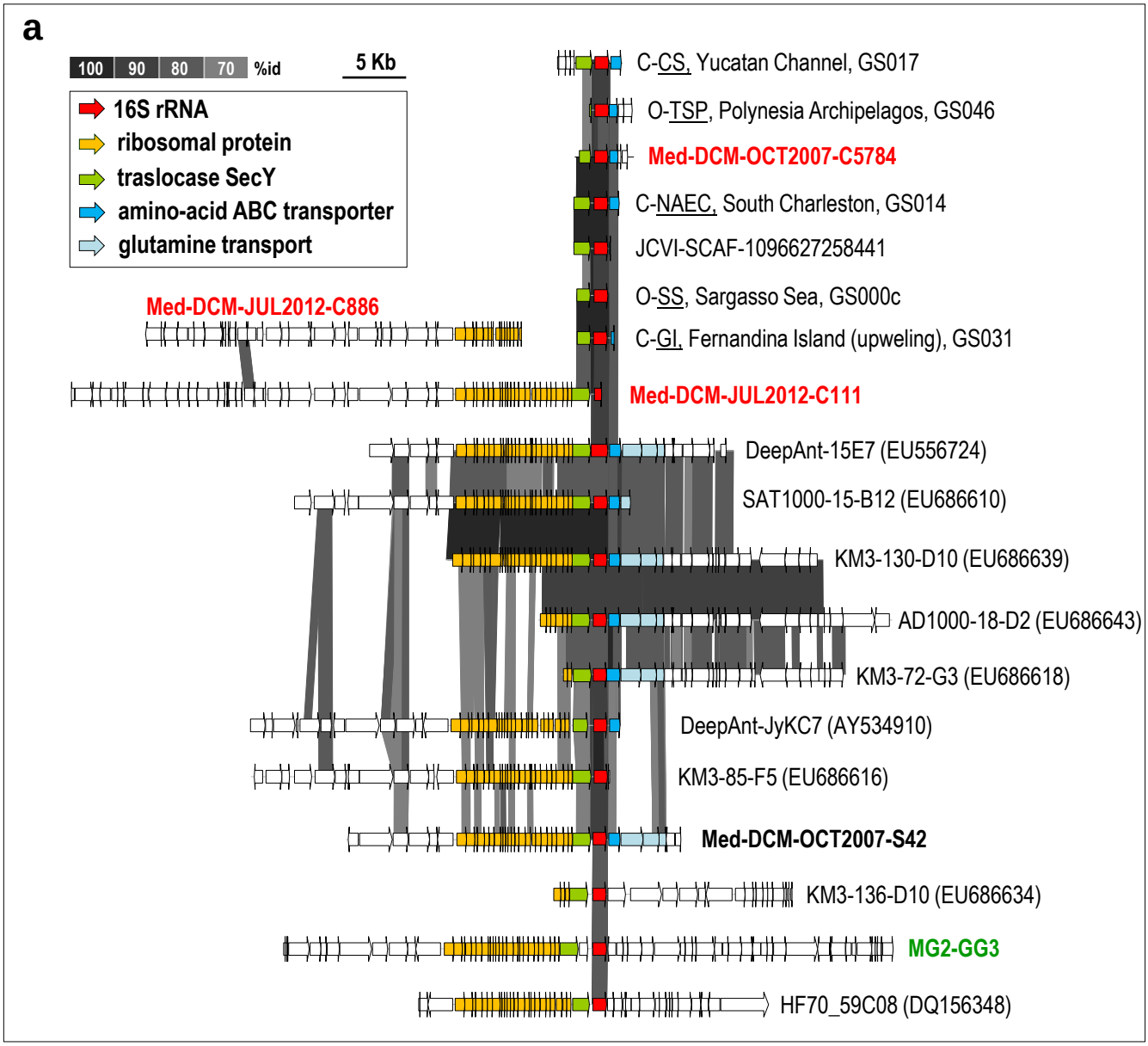
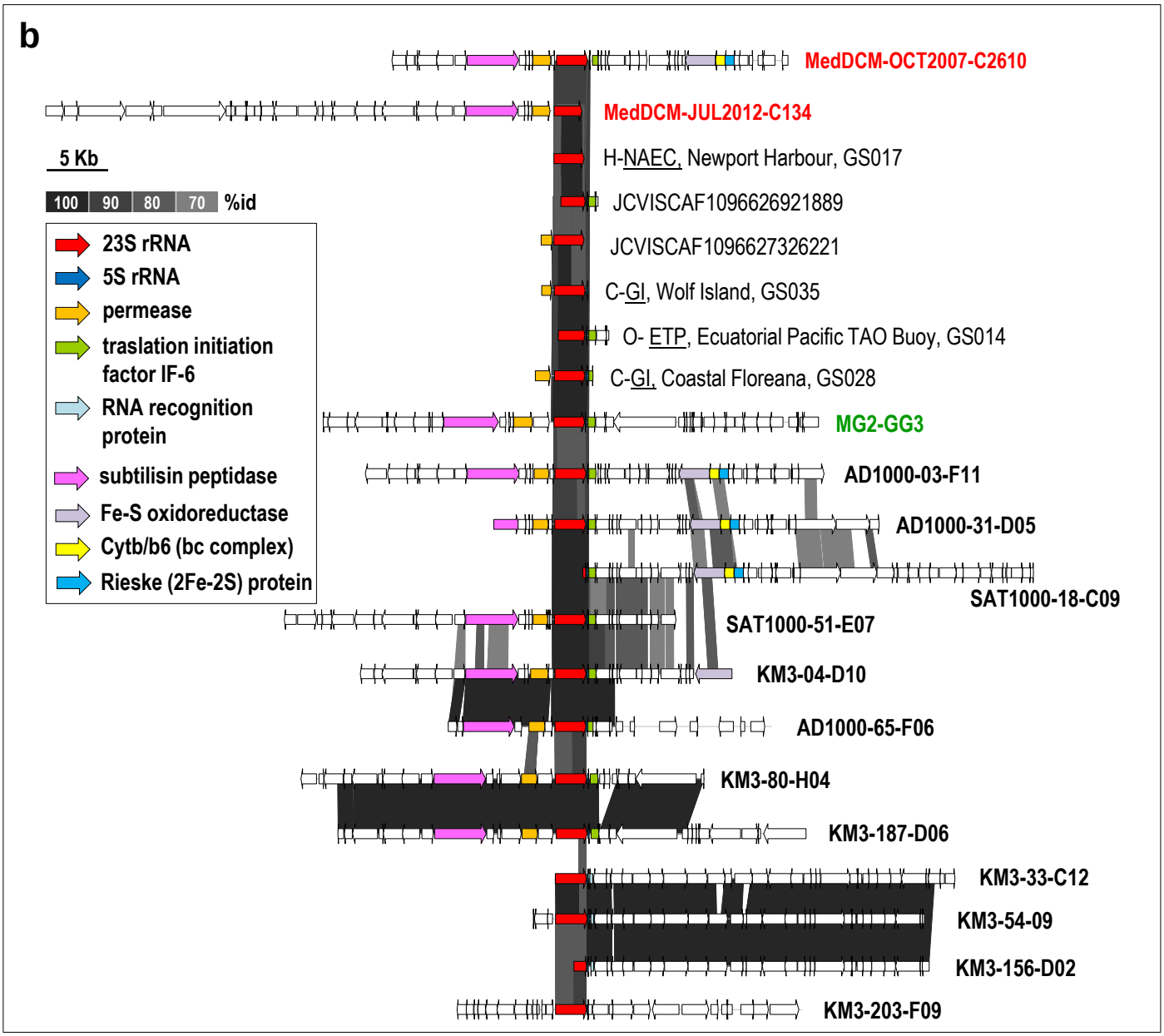
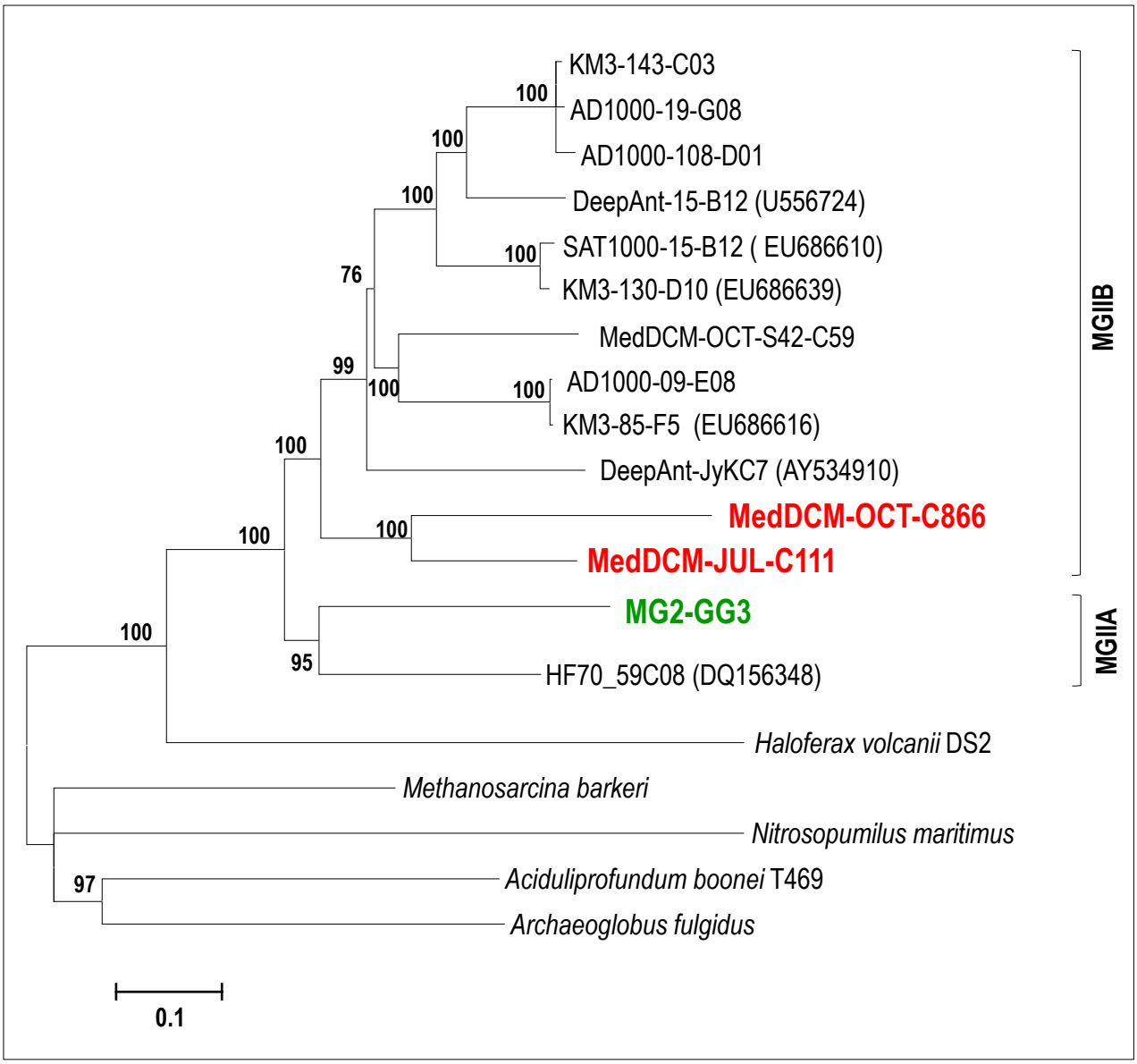


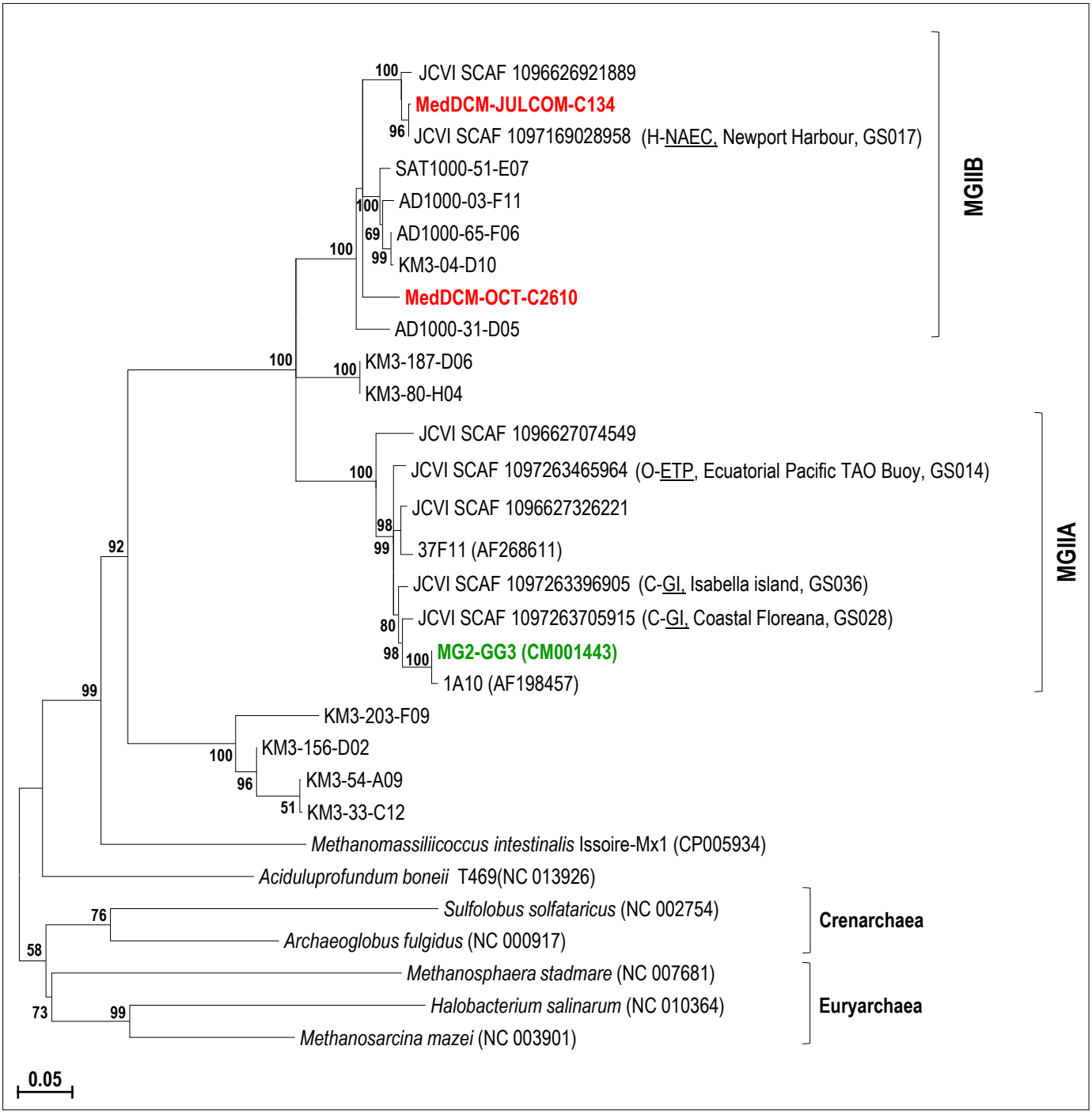
Supplementary Figure 1. Comparison of marine low GC thalassoarchaeal contigs containing rRNA genes (bold and red) to scaffolds from the Global Ocean Sampling (GOS) dataset and other MGII described in literature literature (MG2-GG3 in bold and green) (using BLASTN). The oceanic habitat (C-Coastal, CRA-Coral Reef Atoll, O-Open Ocean, E-Estuary), sampling locations (NAEC-North American East Coast, GI-Galapagos Islands, ETP-Eastern Tropical Pacific, PA-Polynesia Archipelagos, SS-Sargasso Sea, CS-Caribbean Sea) and the GOS dataset identifier are shown next to each GOS scaffold. All ribosomal RNA genes are highlighted in red color and sequence identity amongst the contigs is shown in shades of grey (see color scale). **(a)** Contigs containing 16S rRNA. **(b)** Contigs containing 23S rRNA.



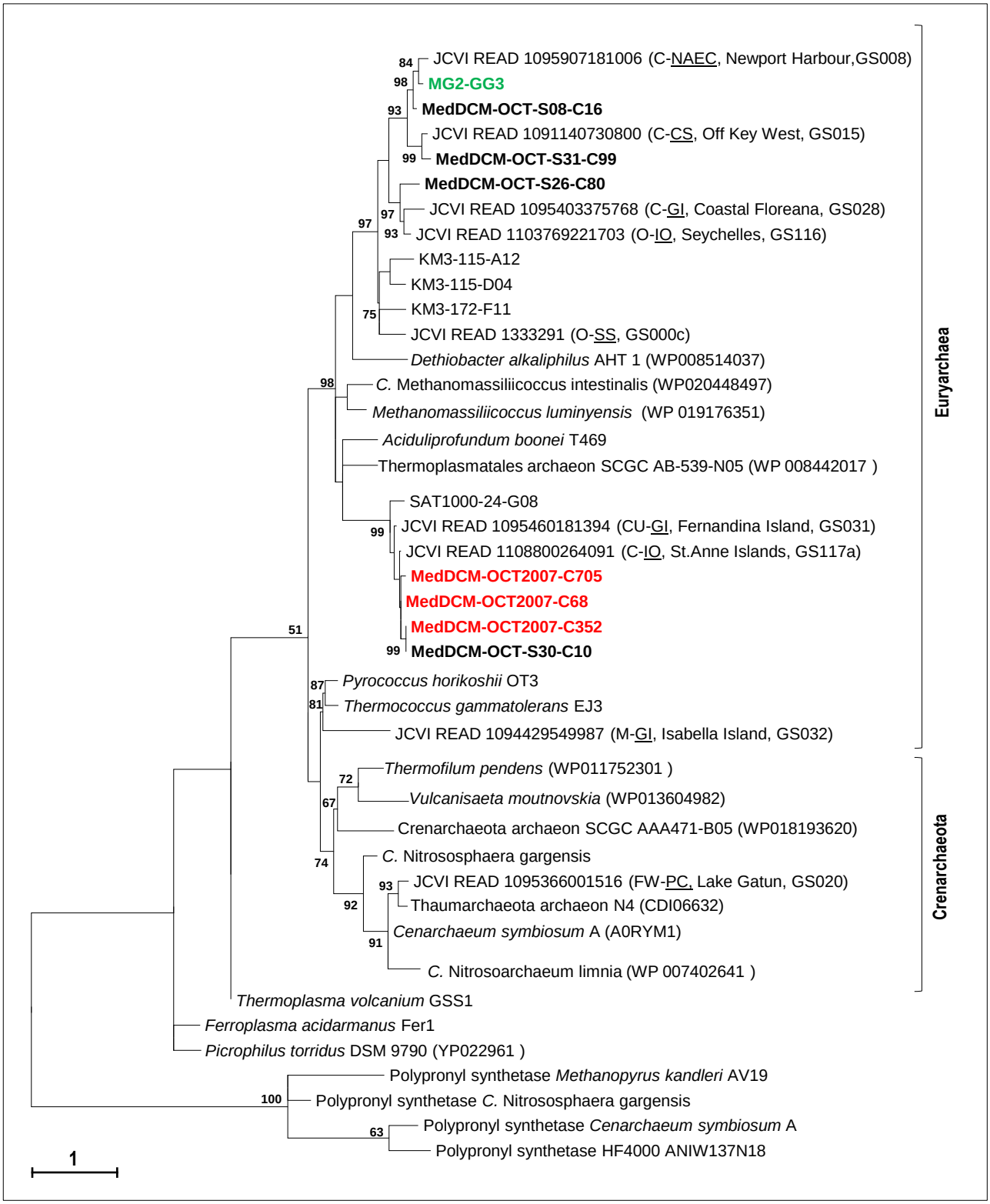
Supplementary Figure 1. Comparison of marine low GC thalassoarchaeal contigs containing rRNA genes (bold and red) to scaffolds from the Global Ocean Sampling (GOS) dataset and other MGII described in literature (MG2-GG3 in bold and green) (using BLASTN). The oceanic habitat (C-Coastal, CRA-Coral Reef Atoll, O-Open Ocean, E-Estuary), sampling locations (NAEC-North American East Coast, GI-Galapagos Islands, ETP-Eastern Tropical Pacific, PA-Polynesia Archipelagos, SS-Sargasso Sea, CS-Caribbean Sea) and the GOS dataset identifier are shown next to each GOS scaffold. All ribosomal RNA genes are highlighted in red color and sequence identity amongst the contigs is shown in shades of grey (see color scale). **(a)** Contigs containing 16S rRNA. **(b)** Contigs containing 23S rRNA.



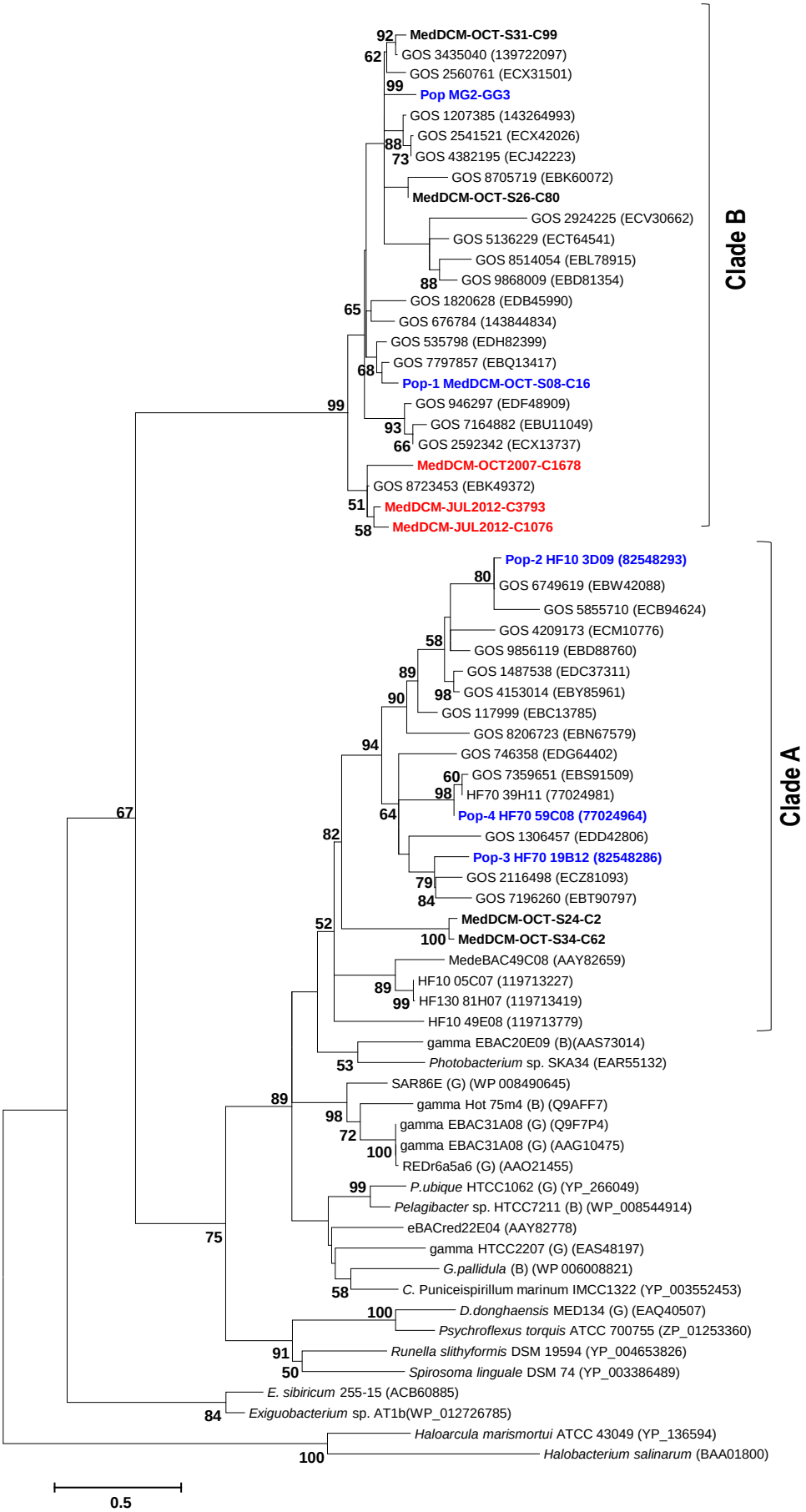
Supplementary Figure 2. Ribosomal proteins concatenate. Maximum-likelihood phylogenomic tree of 22 concatenated universally conserved ribosomal proteins (13116 unambiguously aligned base pair positions) from the thalassoarchaeal contigs (bold and red) compared with other Euryarchaea genomes representatives (MG2-GG3 in bold and green). Numbers at nodes represent bootstrap values from 1000 replications implemented in MEGA 5.10. Scale bar indicates estimated number of substitutions per site.



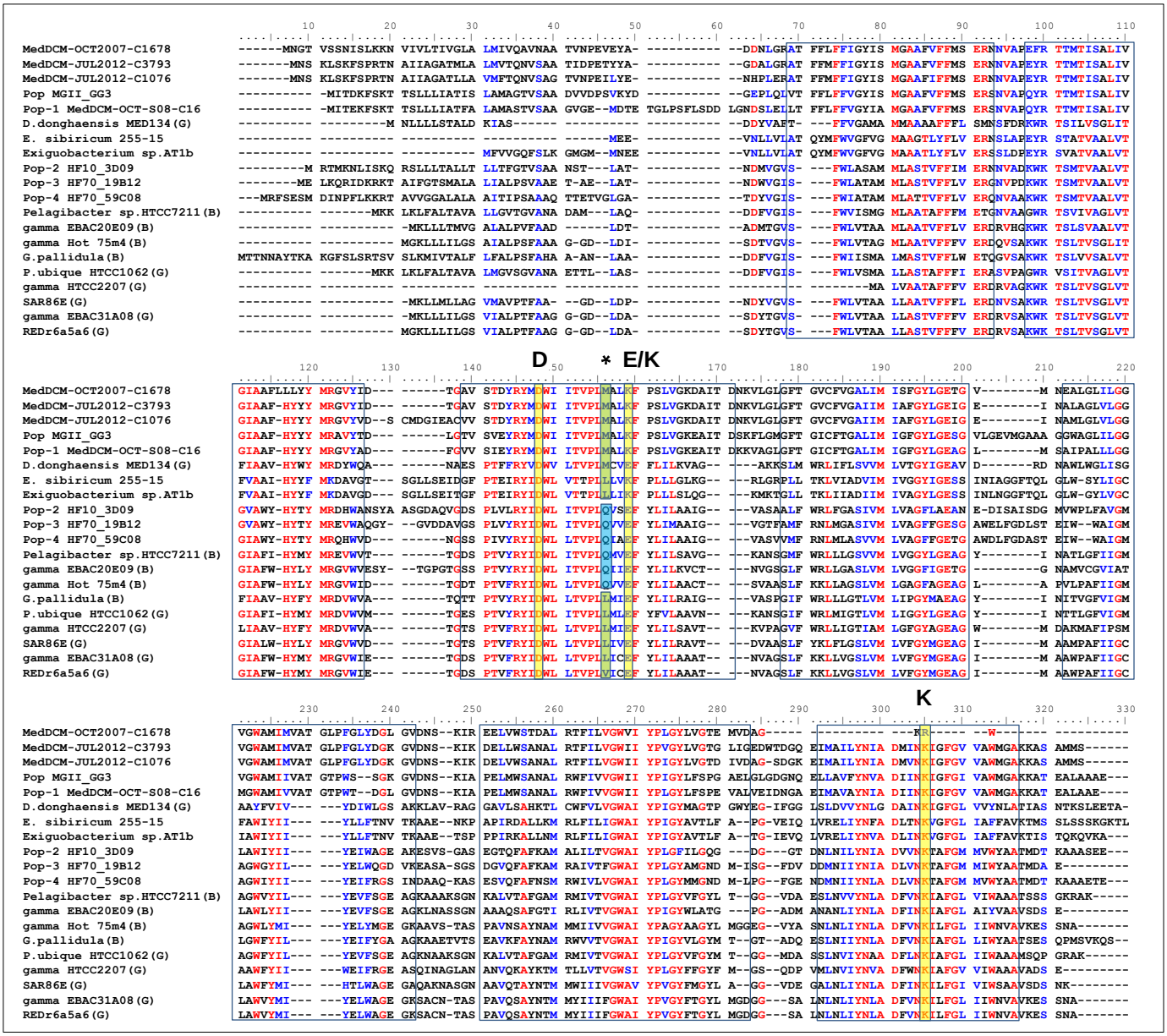
Supplementary Figure 3. 23S ribosomal RNA phylogeny. Maximum-likelihood 23S rRNA gene tree (2520 unambiguously aligned nucleotides) showing the relationship of the Thalassoarchaea (bold and red) with other MGII (MG2-GG3 in bold and green). Along with the rRNA sequences from the fosmids from the Mediterranean DCM, we have also included several rRNA sequences from GOS scaffolds where we could identify complete 23S genes. Numbers at nodes in major branches indicate bootstrap support (shown as percentages and only those >50%) by ML in MEGA 5.10. Scale bar represents the estimated number of substitutions per site. Sampling locations: MED, Mediterranean Sea, ETP-Eastern Tropical Pacific, NAEC-North American East Coast, GI-Galapagos Islands. The GOS dataset identifiers are shown next to each GOS scaffold.



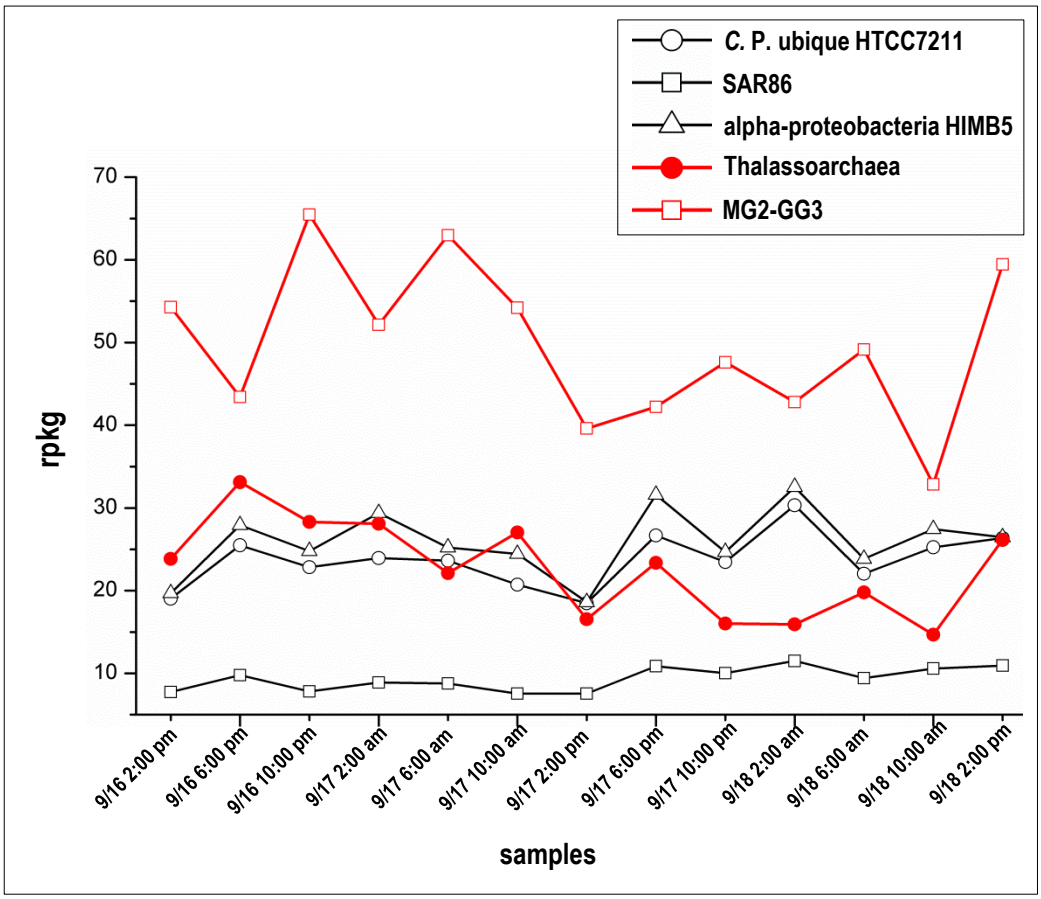
Supplementary Figure 4. Geranylgeranylglyceryl phosphate synthase phylogenetic tree. Thalasoarchaeal sequences are mark in bold and red. In bold and black, sequences previously described in Rohit et al. (2011) and in green, the MGIIA MG2-GG3 homolog.



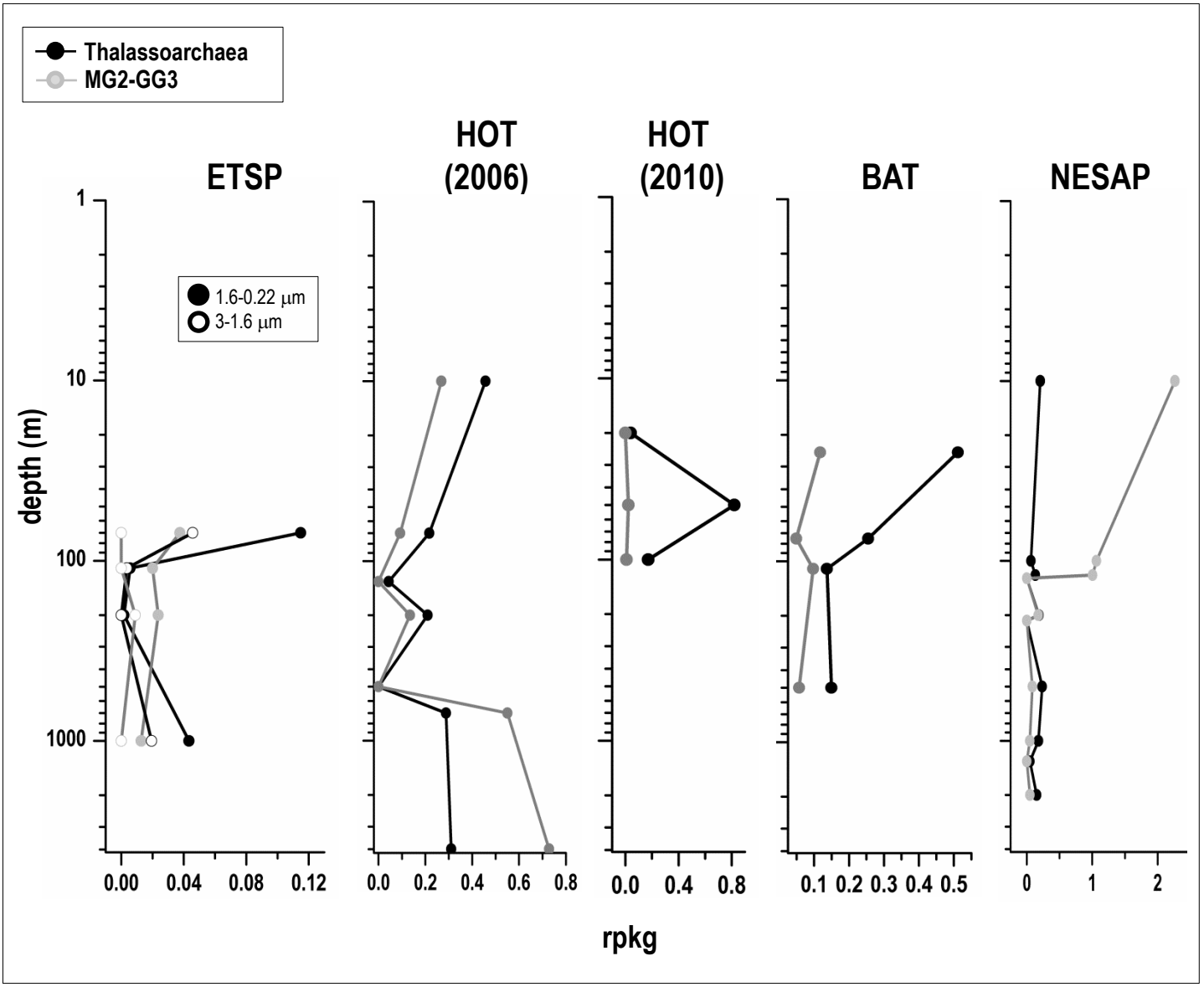
Supplementary Figure 5. Rhodopsin phylogeny. Maximum-likelihood rhodopsin gene tree showing the relationship of the thalassorhodopsin (bold and red) with other rhodopsins. In blue are marked Pop, Pop-1, Pop-2, Pop-3 and Pop-4 euryarchaeal rhodopsins previously described. Along with the sequences from the fosmids from the Mediterranean DCM, we have also included several rhodopsins sequences from GOS scaffolds where we could identify the complete gene. Numbers at nodes in major branches indicate bootstrap support (shown as percentages and only those >50%) by ML in MEGA 5.10. Scale bar represents the estimated number of substitutions per site. The GOS dataset identifiers are shown next to each GOS scaffold.



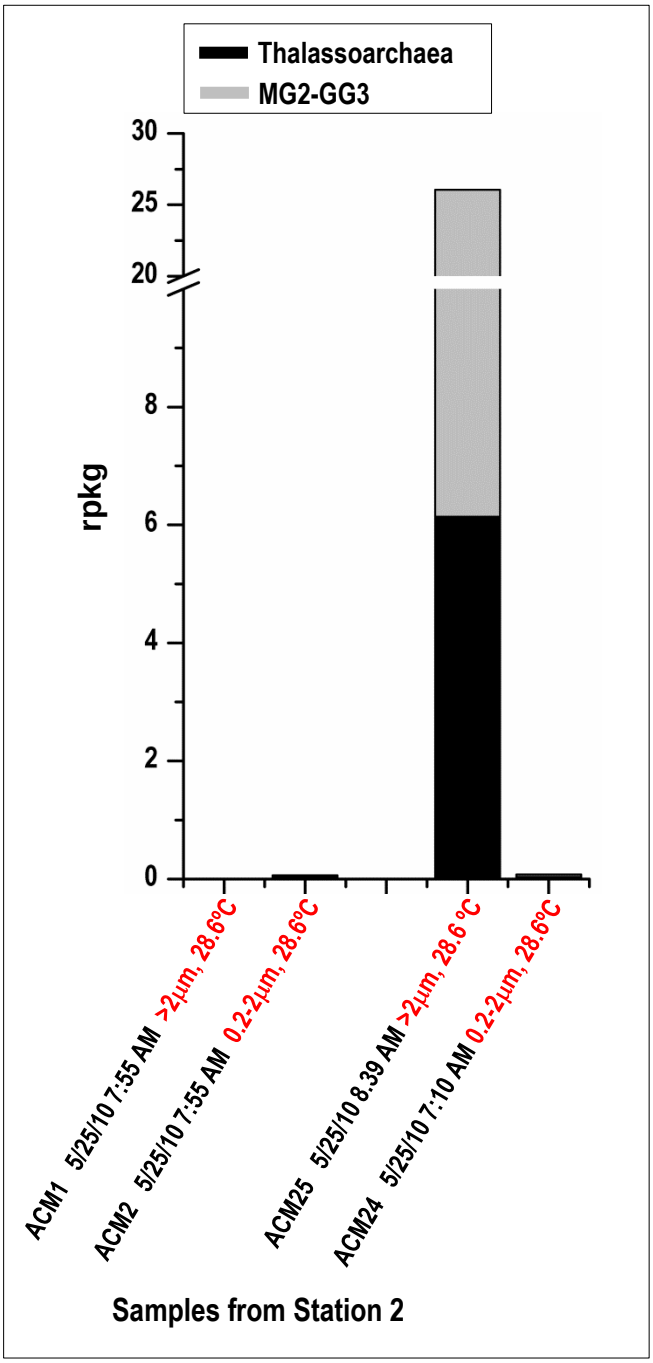
Supplementary Figure 6. Alignment of the thalassorhodopsins with the rhodopsin of *D. donghaensis* MED134 and other cloned rhodopsins sequences. Predicted transmembrane helices are marked by boxes. Identical residues are indicated in red. Residues in blue are conserved in more than 70% of the sequences. Key amino acids for rhodopsins functionality (listed herein with EBAC31A08 numbering) are marked by colors: Lys131 (K) binds retinal, and Asp97 (D) and Glu108 (E) function as Schiff base proton acceptor and donor, respectively. Metionin (M) in position 105 (*) in the thalassorhodopsins sequences indicate an absorption maxima at the green spectrum range. Letters (G) and (B) in the name of the sequences indicate the range of the spectrum. (The GenBank accession numbers of the sequences used for the alignment are as follows: *Dokdonia donghaensis* MED134, ZP_01049273; Pop-2 HF10_3D09, 82548293; Pop-3 HF70_19B12, 82548286; Pop-4 HF70_59C08, 77024964; *G. pallidula*, WP_006008821; *C.* Pelagibacter ubique HTCC1062 (SAR11), YP_266049; Pelagibacter sp. HTCC7211, WP_008544914; eBAC20E09, AAS73014; gammaproteobacteria HTCC2207 (SAR92), EAS48197; eBAC31A08 (SAR86), WP008490645; HOT75m4, AAK30179; REDr6a5a6, AAO21455).



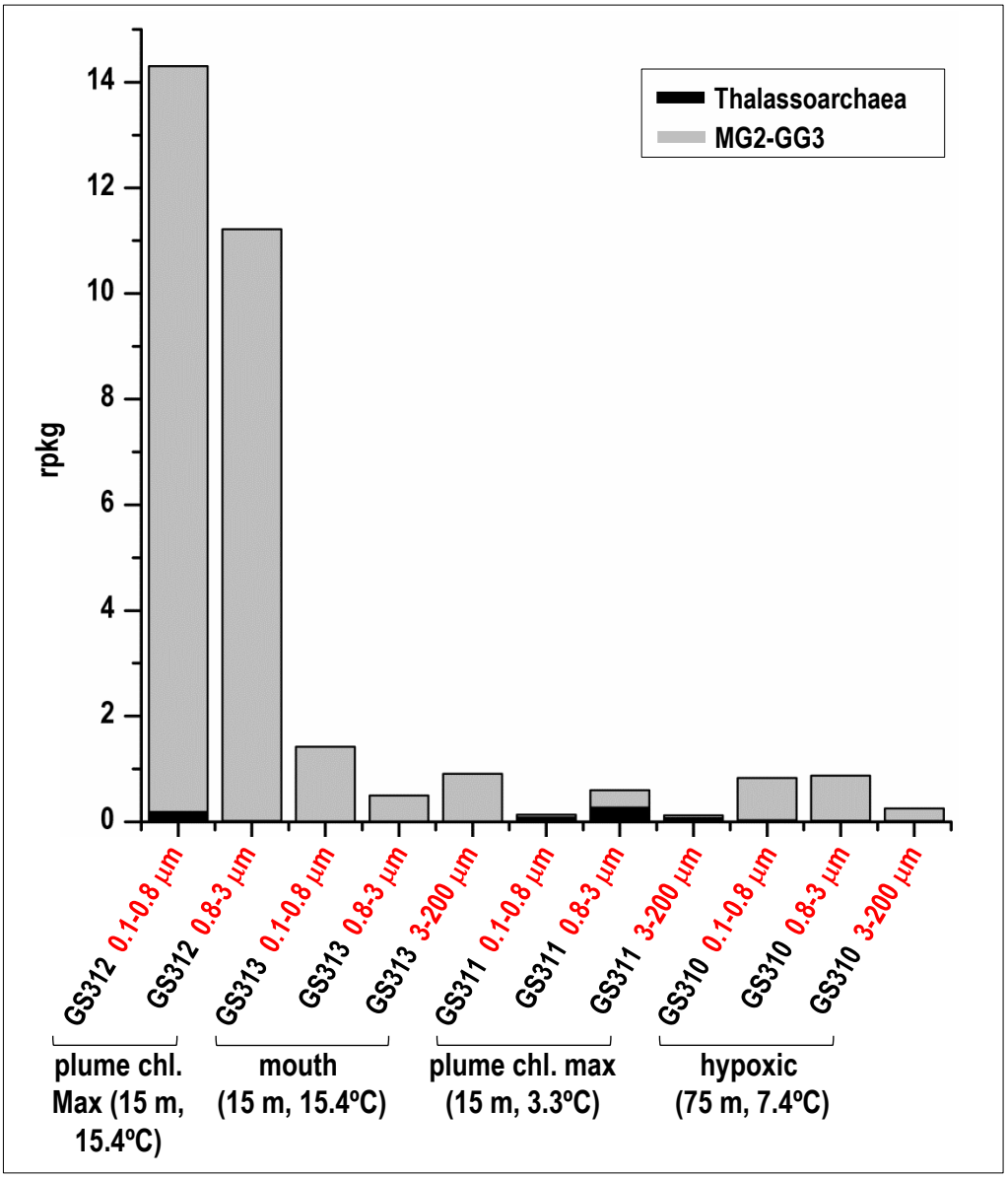
Supplementary Figure 7. Relative abundance of Thalassoarchaea and other marine microbes transcripts in the serial metatranscriptomes recovered in a coastal upwelling in the Coast from North California (Ottesen et al. 2013).



Supplementary Figure 8. Depth distribution of Thalassoarchaeales and MG2-GG3 determined by metagenomic fragment recruitment (BLASTN-based, see Methods). Abundance is expressed as number of reads per kilobase per gigabase of the collection (rpkg). HOT- Hawaii Ocean Time Series station ALOHA, NESAP-North Eastern Subarctic Pacific, ETSP-Eastern Tropical South Pacific and BATS-Bermuda Atlantic Time Series



Supplementary Figure 9. Depth distribution of Thalassoarchaeales and MG2-GG3 determined by metagenomic fragment recruitment (BLASTN-based, see Methods) in the Station 2 of the Amazon River plume (Satinsky et al. 2014). Abundance is expressed as number of reads per kilobase per gigabase of the collection (rpkg).



Supplementary Figure 10. Depth distribution of Thalassoarchaeales and MG2-GG3 determined by metagenomic fragment recruitment (BLASTN-based, see Methods) in the Columbia River stations (Smith et al. 2013). Abundance is expressed as number of reads per kilobase per gigabase of the collection (rpkg).