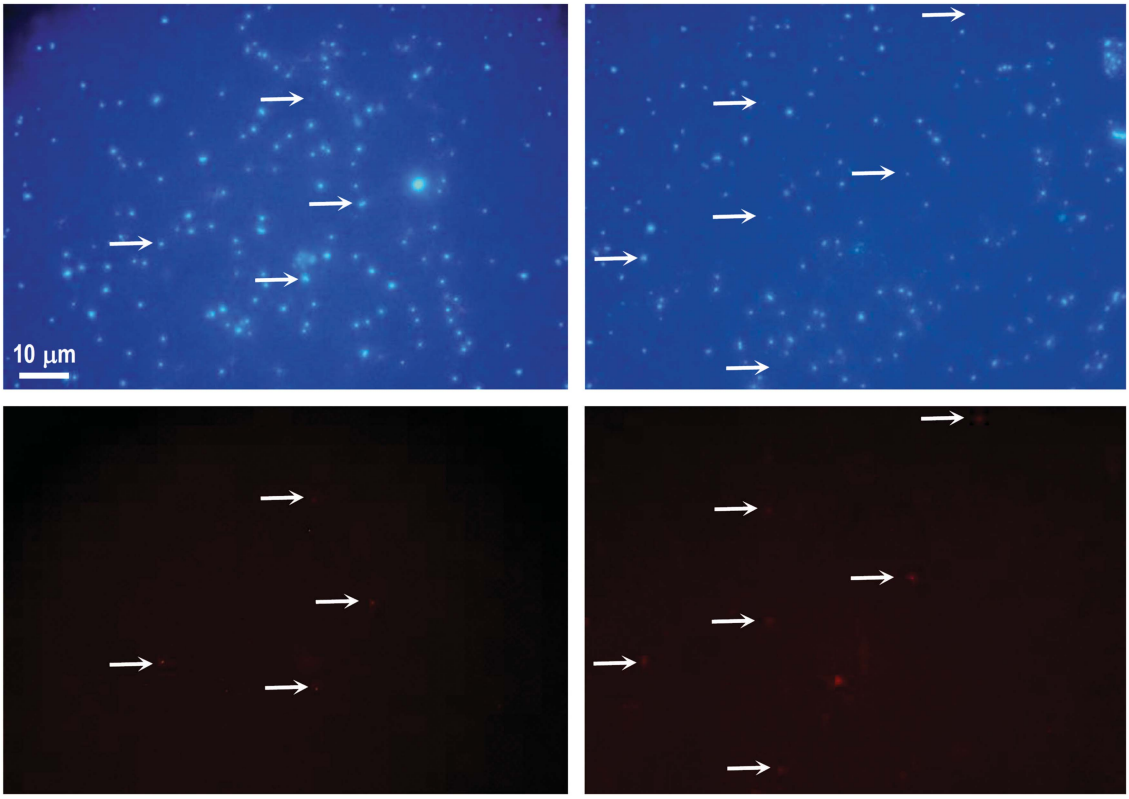

Figure 2 Microscopy of MGII cells. DCM seawater sample from Mediterranean Sea hybridized with fluorescein-labeled 16S rRNA group II archaeal probes (left panel, 'ThaMar' probe and right panel, 'Eury 806' probe). Images of the same field were captured using the fluorescein filter set (lower panels) and the DAPI filter set (upper panels). Scale bars, 10 µm.

The identity of the two thalassoarchaeal 16S rRNAs (94%) indicates that more than one species were present simultaneously in the fosmid collection. In addition, we found several contigs from overlapping genome regions that were largely syntenic but which average nucleotide identity (ANI) ranged from 98% to 80% (Figure 4), i.e. consistent with more than one species being present. To establish how many we searched among the DNA fragments for different versions of 36 housekeeping genes (Supplementary Table 3). The results indicate that although there are clearly more than one abundant species, their number is likely not very high, probably two, and diverging less than 20% ANI, i.e. within a single genus. The use of other sequence parameters such as codon usage, pentanucleotide frequencies, %GC or the coverage in the metagenomes did not permit to separate them.

We analyzed the contigs for the presence of 35 (Raes *et al.*, 2007) or 100 (Albertsen *et al.*, 2013) previously defined orthologous markers to estimate the completeness of the assembled genomes. Using these criteria, the thalassoarchaeal genomic fragments retrieved represent between 78 and 100% of a complete genome. By the same extrapolation, the estimated genome length of an individual genome of this group would be ~2 Mb, very similar to the estimated size of the MG2-GG3 genome (Iverson *et al.*, 2012). A total of 4074 ORFs could be identified in the thalassoarchaeal fosmids. Most hits (61%) against the nr database (Genbank) were to Euryarchaeota, 16% to Bacteria and 2% to eukaryotes (the remaining 21% were unclassified). The corresponding proteins were clustered using CD-HIT at 50% similarity and coverage resulting in a smaller data set of 2435 non-redundant proteins. This set was compared to the 1698 proteins of the MG2-GG3 genome (using a reciprocal best blast hit analysis) and 1427 proteins were found to share more than 50% similarity. However, the average similarity was low (65%). Only 639 homologs were found within the 1544 proteins of *A. boonei* T469 (average similarity 60%). The SAG SCGC-AAA-288-C18 from 700 m deep in the central North Pacific also had a low similarity (average 70%). The highest similarities were found with the proteins of the large set of MGII fosmids retrieved from bathypelagic Mediterranean samples (Ionian Sea 3000 m, 324 fosmids; and Adriatic Sea 1000 m, 139 fosmids) (Deschamps *et al.*, 2014). Of the non-redundant thalassoarchaeal proteins 1599 had a hit with Adriatic (38% of their total) and 1850 with the Ionian fosmids. However, synteny was seldom conserved and average similarity was of 74 and 73% respectively, suggesting that the deep MGII belong to very different taxa.

Inferred metabolic and ecological features

We found genes coding for enzymes for glycolysis, the tricarboxylic acid cycle and oxidative