doi:10.1038/nature12033

Predominant archaea in marine sediments degrade detrital proteins

Karen G. Lloyd[1,2]*, Lars Schreiber[1]*, Dorthe G. Petersen[1], Kasper U. Kjeldsen[1], Mark A. Lever[1], Andrew D. Steen[2], Ramunas Stepanauskas[3], Michael Richter[4], Sara Kleindienst[5], Sabine Lenk[5], Andreas Schramm[1] & Bo Barker Jørgensen[1]

Half of the microbial cells in the Earth's oceans are found in sediments[1]. Many of these cells are members of the Archaea[2], single-celled prokaryotes in a domain of life separate from Bacteria and Eukaryota. However, most of these archaea lack cultured representatives, leaving their physiologies and placement on the tree of life uncertain. Here we show that the uncultured miscellaneous crenarchaeotal group (MCG) and marine benthic group-D (MBG-D) are among the most numerous archaea in the marine sub-sea floor. Single-cell genomic sequencing of one cell of MCG and three cells of MBG-D indicated that they form new branches basal to the archaeal phyla Thaumarchaeota[3] and Aigarchaeota[4], for MCG, and the order Thermoplasmatales, for MBG-D. All four cells encoded extracellular protein-degrading enzymes such as gingipain and clostripain that are known to be effective in environments chemically similar to marine sediments. Furthermore, we found these two types of peptidase to be abundant and active in marine sediments, indicating that uncultured archaea may have a previously undiscovered role in protein remineralization in anoxic marine sediments.

In the cold anoxic sediments underlying most of the Earth's oceans, the only metabolisms known for cultured archaea are methane production from simple carbon substrates, and methane consumption[5]. Recent isotopic evidence, however, has shown that sedimentary archaea can be heterotrophic[6], but potential carbon substrates remain unknown. Intriguingly, detrital proteins are the largest components of marine organic matter that can be chemically characterized[7], and they are degraded slowly by thus-far-unidentified microbes in anoxic sediments[8]. If heterotrophic archaea were able to degrade proteins, such a finding would change our basic conception of the marine sedimentary carbon cycle, as it is generally assumed that bacteria drive the primary remineralization of complex organic matter.

Globally, marine subsurface archaea are often dominated by members of the MCG and MBG-D[9] (Fig. 1). We analysed these groups of archaea present in Aarhus Bay, Denmark (Supplementary Fig. 1), in organic-rich marine sediments where microbial activity is high at the surface but drops to low values similar to those found in deep oceanic sediments a few metres below the sea floor[10]. Here, MCG and MBG-D are abundant, based on 16S ribosomal RNA gene polymerase chain reaction (PCR) amplicon sequence libraries and quantitative PCR (Supplementary Fig. 2 and Supplementary Tables 1–3). Their abundance is independent of the major biogeochemical zones of sulphate reduction and methanogenesis, as has been noted previously for MCG[9], suggesting that they have different metabolic pathways from these types of anaerobic respiration.

However, the taxonomic marker 16S rRNA gene cannot be used to infer physiologies for MCG and MBG-D because no member or close relative of these groups has ever been grown in laboratory culture. We therefore obtained large genomic sequences from single cells to couple taxonomic genes to those encoding other cellular functions[11,12]. We used density gradient centrifugation to extract intact cells from a

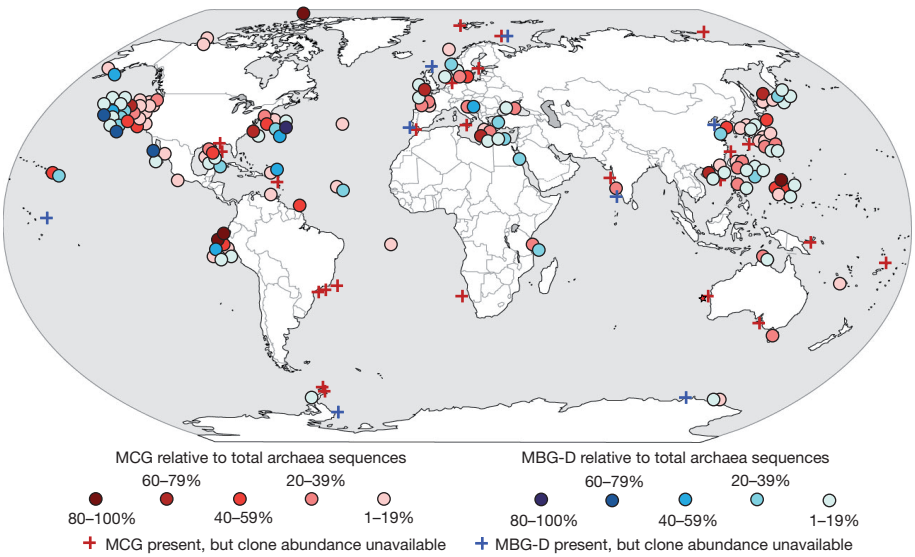

Figure 1 | Global marine occurrence of miscellaneous crenarchaeotal group (MCG) and marine benthic group D (MBG-D). Relative abundance of 16S rRNA gene sequences in clone libraries from marine sediments for MCG (red) and MBG-D (blue). For some (crosses), sequence abundance information was unavailable.

[1]Center for Geomicrobiology, Department of Bioscience, Aarhus University, Aarhus 8000, Denmark. [2]University of Tennessee, Knoxville, Tennessee 37996, USA. [3]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine 04544, USA. [4]Ribocon GmbH, Bremen 28359, Germany. [5]Max Planck Institute for Marine Microbiology, Bremen 28359, Germany.
*These authors contributed equally to this work.

©2013 Macmillan Publishers Limited. All rights reserved

sediment depth of 10 cm, sorted individual cells with fluorescence-activated cell sorting (FACS), amplified their genomic DNA with multiple displacement amplification (MDA), and screened for taxonomic identification via 16S rRNA gene sequences (Methods and Supplementary Figs 3 and 4). We selected a genome amplified from a single MCG cell (MCG_E09, which has NCBI name Thaumarchaeota archaeon SCGC AB-539-E09) and three genomes from single MBG-D cells (MBG-D_N05, MBG-D_F20 and MBG-D_C06, which have NCBI names Thermoplasmatales archaeon SCGC AB-539-N05, SCGC AB-539-F20 and SCGC AB-539-C06 in the NCBI database) for whole-genome sequencing (Supplementary Tables 4–6). Total *de novo* assembly sizes for the single amplified genomes (SAGs) ranged from 0.593 to 1.037 megabases (Mb) with 73–172 contigs. Quality control showed no evidence of contamination (Methods and Supplementary Fig. 5). Estimated genome coverage was 32–70% (Table 1), due to uneven or biased genomic amplification during MDA[11].

Phylogenetic analyses based on a concatenation of genes conserved in single copies in all archaea in the Integrated Microbial Genomes (IMG) database (Supplementary Fig. 6 and Supplementary Table 7) placed the MCG cell on a deep branch within the phyla Thaumarchaeota[3] and Aigarchaeota[4], and the MBG-D cells basal to or just inside group MG-II in the order Thermoplasmatales in the phylum Euryarchaeota (Fig. 2 and Supplementary Fig. 7). The MCG-E09 placement agrees with phylogenies based on single taxonomic marker genes that show that MCG are distinct from the Crenarchaeota[13]. However, the three MBG-D SAGs are not monophyletic relative to MG-II, the only other uncultured group of Thermoplasmatales with genomic information. This suggests either that partial complements of archaeal conserved genes cannot resolve fine-scale phylogenies or that the evolutionary history of the MBG-D is more complex than that derived from 16S rRNA genes. The 16S rRNA gene sequences from the single cells are >98% similar to other environmental sequences, making the cells representatives of these uncultured groups (Supplementary Fig. 8).

All four single cells contained predicted extracellular cysteine peptidases found in anaerobic protein-degrading bacteria: clostripain (Merops family C11, 1 copy in MCG_E09 and 2 copies in MBG-D_N05), gingipain (Merops family C25, 4, 12 and 6 copies in MBG-D_N05, MBG-D_F20 and MBG-D_C06, respectively), papain (Merops family C1A, 2 and 1 copies in MBG-D_N05 and MBG-D_F20, respectively) and pyroglutamyl peptidase (Merops family C15, 1 copy in MBG-D_F20) (Fig. 3 and Supplementary Tables 8 and 9)[14]. Clostripain and gingipain are secreted endopeptidases that are specific for arginine (or sometimes lysine, for gingipain) in the primary amino acid position of the substrate[15]. Papain and pyroglutamyl peptidase are often cytosolic or lysosomal; the first cleaves a wide variety of substrates and the second removes a pyroglutamate residue from the amino terminus of a peptide[15]. Each SAG gene homologue encodes a complete active site and signal sequences for extracellular transport; some may be coexpressed as they cluster on the genome (Supplementary Table 9). The clostripain from MCG_E09 is closely related to that of Clostridia spp. (Supplementary Fig. 9), and contains binding sites for the cofactor Ca^{2+} (Fig. 3d)[16]. In MBG-D_N05, clostripain is closely related to that of *Aciduliprofundum boonei*, a hyperthermophilic protein-degrading member of the Thermoplasmatales[17], and is adjacent to two copies of gingipain which have the domain architecture of gingipain in *A. boonei*[18] (Fig. 3c and Supplementary Fig. 9). However, 15 of the 16 gingipain copies from all three MBG-D SAGs are monophyletic and distinct from other groups (Supplementary Fig. 9). Cysteine peptidase activity requires chemically reducing, moderate-pH environments with high calcium concentrations[19], which are conditions commonly found in marine sediments. MCG_E09 has two copies of M19, one of the few peptidases known to target D-amino acids. The D enantiomer is highly abundant in the peptidoglycan of bacterial cell walls, which comprise the most persistent sedimentary detrital matter[20]. MCG_E09 may therefore be specially adapted to degrade these recalcitrant components of cell walls. A comparison with all 4,888 genomes in the IMG database[21] shows that only mesophilic or moderately thermophilic protein-degrading bacteria share all of

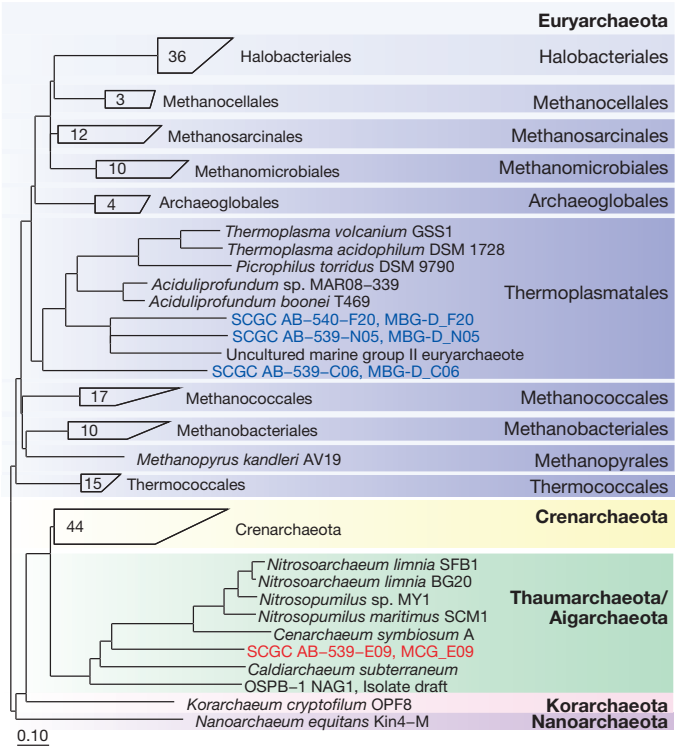

Figure 2 | Evolutionary placement of SAGs. Consensus of maximum likelihood (RAxML) trees of concatenated core archaeal conserved single-copy genes (individual trees shown in Supplementary Fig. 7). Phyla (bold) and orders are labelled. Numbers of genomes in collapsed clades are written on the boxes.

Table 1 | Assembly statistics for each SAG

Statistic	SCGC AB-539-C06	SCGC AB-539-E09	SCGC AB-539-N05	SCGC AB-540-F20
Total generated sequence data (Mb)	491	1,437	2,666	440
Assembly size (Mb)*	0.593	0.627	0.801	1.037
Number of contigs*	104	73	99	172
Maximum contig length (kb)*	48.2	59.3	73.3	77.0
N50 value (kb)*	10.4	27.3	27.1	12.9
Number of predicted genes*	792	787	879	1,272
Detected tRNAs	9	7	22	13
Predicted genome size (tRNA-based)† (Mb)	3.03	4.12	1.67	3.67
Achieved genome coverage (tRNA-based)† (%)	19.6	15.2	48.0	28.3
Detected CSCG	21	18	31	14
Predicted genome size (CSCG-based)‡ (Mb)	1.24	1.53	1.14	3.26
Achieved genome coverage (CSCG-based)‡ (%)	47.8	41.0	70.3	31.8

* Only contigs longer than 1,000 base pairs were considered.
† Based on expected average number of 46 transfer RNAs per genome (Supplementary Fig. 6).
‡ Based on 44 determined CSCGs (conserved single-copy genes) common to all archaeal genomes (Supplementary Fig. 6).

©2013 Macmillan Publishers Limited. All rights reserved

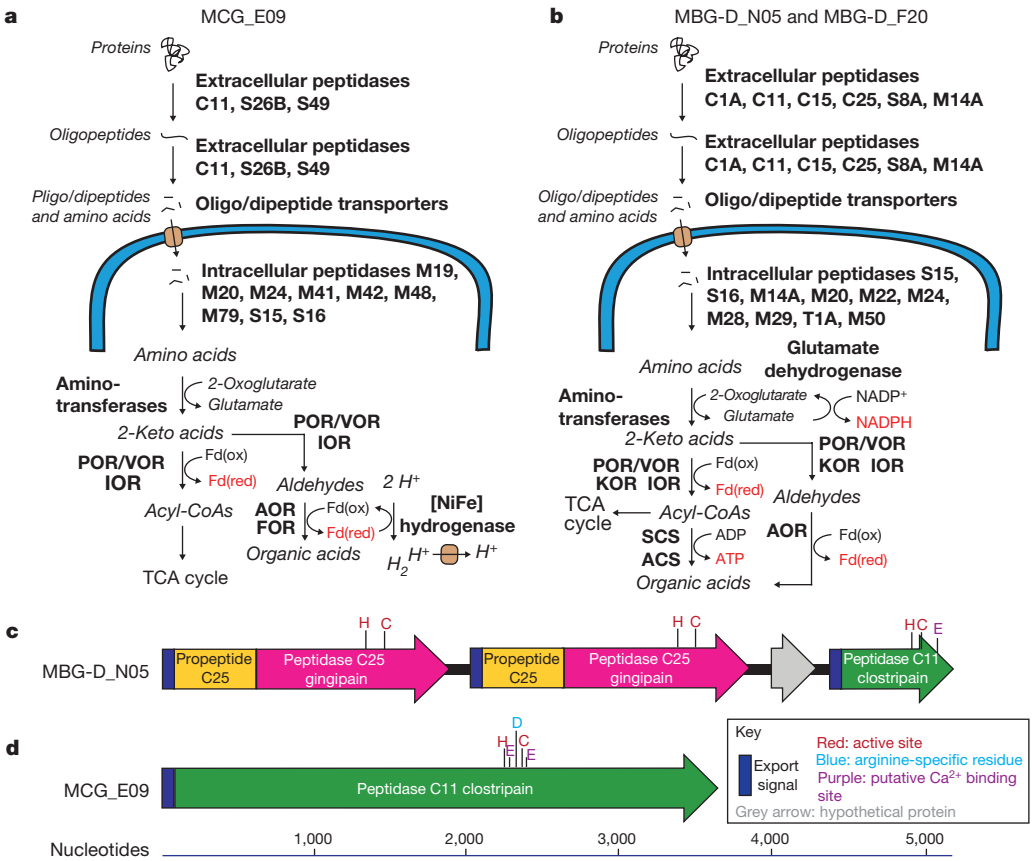

Figure 3 | Proposed protein degradation pathway for MCG_E09 (a) and MBG-D_N05 and MBG-D_F20 (b), and gene architecture for selected extracellular peptidases (c, d). Substrates and products are in black italic font, energetic molecules are red, enzymes are in black bold font, and blue lines indicate the cell membrane. ACS, acetyl-CoA synthetase; SCS, succinyl-CoA synthetase. Other acronyms are defined in the text. **c**, **d**, Gene architecture for gingipain and clostripain in MBG-D_N05 (**c**), and clostripain in MCG_E09 (**d**). MBG-D_C06 had a partial representation of the pathways present in **b**.

the MCG and MBG-D peptidases (Supplementary Table 10). Thus, the MCG and MBG-D single cells are probably capable of degrading the detrital proteins that are present in the Aarhus Bay sediment[8].

All four single cells contained di- and tripeptide transporters as well as genomic pathways for the intracellular breakdown of amino acids. These include aminotransferases, ATP-yielding acetyl-CoA synthetase (in MBG-D), as well as ferredoxin-reducing oxidoreductases specific for aldehydes (AOR, absent in MBG-D_C06), formaldehyde (FOR, only in MCG_E09) and pyruvate/2-ketoisovalerate (POR/VOR, in MCG_E09 and MBG-D_N05), which are intermediates in the breakdown of non-aromatic amino acids in hyperthermophilic archaea[22]. MCG_E09 and MBG-D_F20 also have indolepyruvate ferredoxin oxidoreductase (IOR), which targets intermediates of aromatic amino acid breakdown[22]. These oxidoreductases are highly oxygen-labile and use ferredoxin, which has the lowest redox potential of all known electron carriers (−500 mV)[23]. Tungsten (the AOR and FOR cofactor) has the lowest redox potential of any biologically relevant metal ligand, is plentiful in the Earth's crust (but not in sea water), and confers slower kinetics than observed for the more common molybdenum-containing oxidoreductases[24]. Most genomes in the IMG database with Blast hits ($E < 10^{-5}$) to all the ferredoxin-dependent oxidoreductases present in MCG_E09 and MBG-D_N05 (the other two SAGs had fewer oxidoreductases) were hyperthermophilic protein-degrading archaea, with hyperthermophilic protein-degrading bacteria making up most of the rest (Supplementary Table 11). The presence of oxidoreductases normally associated with hyperthermophiles in MCG and MBG-D, which inhabit permanently cold sediments (2–16 °C seasonally[25]), may be ancestral or confer enhanced molecular stability in this reducing, energy-limited environment.

Reduced ferredoxin produced during protein degradation may be used to create a proton motive force in MCG_E09 by its membrane-bound [NiFe]-hydrogenase, analogously to the mechanism present in *Pyrococcus* sp.[26]. (Supplementary Table 8 and Fig. 3a). MBG-D_N05 and MBG-D_F20 contain heterodisulphide reductase subunits A, B and C (*hdrABC*), methyl-viologen-reducing hydrogenase subunits A, G and D (*mvhAGD*), and N5-methyltetrahydromethanopterin-coenzyme M methyltransferase subunits A and H (*mtrAH*) (Supplementary Table 8). In some methanogens, these enzymes couple hydrogenotrophic methane production to the creation of a sodium motive force that drives ATP formation[27], but the enzymes are also found in non-methanogenic microorganisms. MBG-D_N05 may therefore have a sodium-based energy conservation mechanism.

We observed high extracellular peptidase activity consistent with gingipain and clostripain at 600-cm depth in Aarhus Bay sediments. Gingipain substrates indicated V_{max} (velocity of enzyme-catalysed reaction at saturating substrate concentrations) = 9.9 ± 1.6 μmol 7-amino-4-methylcoumarin (AMC) h^{-1} g^{-1} sediment and K_m (Michaelis constant) = 51 ± 30 μM substrate. Clostripain substrates indicated V_{max} = 15 ± 5.8 μmol AMC h^{-1} g^{-1} sediment, K_m = 188 ± 153 μM substrate (Supplementary Fig. 10). Leucyl aminopeptidase, which to our knowledge is the only other peptidase substrate that has been assayed in marine sediments[28], had much lower potential activity (Supplementary Fig. 10). Archaeal peptidases seem to be numerous in marine sediments, as metagenomes from two geographically disparate marine sediments (California, Gold ID Gm00260, and Alaska, Gold ID Gm00257) are replete with homologues of all extracellular peptidases found in the single cells (up to 2.3 peptidase homologues per genome; $E \leq 10^{-10}$; Supplementary Table 12). This type of extracellular protein degradation within archaea therefore seems to be active, abundant and geographically widespread.

Archaea degrade detrital proteins in extreme environments[17]. The partial genomes obtained for MCG and MBG-D suggest that archaea have a similar function in cold anoxic marine sediments, which comprise the largest organic carbon sink on Earth[29]. Each partial genome contains genes for complete degradation pathways of extracellular proteins, including enzymes whose homologues occur together only in cultured protein-degrading prokaryotes. MCG may represent a new phylum in the Archaea and MBG-D may represent a new order in the

©2013 Macmillan Publishers Limited. All rights reserved

Euryarchaeota because (1) their evolutionarily conserved genes place them basal to established phyla and orders; (2) their environmental distributions differ greatly from their nearest neighbours, Thaumarchaeota, Aigarchaeota and Thermoplasmatales, which are primarily found in oxic and/or hot environments[4,30]; and (3) MCG and MBG-D single-cell genomes seem to be capable of exogenous protein degradation in cold anoxic environments, a process that has never been observed in other archaea. More single-cell genomes or cultures from these and other uncultured groups may further establish them as taxonomic levels that should be given more formal and accurate names than MCG and MBG-D. The ubiquity and frequent dominance of these archaeal groups, as well as the high abundance and potential activity of the type of peptidases that they encode, indicate the importance of these archaea in protein remineralization in marine sediments. However, the broad 16S rRNA gene diversity within these archaeal groups[9] indicates that their impacts on marine biogeochemical cycles probably extend beyond their involvement in detrital protein degradation.

METHODS SUMMARY

A sediment core was collected on 22 March 2011, in a shallow gas area at Aarhus Bay, Denmark (56° 9′ 35.889 N, 10° 28′ 7.893 E), water depth 16.3 m and 2.5 °C (Supplementary Fig. 1). Cells were extracted from 10 cm sediment depth by ultrasonic treatment followed by density gradient centrifugation. Single-cell sorting, whole-genome amplification, and PCR screening of single cells were performed at the Bigelow Laboratory Single Cell Genomics Center (SCGC; http://www.bigelow.org/scgc) by FACS using the SYTO-9 DNA stain. The sorted cells were lysed using five cycles of freeze–thaw, followed by further lysis and DNA denaturing with cold KOH. Genomic DNA from the lysed cells was amplified using MDA, resulting in SAGs. MDA products were screened by quantitative PCR with primer sets targeting 16S rRNA genes. Sequencing of the SAGs was performed using the 454-GS-FLX Titanium, the Illumina HiSeq 2000, and the Ion Torrent PGM platforms. Amplicon sequencing (Supplementary Fig. 2) as well as quantitative PCR was performed on DNA from station M1, a sampling site in close proximity to the SAG sampling site (Supplementary Fig. 1). Two different archaea-targeted primer pairs were used for PCR amplification of 16S rRNA gene fragments and subsequent 454 pyrosequencing. The quantitative PCR was performed by using published primer sets for archaea, bacteria and the MCG group as well as newly designed MBG-D primers (Supplementary Table 2). Extracellular peptidase activities were assayed with fluorogenic substrates. L-leucine-7-amino-5-methylcoumarin (Leu-AMC), Z-Phe-Arg-AMC and Z-Phe-Val-Arg-AMC were used as substrates for leucyl aminopeptidase, gingipain and clostripain, respectively. Autoclaved sediment served as a killed control for each substrate and concentration.

Full Methods and any associated references are available in the online version of the paper.

Received 17 October 2012; accepted 20 February 2013.
Published online 27 March 2013.

1. Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C. & D'Hondt, S. Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc. Natl Acad. Sci. USA* **109,** 16213–16216 (2012).
2. Schippers, A., Köweker, G., Höft, C. & Teichert, B. M. A. Quantification of microbial communities in forearc sediment basins off Sumatra. *Geomicrobiol. J.* **27,** 170–182 (2010).
3. Brochier-Armanet, C., Boussau, B., Gribaldo, S. & Forterre, P. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Rev. Microbiol.* **6,** 245–252 (2008).
4. Nunoura, T. *et al.* Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res.* **39,** 3204–3223 (2011).
5. Jarrell, K. F. *et al.* Major players on the microbial stage: why archaea are important. *Microbiology* **157,** 919–936 (2011).
6. Biddle, J. F. *et al.* Heterotrophic archaea dominate sedimentary subsurface ecosystems off Peru. *Proc. Natl Acad. Sci. USA* **103,** 3846–3851 (2006).
7. Wakeham, S. G., Lee, C., Hedges, J. I., Hernes, P. J. & Peterson, M. J. Molecular indicators of diagenetic status in marine organic matter. *Geochim. Cosmochim. Acta* **61,** 5363–5369 (1997).
8. Pedersen, A.-G. U., Thomsen, T. R., Lomstein, B. A. & Jørgensen, N. O. G. Bacterial influence on amino acid enantiomerization in a coastal marine sediment. *Limnol. Oceanogr.* **46,** 1358–1369 (2001).
9. Kubo, K. *et al.* Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J.* **6,** 1949–1965 (2012).
10. Holmkvist, L., Ferdelman, T. G. & Jørgensen, B. B. A cryptic sulfur cycle driven by iron in the methane zone of marine sediment (Aarhus Bay, Denmark). *Geochim. Cosmochim. Acta* **75,** 3581–3599 (2011).
11. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71,** 3342–3347 (2005).
12. Stepanauskas, R. & Sieracki, M. E. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl Acad. Sci. USA* **104,** 9052–9057 (2007).
13. Li, P. *et al.* Genetic structure of three fosmid-fragments encoding 16S rRNA genes of the Miscellaneous Crenarchaeotic Group (MCG): implications for physiology and evolution of marine sedimentary archaea. *Environ. Microbiol.* **14,** 467–479 (2012).
14. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **40,** D343–D350 (2012).
15. Barrett, A. J. & Rawlings, N. D. Evolutionary lines of cysteine peptidases. *Biol. Chem.* **382,** 727–733 (2001).
16. Labrou, N. E. & Rigden, D. J. The structure–function relationship in the clostripain family of peptidases. *Eur. J. Biochem.* **271,** 983–992 (2004).
17. Reysenbach, A. *et al.* A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal vents. *Nature* **442,** 444–447 (2006).
18. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40,** D290–D301 (2012).
19. Kembhavi, A. A., Buttle, D. J. & Barrett, A. J. Clostripain: characterization of the active site. *FEBS Lett.* **283,** 277–280 (1991).
20. Lomstein, B. A., Langerhuus, A. T., Hondt, S. D., Jørgensen, B. B. & Spivack, A. J. Endospore abundance, microbial growth and necromass turnover in deep sub-seafloor sediment. *Nature* **484,** 101–104 (2012).
21. Markowitz, V. M. *et al.* IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* **40,** D123–D129 (2012).
22. Schut, G. J., Menon, A. L. & Adams, M. W. W. 2-keto acid oxidoreductases from *Pyrococcus furiosus* and *Thermococcus litoralis*. *Methods Enzymol.* **331,** 144–158 (2001).
23. Hall, D. O., Cammack, R. & Rao, K. K. Role for ferredoxins in the origin of life and biological evolution. *Nature* **233,** 136–138 (1971).
24. Johnson, M. K., Rees, D. C. & Adams, M. W. W. Tungstoenzymes. *Chem. Rev.* **96,** 2817–2840 (1996).
25. Dale, A. W. *et al.* Seasonal dynamics of the depth and rate of anaerobic oxidation of methane in Aarhus Bay (Denmark) sediments. *J. Mar. Res.* **66,** 127–155 (2008).
26. Sapra, R., Bagramyan, K. & Adams, M. A simple energy-conserving system: Proton reduction copuled to proton translocation. *Proc. Natl Acad. Sci. USA* **100,** 7545–7550 (2003).
27. Thauer, R., Kaster, A.-K., Seedorf, H., Buckel, W. & Hedderich, R. Methanogenic archaea: ecologically relevant differences in energy conservation. *Nature Rev. Microbiol.* **6,** 579–591 (2008).
28. Coolen, M. J. L. & Overmann, J. Functional exoenzymes as indicators of metabolically active bacteria in 124,000-year-old sapropel layers of the Eastern Mediterranean Sea. *Appl. Environ. Microbiol.* **66,** 2589–2598 (2000).
29. Hedges, J. I. & Keil, R. G. Sedimentary organic matter preservation: an assessment and speculative synthesis. *Mar. Chem.* **49,** 81–115 (1995).
30. Brochier-Armanet, C., Gribaldo, S. & Forterre, P. Spotlight on the Thaumarchaeota. *ISME J.* **6,** 227–230 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors thank the captain and crew of the R/V *Tyra* for sampling; T. B. Søgaard, A. Stentebjerg and B. Poulsen for technical work; F. Löffler for laboratory space; and D. Kirchman and S. Hallam for sharing their unpublished metagenomic data sets. This work was funded by the Danish National Research Foundation, the German Max Planck Society, NSF Center for Dark Energy Biosphere Investigations NSF-157595 (K.G.L.), The Danish Council for Independent Research–Natural Sciences (D.G.P.), the Villum Kann Rasmussen Foundation, an EU Marie Curie fellowship (M.A.L.), the German Research Foundation (S.L.) and the USA National Science Foundation awards EF-826924, OCE-821374 and OCE-1019242 (R.S.).

Author Contributions K.G.L., L.S., D.G.P., K.U.K., R.S., A.S. and B.B.J. worked together to design experiment and develop the method for single cell sorting from sediments. K.G.L. wrote the main paper and developed the protein degradation hypothesis. L.S. wrote the Supplementary Information, designed and performed bioinformatic analyses. K.G.L. and L.S. performed phylogenetic tests. K.G.L. and D.G.P. reconstructed metabolic pathways with SAG genes. R.S. performed cell sorting and amplification. S.K., S.L., D.G.P. and L.S. developed protocols for cell separation from sediments. K.U.K. performed and analysed 16S rRNA gene amplicon sequencing; M.A.L., L.S. and K.G.L. performed quantitative PCR; A.D.S. performed enzyme activity measurements; and M.R. gave bioinformatic support and added quality control tests. A.S. and B.B.J. obtained the major funding for this work. All co-authors commented on and provided substantial edits to the manuscript.

Author Information This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank as Thaumarchaeota archaeon SCGC AB-539-E09 (accession number ALXK00000000), Thermoplasmatales archaeon SCGC AB-539-C06 (AOSH00000000), Thermoplasmatales archaeon SCGC AB-539-N05 (ALXL00000000) and Thermoplasmatales archaeon SCGC AB-540-F20 (AOSI00000000). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.G.L. (klloyd@utk.edu).

©2013 Macmillan Publishers Limited. All rights reserved

METHODS

Sequencing of SAGs, sequence assembly and annotation. A sediment core was collected on 22 March 2011, in a shallow gas area at Aarhus Bay, Denmark (56° 9′ 35.889 N, 10° 28′ 7.893 E), water depth 16.3 m and 2.5 °C (Supplementary Fig. 1). Cells were extracted from 10 cm sediment depth by ultrasonic treatment followed by density gradient centrifugation. Single-cell sorting, whole-genome amplification, and PCR screening of single cells were performed at the Bigelow Laboratory Single Cell Genomics Center (SCGC; http://www.bigelow.org/scgc) by FACS using the SYTO-9 DNA stain. The sorted cells were lysed using five cycles of freeze–thaw, followed by further lysis and DNA denaturing with cold KOH. Genomic DNA from the lysed cells was amplified using MDA, resulting in SAGs. MDA products were screened by quantitative PCR with primer sets targeting 16S rRNA genes. To obtain sufficient quantity of DNA for sequencing, some SAGs were re-amplified by a second round of MDA at the SCGC. Sequencing with the 454-GS-FLX Titanium and the Illumina HiSeq 2000 was performed at GATC Biotech (Germany) with re-amplified SAGs as template. Sequencing with the Ion Torrent PGM platform was performed at Aarhus University using the original (non-re-amplified) SAGs as template. The sequence data for the individual SAGs are summarized in Supplementary Table 4.

Reads from the 454-GS-FLX Titanium and the Ion Torrent PGM platforms were quality trimmed using the prinseq-lite.pl script[31]. The 454 and Ion Torrent reads were de-replicated using cd-hit-454[32] and subsequently assembled using gsAssembler version 2.6 (Roche). In parallel, we used the SPAdes assembler version 2.2.1[33] to assemble the Illumina reads of SCGC AB-539-E09 and SCGC AB-539-N05, as well as the Ion Torrent reads of SCGC AB-539-C06 and SCGC AB-540-F20. The gsAssembler and SPAdes assemblies of each SAG were combined using Sequencher version 5.0.1 (Genecodes). The SAG assemblies were auto-annotated using the Joint Genome Institute (JGI) IMG-ER pipeline[21]. Annotations were manually curated using GenDB[34] supplemented by JCoast[35]. All peptidases were aligned in ARB[36] against the Merops alignment[14] (Supplementary Fig. 9).

Estimation of genome size and purity control of SAGs. The genome sizes of the SAGs were estimated using a conserved single-copy gene (CSCG) analysis similar to ref. 37. CSCGs present in all archaea were identified using the JGI IMG site[38]. The CSCG-based approach was supplemented by a tRNA-based approach[39] where SAG tRNA numbers were compared to the numbers of tRNAs of complete archaeal genomes.

The purity of the SAGs was tested by PCR-screening for archaeal and bacterial 16S rRNA genes using multiple primer sets. Additionally, all predicted ORFs were checked with Blastp[40] against NCBI-nr for amino acid identity observations of over 96% to detect common contaminants. Finally, contigs showing a k-mer pattern divergent from the rest of the SAG sequences[41] were manually checked for a possible contamination by examining their closest Blastp hits.

Phylogenetic reconstruction using conserved single-copy genes and 16S rRNA genes. Similar to the studies performed by refs 42 and 43, we used archaeal CSCGs for inferring the phylogenetic affiliation of the inspected SAGs. The identified CSCG amino acid sequences of completed archaeal genomes and our SAGs were extracted from the JGI IMG-ER site. The sequences were individually aligned using MAFFT version 6.864[44]. All alignments were manually curated using ARB[36]. The alignments were concatenated using the Perl script catfasta2phyml.pl (http://www.abc.se/~nylander/catfasta2phyml/). Maximum likelihood bootstrap trees were calculated using RAxML-HPC2 (RAxML version 7.2.7[45]) as provided by the CIPRES cluster (http://www.phylo.org/46). The 16S rRNA gene tree was created using RAxML-HPC[45] at the CIPRES cluster[46] (Supplementary Fig. 8).

Aarhus Bay 16S rRNA amplicon sequencing and quantitative PCR. A sediment core was taken at station M1, (56° 07.07′ N, 10° 20.80′ E; see Supplementary Fig. 1), a well-characterized site near the site from which SAGs were derived in Aarhus Bay[47]. DNA was extracted[48] from five depths. Two different archaea-targeted primer pairs[49–52] were used for PCR amplification of 16S rRNA gene fragments. The resultant PCR products were sequenced on a 454-GS-FLX Titanium machine as previously described[53]. Sequence analysis was performed in Mothur[54]. Sequence reads were classified according to Silva taxonomy (release 102 (ref. 55)) and new MCG subgroups[9] (Supplementary Fig. 2).

The DNA for quantitative PCR was extracted from the same sediment used for 16S rRNA amplicon sequencing, station M1, using a novel, chemical lysis-based method (M.A.L., manuscript in preparation). The quantitative PCR was performed with primers listed in Supplementary Table 2 (Supplementary Fig. 2). Primers for MBG-D (Supplementary Tables 2 and 3) were designed using PRIMROSE 2.17 (ref. 56) and 839 MBG-D sequences included in the SILVA SSURef database release 106 (ref. 55).

Enzymatic activity assays. Extracellular peptidase activities were assayed using fluorogenic substrates according to a protocol loosely based on that of ref. 57 (Supplementary Fig. 10). L-leucine-7-amino-4-methylcoumarin (Leu-AMC, Sigma-Aldrich) was used to assay leucyl aminopeptidase activity, Z-Phe-Arg-AMC (Sigma-Aldrich) was used to assay gingipain[58] and Z-Phe-Val-Arg-AMC (Bachem) was used to assay clostripain[59]. Sediments were collected by gravity core from Aarhus Bay site M1. Sediments from 600 to 630 cm below sea floor were homogenized, placed in 5-ml serum vials (~0.5 g each), mixed with 4.0 ml anoxic artificial sea water (salinity 30 practical salinity units, pH 7.8) and the precise mass of wet sediment was recorded. Serum vials were then capped, purged with N_2 and vortexed to mix completely. For each substrate and concentration, autoclaved sediment was used as a killed control. Immediately after vortexing, 2 ml slurry was removed to measure initial fluorescence.

To measure fluorescence, each sediment subsample was transferred into a microcentrifuge vial and centrifuged for 10 min at 9,300*g*. One millilitre of supernatant was removed and transferred to a methacrylate 1.5-ml fluorescence cuvette. Fluorescence was measured with a Promega QuantiFluor ST fluorimeter. To calibrate fluorescence values, 7-amino-4-methylcoumarin (AMC) was added directly to sediment slurries, mixed thoroughly, and then processed in the same way as samples for enzyme assays. Enzyme activities are reported as rates of AMC liberation per hour per gram wet sediment.

Sediments were incubated in the dark in a shaking incubator for approximately 8 h (precise time was recorded). After incubation, fluorescence was measured as described above. Enzyme activity was calculated from the change in fluorescence for each sample divided by the incubation time. Kinetic parameters were calculated using the R statistical package[60] by a nonlinear least-squares fit to the activity data (Supplementary Fig. 10).

31. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).
32. Niu, B., Fu, L., Sun, S. & Li, W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11,** 187 (2010).
33. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
34. Meyer, F. *et al.* GenDB-an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* **31,** 2187–2195 (2003).
35. Richter, M. *et al.* JCoast—A biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. *BMC Bioinformatics* **9,** 177 (2008).
36. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Database* **32,** 1363–1371 (2004).
37. Woyke, T. *et al.* Assembling the marine metagenome, one cell at a time. *PLoS ONE* **4,** e5299 (2009).
38. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids Res.* **34,** D344–D348 (2006).
39. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 0955–0964 (1997).
40. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
41. Swan, B. K. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333,** 1296–1300 (2011).
42. Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **19,** 631–639 (2002).
43. Brochier, C., Forterre, P. & Gribaldo, S. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol. Biol.* **5,** 36 (2005).
44. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9,** 286–298 (2008).
45. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* **57,** 758–771 (2008).
46. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computing Environments Workshop 1–8 (2010).
47. Jørgensen, B. B. Case study—Århus Bay. *Eutrophication in Coastal Marine Systems* 137–154 (American Geophysical Union, 1996).
48. Kjeldsen, K. U. *et al.* Diversity of sulfate-reducing bacteria from an extreme hypersaline sediment, Great Salt Lake (Utah). *FEMS Microbiol. Ecol.* **60,** 287–298 (2007).
49. DeLong, E. F. Archaea in coastal marine environments. *Proc. Natl Acad. Sci. USA* **89,** 5685–5689 (1992).
50. Stahl, D. A. & Amann, R. *Development and Application of Nucleic Acid Probes* (Wiley, 1991).
51. Takai, K. E. N. & Horikoshi, K. Rapid detection and quantification of members of the archaeal community by quantitative PCR using fluorogenic probes. *Appl. Environ. Microbiol.* **66,** 5066–5072 (2000).
52. Teske, A. & Sørensen, K. B. Uncultured archaea in deep marine subsurface sediments: have we caught them all? *ISME J.* **2,** 3–18 (2008).
53. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE* **5,** e9085 (2010).
54. Schloss, P. D. *et al.* Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Society* **75,** 7537–7541 (2009).
55. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).

©2013 Macmillan Publishers Limited. All rights reserved

56. Ashelford, K. E., Weightman, A. J. & Fry, J. C. PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Res.* **30,** 3481–3489 (2002).
57. King, G. M. Characterization of β-glucosidase activity in intertidal marine sediments. *Appl. Environ. Microbiol.* **51,** 373–380 (1986).
58. Nakayama, K., Kadowaki, T., Okamoto, K. & Yamamoto, K. Construction and characterization of arginine-specific cysteine proteinase (Arg-gingipain)-deficient mutants of *Porphyromonas gingivalis*: Evidence for significant contribution of Arg-gingipain to virulence. *J. Biol. Chem.* **270,** 23619–23626 (1995).
59. Rauber, P., Walker, B., Stone, S. & Shaw, E. Synthesis of lysine-containing sulphonium salts and their properties as proteinase inhibitors. *Biochem. J.* **250,** 871 (1988).
60. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing http://www.rproject.org (2012).

©2013 Macmillan Publishers Limited. All rights reserved