportions of SAR11 rRNA-containing BAC ends (1.9%) and the high proportion of SAR11 seen in the BLAST hits was also observed in surface stations of the ALOHA community genomics fosmid project (DeLong *et al.*, 2006). As already proposed by us in the year 2000, genes linked to the rRNA operon in SAR11 chromosomal DNA might be toxic to *E. coli*, even when present as a single copy on the BAC vector (Béjà *et al.*, 2000a). However, it is important to note that no evident toxic protein candidate emerged in the different SAR11 genomes that are currently available and there is yet no proven mechanistic explanation for these observations.

On the basis of our analyses, we suggest that the SAR11 group might be under-represented in fosmid and BAC libraries due to possible bias against toxic effects of some of their proteins and not as a result of the very low GC content of the SAR11 core genomes (29.1–29.7% G+C (Giovannoni *et al.*, 2005; Wilhelm *et al.*, 2007)) as suggested by Temperton *et al.* (2009). As there are many uncertainties in our current understanding of cloning biases as well as possible biases embedded in pyrosequencing, further sequencing and analyses of other BAC and fosmid libraries will be needed to resolve the SAR11-cloning bias enigma.

Acknowledgements

This work was partially supported by grants from the state of Lower Saxony and the Volkswagen Niedersachsen foundation.

R Feingersch and O Béjà are at the Faculty of Biology, Technion—Israel Institute of Technology, Haifa, Israel
E-mail: beja@tx.technion.ac.il

References

Alonso-Sáez L, Balagué V, Sà EL, Sánchez O, González JM, Pinhassi J *et al.* (2007). Seasonality in bacterial diversity in north-west Mediterranean coastal waters: assessment through clone libraries, fingerprinting and FISH. *FEMS Microbiol Ecol* **60**: 98–112.

Béjà O. (2004). To BAC or not to BAC: marine ecogenomics. *Curr Opin Biotechnol* **15**: 187–190.

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

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

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U *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, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242–1245.

Legault BA, Lopez-Lopez A, Alba-Casado JC, Doolittle WF, Bolhuis H, Rodriguez-Valera F *et al.* (2006). Environmental genomics of '*Haloquadratum walsbyi*' in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics* **7**: 171.

Martín-Cuadrado AB, López-García P, Alba JC, Moreira D, Monticelli L, Strittmatter A *et al.* (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* **2**: e914.

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.

Rappé MS, Giovannoni SJ. (2003). The uncultured microbial majority. *Annu Rev Microbiol* **57**: 369–394.

Rondon MR, Raffel SJ, Goodman RM, Handelsman J. (1999). Toward functional genomics in bacteria: analysis of gene expression in *Escherichia coli* from a bacterial artificial chromosome library of *Bacillus cereus*. *Proc Natl Acad Sci USA* **96**: 6451–6455.

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR *et al.* (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol* **66**: 2541–2547.

Rusch DB, Halpern AL, Heidelberg KB, Sutton G, Williamson SJ, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: I, The northwest Atlantic through the eastern tropical Pacific. *PLoS Biol* **5**: e77.

Sabehi G, Loy A, Jung KH, Partha R, Spudich JL, Isaacson T *et al.* (2005). New insights into metabolic properties of marine bacteria encoding proteorhodopsins. *PLoS Biol* **3**: e173.

Sorek R, Zhu YW, Creevey CJ, Francino MP, Bork P, Rubin EM. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**: 1449–1452.

Suzuki MT, Preston CM, Béjà O, de la Torre RJ, Steward GF, DeLong EF. (2004). Quantitative phylogenetic screening of ribosomal RNA gene-containing clones in Bacterial Artificial Chromosome (BAC) libraries from different depths in Monterey Bay. *Microbial Ecol* **48**: 473–488.

Temperton B, Field D, Oliver A, Tiwari B, Mühling M, Joint I *et al.* (2009). Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. *ISME J* (doi:10.1038/ismej.2009.32).

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.