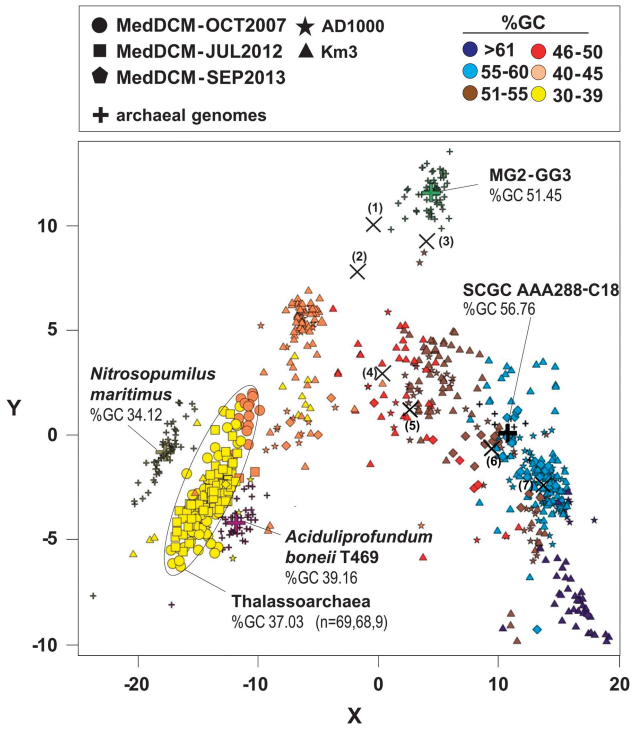

Figure 3 Comparison of MGII fosmids and contigs with known archaeal genomes. Principal component analysis of tetranucleotide frequencies of the low-GC MGII MedDCM-OCT2007 fosmids and the assembled low-GC MGII contigs from the MedDCM-JUL2012 and MedDCM-SEP2013 data sets. Reference genomes are shown as crosses (*N. maritimus*, *A. boonei* T469, MG2-GG3 and the SAG SCGC_AAA288-C18). Smaller crosses around the larger ones represent the tetranucleotide frequencies of 35 kb fragments from the same genome. Colors correspond to the %GC accordingly with the legend. The contigs classified as Thalassoarchaea are marked by a circle and the number of contigs from each of the data sets (MedDCM-OCT2007, MedDCM-JUL2012 and MedDCM-SEP2013) is indicated. The length of the assembled DNA fragments and %GC average are also shown. Numbers in brackets indicate MGII fosmids containing 16S rRNA sequences previously described: (1) *HF10-29C11*, %GC 45.56; (2) *HF10-3D09*, %GC 46.84; (3) *HF70-59C08*, %GC 51.36; (4) *HF70-19B12*, %GC 48.75; (5) *EF100-57A08*, %GC 50.66; (6) *DeepAnt-15E7*, %GC 56.03; (7) *HF4000-APKG2H5*, %GC 56.08).

phosphorylation. However, due perhaps to the incomplete nature of these genomes, not all genes could be found, e.g. Thalassoarchaea cells appear to contain many of the enzymes of the Embden–Meyerhof–Parnas (EMP) pathway for the metabolism of hexose sugars, with the exception of the first (glucokinase) and the last (pyruvate kinase). For the first, several carbohydrate kinases of unknown specificity found could serve as alternatives. For the second, a pyruvate phosphate dikinase was found that could operate in the catabolic direction, as have been described in some archaea (Tjaden *et al.*, 2006). Similar proteins were also found among the fosmids from the deep Mediterranean libraries (Deschamps *et al.*, 2014) but they were absent in the MG2-GG3 or any other MGII genomic fragment. Therefore, further investigations must be performed in this direction to clarify if the EMP functions in the gluconeogenic direction rather than the glycolytic pathway, as has been proposed for other Archaea (Hutchins *et al.*, 2001; Hallam *et al.*, 2006). Along these lines, typical gluconeogenesis enzymes such as pyruvate carboxylase (subunits A and B) and a gene coding for a phosphoenolpyruvate carboxykinase, both typical gluconeogenic enzymes, were found. As in other Euryarchaeota (Makarova *et al.*, 1999; Makarova and Koonin 2003; Hallam *et al.*, 2006) glucose 1-dehydrogenase, gluconolactonase and 2-keto-3-deoxy gluconate aldolase homologues are absent, suggesting that the Entner–Duodoroff hexose catabolic pathway is not present. In addition, we identified a complete non-oxidative pentose phosphate pathway, but the irreversible oxidative branch was missing. The reactions of the oxidative branch are important for generating NADPH, which is a source of reducing energy required by many enzymes in central biosynthetic pathways. However, we found genes for enzymes such as a 2,5-didehydrogluconate reductase, or a malate dehydrogenase that could act as alternatives for reducing $NADP^+$ to NADPH.

One of the metabolic features that we can infer is that these representatives of group IIB are facultative photoheterotrophs as seems to be the case of MG2-GG3. Three different rhodopsin genes were found among our contigs, sharing a similarity between 76 and 92% (Figure 5). Rhodopsin genes are widespread in the open oceans (Beja *et al.*, 2000) and are used as back up for heterotrophic energy generation using sunlight. These rhodopsin genes were neither adjacent to the 16S rRNA nor to the archaeal geranylgeranylglyceryl phosphate synthase genes (Figures 4 and 5) as found previously for other MGII fosmids (Frigaard *et al.*, 2006), including those assembled previously from the Mediterranean DCM (Ghai *et al.*, 2010). Two different genomic contexts were found for the thalassoarchaeal rhodopsins, likely corresponding to different species (Figure 5). Several GOS scaffolds were found to be syntenic to these rhodopsin containing fosmids, but none larger than 5 kb. The phylogenetic tree (Supplementary Figure 5) shows the rhodopsin of the thalassoarchaeal representatives to be closely related to the one found in MG2-GG3. Both are in a cluster (clade B) separate from other rhodopsins from bacterial or eukaryotic origin (clade A). We would like to suggest the term thalassorhodopsin to designate this group. Their key residues (listed herein with EBAC31A08 numbering), indicate some important differences with other related rhodopsins (Supplementary Figure 6). Residue 105, involved in spectral tuning, was a methionine, characteristic of green absorbing rhodopsins from shallow depths (Fuhrman *et al.*, 2008) of the Bacteriodetes proteorhodopsin isoform, such as in the marine flavobacterium *Dokdonia donghaensis* MED134 (Gomez-Consarnau *et al.*, 2007; Riedel *et al.*, 2010). It is known that different amino acid substitutions in residues aspartic-97 (D) and glutamic-108 (E), which function as Schiff base proton acceptor and donor in