






ARTICLE

DOI: 10.1038/s41467-017-01376-9

OPEN

# Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients

Alyse K. Hawley <sup>1</sup>, Masaru K. Nobu<sup>2,3</sup>, Jody J. Wright<sup>1</sup>, W. Evan Durno<sup>4</sup>, Connor Morgan-Lang<sup>4</sup>, Brent Sage<sup>4</sup>, Patrick Schwientek<sup>5</sup>, Brandon K. Swan<sup>6,11</sup>, Christian Rinke <sup>7</sup>, Monica Torres-Beltrán<sup>1</sup>, Keith Mewis<sup>8</sup>, Wen-Tso Liu<sup>2</sup>, Ramunas Stepanauskas <sup>6</sup>, Tanja Woyke <sup>5</sup> & Steven J. Hallam <sup>1,4,9,10</sup>

Microbial communities drive biogeochemical cycles through networks of metabolite exchange that are structured along energetic gradients. As energy yields become limiting, these networks favor co-metabolic interactions to maximize energy disequilibria. Here we apply single-cell genomics, metagenomics, and metatranscriptomics to study bacterial populations of the abundant “microbial dark matter” phylum Marinimicrobia along defined energy gradients. We show that evolutionary diversification of major Marinimicrobia clades appears to be closely related to energy yields, with increased co-metabolic interactions in more deeply branching clades. Several of these clades appear to participate in the biogeochemical cycling of sulfur and nitrogen, filling previously unassigned niches in the ocean. Notably, two Marinimicrobia clades, occupying different energetic niches, express nitrous oxide reductase, potentially acting as a global sink for the greenhouse gas nitrous oxide.

<sup>1</sup> Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada. <sup>2</sup> Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Avenue, Urbana, IL 61801, USA. <sup>3</sup> Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Central 6, Higashi, Tsukuba, Ibaraki 305-8566, Japan. <sup>4</sup> Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada. <sup>5</sup> Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA. <sup>6</sup> Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. <sup>7</sup> Australian Centre for Ecogenomics, University of Queensland, St Lucia, Brisbane, 4072 QLD, Australia. <sup>8</sup> Genome Science and Technology Graduate Program, University of British Columbia, Vancouver, BC, Canada. <sup>9</sup> ECOSCOPE Training Program, University of British Columbia, Vancouver, BC, Canada. <sup>10</sup> Peter Wall Institute for Advanced Studies, University of British Columbia, Vancouver, BC V6T 1Z2, Canada. <sup>11</sup> Present address: National Biodefense Analysis and Countermeasures Center, Frederick, MD 21702, USA. Alyse K. Hawley and Masaru K. Nobu contributed equally to this work. Correspondence and requests for materials should be addressed to S.J.H. (email: [shallam@mail.ubc.ca](mailto:shallam@mail.ubc.ca))

The laws of thermodynamics apply to all aspects of Life, governing energy flow in both biotic and abiotic regimes. Nicholas Georgescu–Roegen was the first to directly apply the laws of thermodynamics to economic theory, bringing to the forefront the reality of limited natural resources on sustainable growth<sup>1</sup>. Robert Ayers used the term “eco-thermodynamics” to describe the application of thermodynamics and energy flow to economic models with the controversial conclusion that future economic growth necessitates the recycling of goods<sup>2</sup>. Within microbial ecology there is an emerging consensus that these same organizing principles structure microbial community interactions and growth with feedback on global nutrient and energy cycling<sup>3–6</sup>. Indeed, recycling in the common sense may be analogous to metabolite exchange or use of public goods<sup>7</sup>, as the goods from one production stream become available for growth of another. Microbial communities living near-thermodynamic limits where high potential electron acceptors are scarce tend to utilize differential modes of metabolic coupling including obligate syntrophic interactions, maximizing any chemical disequilibria to yield energy for growth<sup>8,9</sup>. Thus, the term eco-thermodynamics takes on new meaning in the context of microbial ecology where thermodynamic constraints directly shape the structure and activity of microbial interaction networks.

Eco-thermodynamic gradients are formed by the distribution of available electron donors and acceptors within the physical environment, creating metabolic niches that are occupied by diverse microbial partners playing recurring functional roles<sup>10,11</sup>. Marine oxygen minimum zones (OMZs) provide a vivid example of eco-thermodynamic gradients shaping differential modes of metabolic coupling at the intersection of carbon, nitrogen, and sulfur cycling in the ocean<sup>12,13</sup>. For example, OMZ microbial communities manifest a modular denitrification pathway that links reduced sulfur compounds to nitrogen loss and nitrous oxide (N<sub>2</sub>O) production<sup>12,14–16</sup>. While many of the most abundant interaction partners are known, recent modeling efforts point to a novel metabolic niche for the terminal step in the denitrification pathway (nitrous oxide reduction to dinitrogen gas) occupied by unidentified community members<sup>5</sup>. By defining the interaction networks coupling microbial processes along eco-thermodynamic gradients it becomes possible to more accurately model nutrient and energy flow at ecosystem scales.

Recent advances in sequencing technologies have opened a genomic window on uncultivated microbial diversity, illuminating the metabolic potential of numerous candidate divisions also known as microbial dark matter (MDM)<sup>17–20</sup>. Many MDM organisms occupy low-energy environments, where they appear to form obligate metabolic dependencies that could help explain resistance to traditional isolation methods. Marinimicrobia (formerly known as Marine Group A and SAR406) is an MDM phylum with no cultured representatives that is prevalent in the ocean. Marine Marinimicrobia have been previously implicated in sulfur cycling via a polysulfide reductase gene cluster<sup>21,22</sup>. In studies of a methanogenic bioreactor, Marinimicrobia have also been identified to rely on syntrophic interactions with metabolic partners to accomplish degradation of amino acids<sup>23</sup>. The global distribution of Marinimicrobia clades implicates a much wider diversity of both metabolic functions and partners than currently described. Here we use shotgun metagenomics, metatranscriptomics and single-cell genomics to investigate energy metabolism within the Marinimicrobia to reveal novel modes of metabolic coupling with important implications for nutrient and energy cycling in the ocean.

## Results

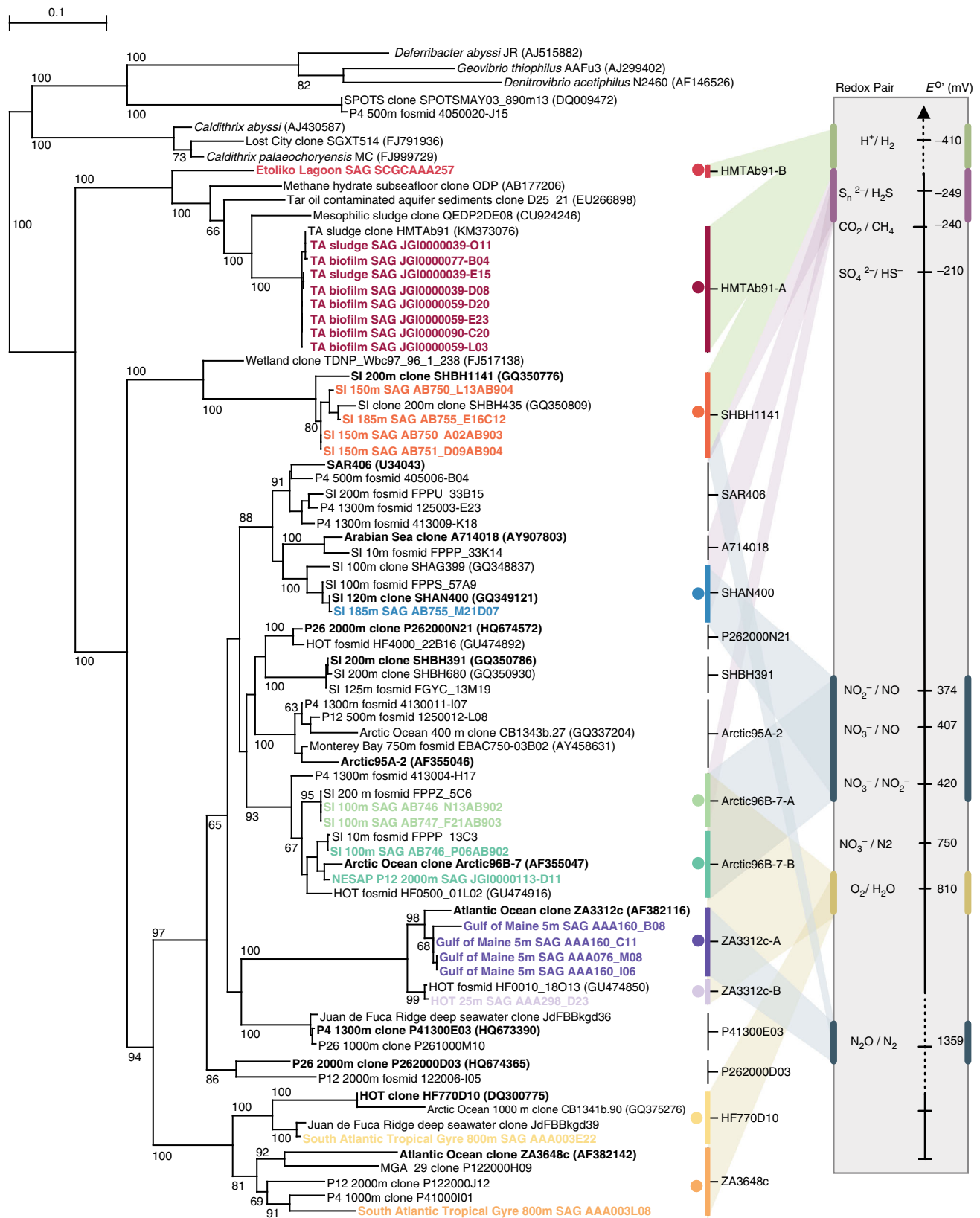
### Marinimicrobia single-cell amplified genomes and phylogeny.

A total of 25 Marinimicrobia single-cell amplified genomes

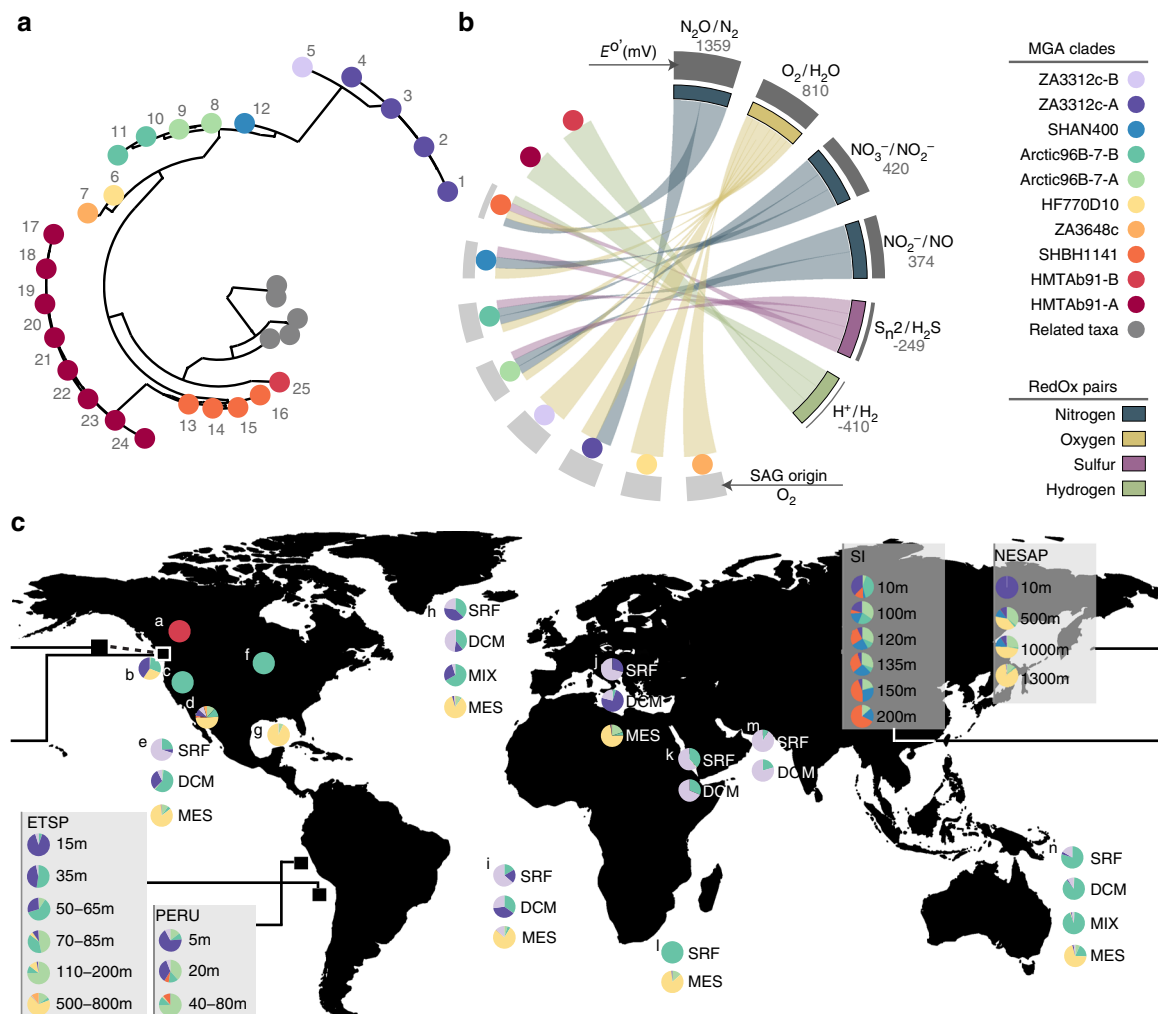
(SAGs) from sources along eco-thermodynamic gradients were identified globally by flow sorting, whole-genome amplification and sequencing (Supplementary Data 1). SAG de novo assemblies ranged in size from 0.39 to 2.01 million bases (Mb) with estimated genome completeness ranging from <10% to >90% (average 45%) (Supplementary Table 1). Most Marinimicrobia SAGs manifested streamlined genomes, with high coding base percentage (89.99–97.13%) and low cluster of orthologous group (COG) redundancy (1.08–1.16) (Supplementary Fig. 1). PhyloPhlAn analysis of conserved marker genes placed Marinimicrobia SAGs within the bacterial domain branching deeply from the closest cultured thermophilic representative *Caldithrix abyssi* (Supplementary Fig. 2). To determine phylogenetic diversity within the Marinimicrobia, we constructed a comprehensive SSU rRNA gene tree resolving 17 clades (Fig. 1). SAG sequences were affiliated with 10 clades spanning the entire breadth of the Marinimicrobia tree (Figs. 1 and 2a, b) providing a broad phylogenetic range with which to assess distribution patterns and energy metabolism within the phylum.

**Biogeography of Marinimicrobia clades.** Using this phylogenetic information, we determined the global biogeographic distribution of Marinimicrobia and specific SAG-affiliated clades along eco-thermodynamic gradients spanning oxic (>90 μmol O<sub>2</sub>), dysoxic (20–90 μmol O<sub>2</sub>), suboxic (1–20 μmol O<sub>2</sub>), anoxic (<1 μmol O<sub>2</sub>), sulfidic and methanogenic conditions. Estimates of Marinimicrobia total abundance and clade distribution were carried out by a robust survey of 594 globally sourced metagenomes (549 assembled Illumina data sets and 45 unassembled 454 data sets) across terrestrial and marine ecosystems, including Northeastern Subarctic Pacific (NESAP, *n* = 43), Saanich Inlet (SI, *n* = 90), Eastern Tropical South Pacific (ETSP, *n* = 6), Peruvian (*n* = 17), and Guaymas Basin (*n* = 2) OMZs; TARA Oceans (*n* = 243) and several other marine (*n* = 141) and terrestrial sites (*n* = 52), (Supplementary Data 2) totaling 127 Gigabases (Gb) of sequence information. To estimate total abundance, we used a sequence similarity recruitment with a cutoff of >70% nucleotide identity over >70% of the metagenomic contig. Globally we recovered 1.3 Gb of Marinimicrobia-affiliated sequence or 1.3 million genome equivalents (assuming 1 Mb average genome) representing ~1% of surveyed data. The recovery of Marinimicrobia-affiliated sequences was highest in coastal OMZs, increasing in relation to decreasing O<sub>2</sub> concentration (Supplementary Fig. 3A). Recovery was more variable in other marine locations and minimal in terrestrial locations. To more fully resolve this sequence information at the level of specific Marinimicrobia clades, we conducted a more stringent recruitment of >95% nucleotide identity across >200 bp intervals (Supplementary Data 4). On a global scale three clades constituted 75% of observed Marinimicrobia with the remaining seven clades making up the difference (Supplementary Fig. 3B). Consistent with previous results, predominantly marine sites were recruited with two hits from terrestrial locations. Sakinaw Lake, a meromictic lake with high methane concentrations<sup>19</sup>, was the only geographic location with recruitment to the HMTAb91 clade. Within marine systems, SAGs recruited sequences from cognate environments and conditions consistent with observed tree branching patterns (Fig. 2a–c; Supplementary Datas 3 and 4). Overall, trends indicated that specific clades inhabit particular energetic niches with potential for metabolic coupling within a given niche.

**Population genome bin construction.** To determine the energy metabolism of Marinimicrobia clades and overcome low genome completion of some SAGs, we leveraged extensive metagenomic



**Fig. 1** Maximum-likelihood small subunit rRNA gene tree and proposed energy metabolism for Marinimicrobia clades. Maximum likelihood phylogenetic tree of small subunit ribosomal rRNA (SSU rRNA) genes from all available studies. SSU rRNA genes from SAGs used in this study are in bold and colored to indicate their membership to population genome bins. Redox pairs are colored consistent with Fig. 1. Energy metabolism redox pairs for each clade explored in this publication are mapped to the electron tower on the right of the tree. The bar represents 1% estimated sequence divergence. Bootstrap values below 50% are not shown



**Fig. 2** Phylogeny and biogeography of Marinimicrobia single-cell-amplified genomes and clades. **a** Unrooted phylogenetic tree based on SSU rRNA genes showing the phylogenetic affiliation of Marinimicrobia SAGs. Each dot represents a SAG in Supplementary Table 1 with the corresponding number. The tree was inferred using maximum likelihood implemented in PhyML. **b** Circular plot indicating the terminal electron acceptors used and their respective  $E^0$  (mV) value by the different Marinimicrobia clades (left). **c** Global distribution of Marinimicrobia SAG-affiliated clades, as determined by metagenomic fragment recruitment using FAST (23) with 594 global metagenomes with a threshold of  $\geq 95\%$  nucleotide sequence identity and alignments  $\geq 200$  bp. Recruited contig lengths were normalized by the length of each SAG assembly in mega base pairs (Mbp) and to the size of the metagenome of origin in Mbp

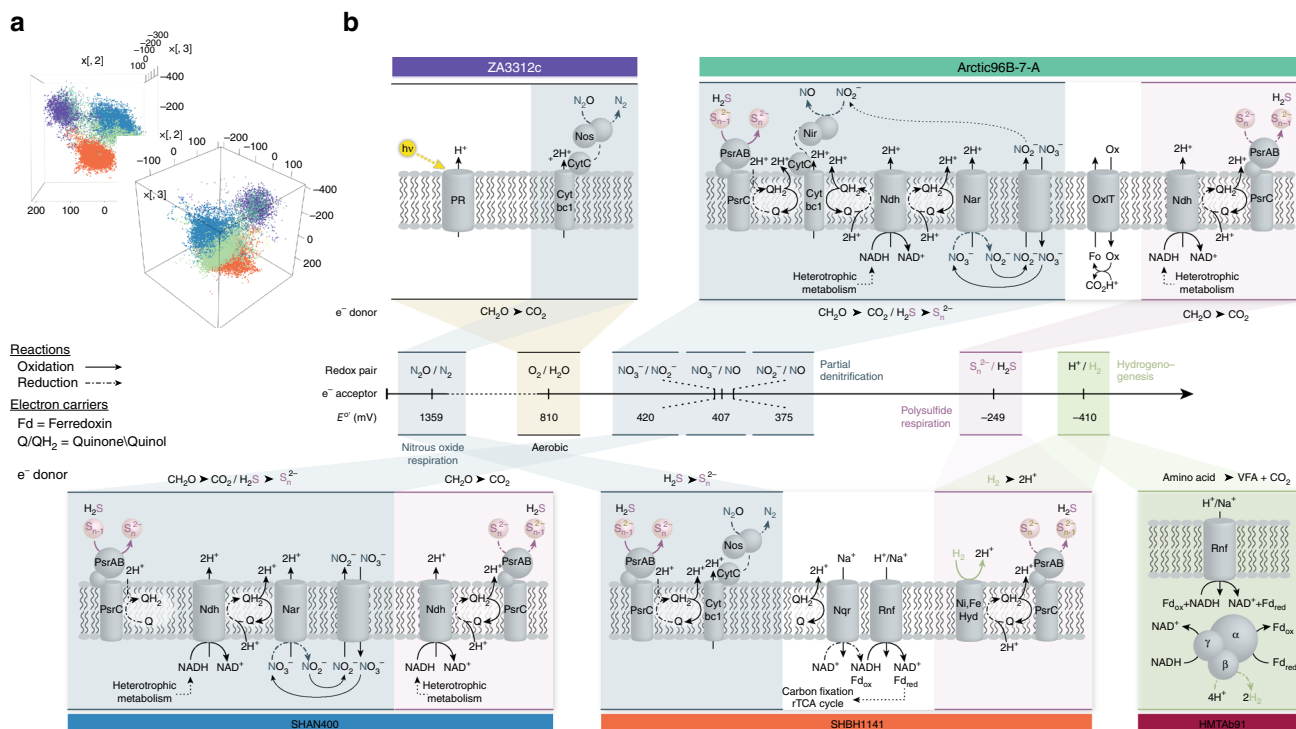
and metatranscriptomic resources from NESAP and Saanich Inlet time series<sup>24, 25</sup> to construct population genome bins, improving estimated genome completion to an average of 87% (Supplementary Data 5). Metagenomic contigs  $>5000$  bp and with  $>95\%$  identity to SAGs were identified followed by tetranucleotide frequency analysis to resolve specific clades (Fig. 3a). A total of five population genomes for Marinimicrobia clades ZA3312c-A/B, HF770D10, Arctic96B-7-A/B, SHAN400, and SHBH1141 spanning oxic, dysoxic, suboxic, anoxic, and anoxic-sulfidic conditions were resolved from Saanich Inlet and NESAP metagenomes, enabling more complete metabolic reconstruction within each clade (Fig. 3a, b). A sixth clade (HMTAb91-A), endemic to a methanogenic bioreactor branching near the base of Marinimicrobia radiation was included in downstream comparisons of metabolic potential to encompass the complete range of electron donor-acceptor pairs. Energy metabolism of Marinimicrobia population genomes was examined in relation to tree branching patterns and environmental disposition. A total of 18 metatranscriptomes from six depths and three time points (Fig. 4b) were used to explore Marinimicrobia gene expression over defined energy gradients including a deep

water renewal event resulting in the influx of oxygenated nutrient rich waters in Saanich Inlet basin waters. This enabled the resolution of metabolic niches and indicated potential modes of metabolic coupling within specific Marinimicrobia clades.

#### Metabolic reconstruction and gene model validation.

Marinimicrobia clades ZA3312c-A/B and HF770D10 were most abundant under oxic water column conditions with extensive genome streamlining comparable to *Ca. Pelagibacter* (Supplementary Fig. 1A). All three clades harbored genes encoding for aerobic respiration, and heterotrophy with no indication for autotrophic  $\text{CO}_2$  fixation. ZA3312c clades also encoded the oxidative tricarboxylic acid (TCA) cycle (Supplementary Data 6) and proteorhodopsin, a proton-pump used to harness light energy (Fig. 3b)<sup>26</sup>. ZA3312c proteorhodopsin transcripts were highly expressed in oxic surface waters of Saanich Inlet, suggesting that ZA3312c are capable of supplementing organotrophy with phototrophy in surface waters, a trait well suited to open-ocean oligotrophic environments (Supplementary Fig. 6A). Interestingly, ZA3312c-A encoded nitrous oxide reductase (*nozZ*)





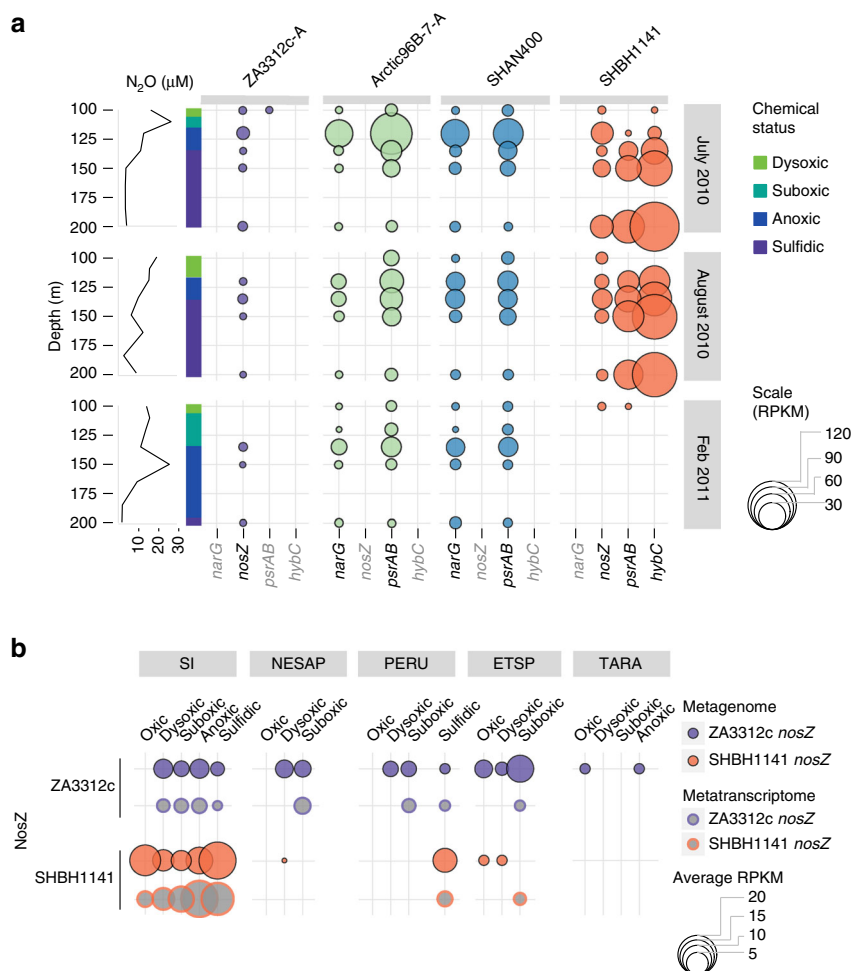
**Fig. 3** Energy metabolism of Marinimicrobia population genome bins. **a** Binning of Marinimicrobia population genomes by Kmer frequency principal component analysis, two rotations of three-dimensional plot, clouds of color coded genome bins are apparent. **b** Summary of co-metabolic and energy metabolism and conservation strategies of Marinimicrobia population genomes from along eco-thermodynamic gradients, for oxygen (beige), nitrogen (blue), sulfur (pink), and hydrogen (green). Enzymes include: proteorhodopsin (PR), sulfur: polysulfide reductase (PsrAB, PsrC); nitrogen: nitrite reductase (Nir), nitrate reductase Nar, nitrate/nitrite antiporter (NirK), nitrous oxide reductase (Nos); hydrogen metabolism: Ni,Fe hydrogenase (Ni,Fe Hyd), hydrogenase complex (HydBD); respiratory elements: cytochrome bc1 complex (Cytbc1), NADH dehydrogenase (Ndh), energy-conserving putative electron transfer mechanisms putative ion-translocating ferredoxin:NADH oxidoreductase (IfoAB); oxalate transporter (OxIT); *Rhodobacter* nitrogen fixation complex (Rnf). Oxidation and reduction indicated by solid or dotted arrows, respectively

and associated maturation factors (*nosL*, *nosD*, and *nosY*) that drive the conversion of N<sub>2</sub>O to N<sub>2</sub> in the terminal step of denitrification. Transcripts for *nosZ* were expressed throughout the Saanich Inlet water column (Fig. 4a; Supplementary Fig. 7) and indicate potential coupling to ammonia oxidizing *Thaumarchaea* that produce N<sub>2</sub>O as a byproduct of ammonia oxidation<sup>27</sup>. ZA3312c-A *nosZ* transcripts were also detected in suboxic waters of the NESAP, Peru, and ETSP OMZs, and four TARA oceans metagenomes contained ZA3312c-A *nosZ* sequences (>80% nucleotide identity) (Fig. 4b) reinforcing a global distribution pattern with functional implications for marine nitrogen budgets and greenhouse gas cycling. Marinimicrobia clades Arctic96B-7-A and B were widespread in dysoxic ocean waters. Arctic96B-7 clades harbored genes encoding for aerobic respiration, organotrophy and oxidative TC cycle with no indication for proteorhodopsin or autotrophic CO<sub>2</sub> fixation (Supplementary Data 6). Arctic96B-7 clades may supplement energy generation in a similar manner to proteorhodopsin through catabolism of the common ocean compound oxalate<sup>28</sup>, coupling a unique oxalate:formate antiporter and oxalate decarboxylase<sup>29</sup>. The Arctic96B-7-A clade also encoded nitrate reductase (*narG*), and polysulfide (polyS) reductase (*psrABC*) (Figs. 2 and 3b; Supplementary Figs. 6A and 8) that were expressed throughout the Saanich Inlet water column. Peak expression corresponded to depths with low NO<sub>3</sub><sup>-</sup> and no detectable H<sub>2</sub>S (Fig. 4a; Supplementary Fig. 6A). Interestingly, the PsrABC enzyme complex can use H<sub>2</sub>S as an auxiliary electron donor through PsrABC-mediated H<sub>2</sub>S oxidation to polyS and stored polyS can serve as an alternative electron sink, regenerating H<sub>2</sub>S. The combination of *narG* and *psrABC* provides Arctic96B-7 clades with versatile energy

metabolism with potential coupling to both sulfur oxidizing bacteria (ARCTIC96-BD19, SUP05) by regenerating H<sub>2</sub>S under non-sulfidic conditions, and anaerobic ammonium (*Planctomycetes*) and nitrite (*Nitrospina*) oxidizing bacteria through the production of NO<sub>2</sub><sup>-</sup> in dysoxic, suboxic, and anoxic waters (Fig. 5a). Thus, Arctic96B-7 clades may form supportive metabolic partnerships with major primary producers in OMZs critical to the biogeochemical cycling of carbon, nitrogen, and sulfur<sup>12</sup>.

Marinimicrobia clade SHAN400 appears to be endemic to Saanich Inlet where it is most abundant below the oxycline (Supplementary Fig. 4). SHAN400 harbored genes encoding for aerobic and anaerobic respiration, heterotrophy and oxidative TCA cycle. SHAN400 also encoded ferredoxin, pyruvate metabolism, and NADH dehydrogenase (Fig. 3b; Supplementary Figs. 8 and 9), potentially providing additional electron shuttles for energy metabolism under anoxic conditions. Similar to Arctic96B-7, SHAN400 encoded *narG* and *psrABC*, potentially linking its energy metabolism to both sulfur-oxidizing bacteria (SUP05) and anaerobic ammonium- (*Planctomycetes*) and nitrite- (*Nitrospina*) oxidizing bacteria in anoxic waters (Figs. 3 and 4; Supplementary Fig. 6A, B). In contrast to Arctic96B-7, SHAN400 transcripts for heme/copper-type cytochrome and NADH dehydrogenase were most highly expressed in anoxic waters (Supplementary Fig. 9A). This is consistent with redox-driven niche partitioning between Arctic96B-7 and SHAN400 clades in the Saanich Inlet water column.

Marinimicrobia clade SHBH1141 was prevalent in anoxic and anoxic-sulfidic OMZ waters (Supplementary Fig. 4). SHBH1141 harbored genes encoding for aerobic and anaerobic respiration, autotrophic CO<sub>2</sub> fixation via the reductive TCA cycle (citrate

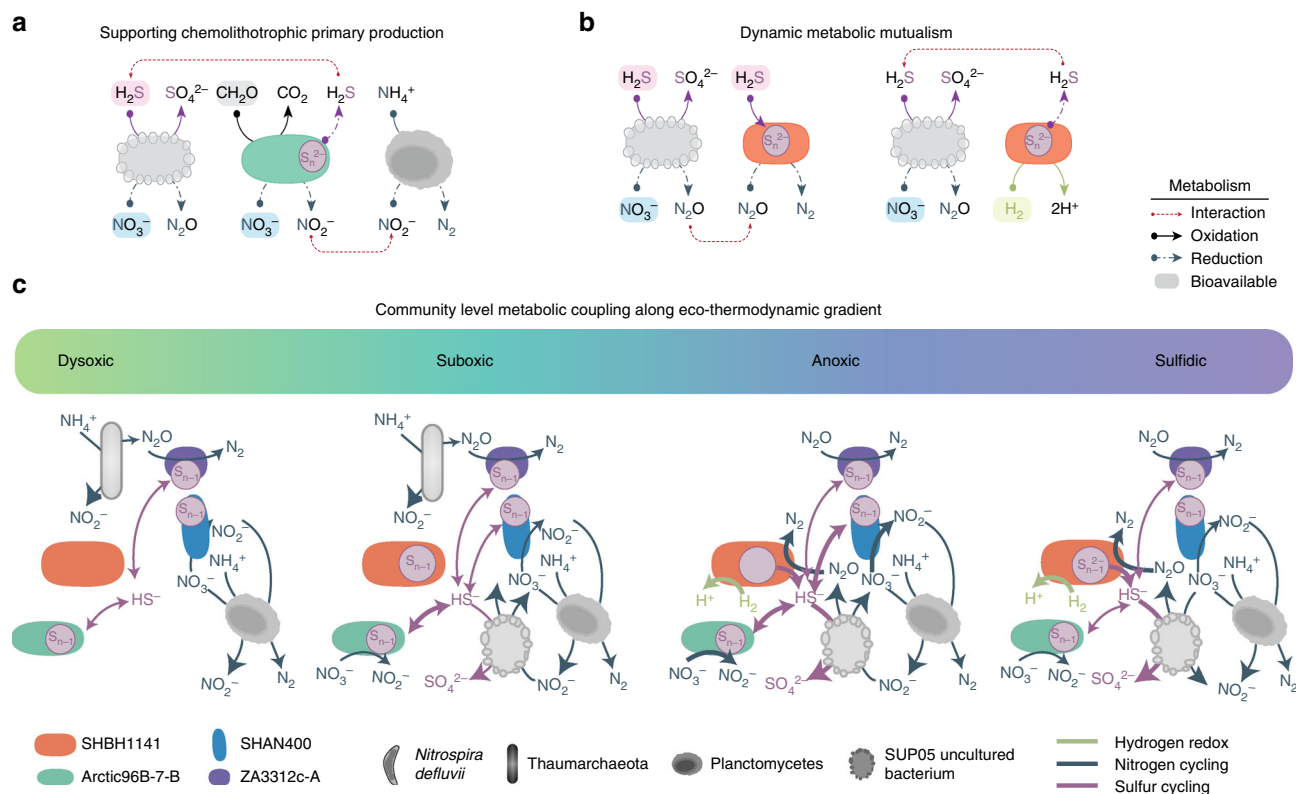


**Fig. 4** Expression of selected Marinimicrobia energy metabolism genes. **a** Expression of selected genes involved in Marinimicrobia energy metabolism in Saanich Inlet station SI03 at three time points and five depths between 100 and 200 m. Size of circle represents reads per kilobase per million mapped (RPKM)<sup>52</sup> for metatranscriptomic reads mapped to the selected genes for the indicated population genomes. Water column redox state for each time point encoded on left axis and nitrous oxide concentration profile for each time point on left. Enzymes: nitrate reductase (*narG*), Nitrous oxide reductase (*nosZ*), polysulfide reductase subunits A and B (*psrAB*) and Ni-Fe hydrogenase subunits A and B (*hybC*). **b** Detected genes and transcripts for Marinimicrobia ZA3312c and SHBH1141 *nosZ* along eco-thermodynamic gradients from oxic (>90  $\mu\text{mol O}_2$ ), dysoxic (20–90  $\mu\text{mol O}_2$ ), suboxic (1–20  $\mu\text{mol O}_2$ ), anoxic (<2  $\mu\text{mol O}_2$ ), and sulfidic conditions in Saanich Inlet (SI) time series, Northeastern Subarctic Pacific (NESAP), Peru, Eastern Tropical South Pacific (ETSP), and TARA Oceans (no transcriptomes available) data sets. For SI and ETSP dot size represents average reads per kilobase per million mapped (RPKM) summed for a given *nosZ* type for each metagenome or metatranscriptome and averaged by the total number of metagenomes or metatranscriptomes for a given water column classification. For ETSP, Peru, and TARA bubble size is the number of reads (ETSP and Peru) or contigs (TARA) with *nosZ* averaged per number of metagenome or metatranscriptomes for a given water column classification

lyase and ferredoxin-dependent 2-ketoacid oxidoreductases), and the *Rhodobacter* nitrogen fixation (Rnf) complex to produce reduced ferredoxin to drive endergonic reductive carboxylation steps, indicating a capacity to perform anaerobic autotrophy (Supplementary Figs. 8 and 9). In addition, SHBH1141 encoded *psrABC*, class I [Ni,Fe] hydrogenases (*hybOABCD*) and *nosZ* with associated maturation factors *nosL* and *nosD* (Fig. 3b; Supplementary Figs. 6A and 8). Gene expression for *psrABC*, *hybOABCD*, and *nosZ* was elevated under anoxic to sulfidic conditions (120 m in July 2010, and 150 m in July and August 2010; Fig. 4). SHBH1141 class I [Ni,Fe] hydrogenase is proposed to operate bidirectionally based on observations in *Escherichia coli* and *Salmonella enterica*, with proposed hydrogen production under more oxidizing conditions<sup>30</sup>. SHBH1141 *nosZ* was recovered on a global scale and expressed under both sulfidic conditions in Peru and suboxic conditions in the ETSP as well as Saanich Inlet (Fig. 4b), positing a central role for SHBH1141 in OMZ  $\text{N}_2\text{O}$  reduction. The expression of these genes in

anoxic–sulfidic waters points to a new mode of dynamic metabolic mutualism in which SHBH1141 may rely on SUP05  $\text{N}_2\text{O}$  generation in anoxic and sulfidic waters<sup>12,31</sup> to store polyS and re-evolve  $\text{H}_2\text{S}$  from polyS to stimulate SUP05  $\text{N}_2\text{O}$  production (Fig. 5b). This would in turn support autotrophic carbon fixation in both partners and sustains N and S biogeochemical cycling under dynamic or unfavorable conditions (e.g., limited  $\text{H}_2\text{S}$  bioavailability; Supplementary Fig. 5). Such mutualism would be highly dependent on either (a) migration along the eco-thermodynamic gradient or (b) seasonal/temporal changes such as renewal or upwelling events.

Marinimicrobia clades HMTAb91-A/B are prevalent in methanogenic locations at the base of the electron tower. Apparently, HMTAb91-A/B did not harbor genes for aerobic respiration and had an incomplete TCA cycle. HMTAb91-A encoded the Embden–Meyerhof–Parnas pathway (Supplementary Data 6) and both HMTAb91-A/B encoded energy-conserving  $\text{H}^+$  respiration through electron-confurcating hydrogenases, the



**Fig. 5** Proposed co-metabolic model along eco-thermodynamic gradient in Saanich Inlet. **a** Proposed metabolic coupling between ARCTIC96B-1, SUP05, and Planctomycetes. **b** Proposed dynamic metabolic mutualism between SUP05 and SHBH1141. **c** Conceptual model for Marinimicrobia co-metabolic activity with other major microbial groups in Saanich Inlet along eco-thermodynamic gradients. Interactions based on expression data for sulfur (pink), nitrogen (blue), and hydrogen (green) for dominant Marinimicrobia clades in Saanich Inlet as well as putative metabolic partners *Nitrosopumulaceae* sp., *Planctomycetes*, and SUP05

energy-conserving (Rnf complex) and putative syntrophic amino-acid metabolism through the ion-translocating ferredoxin:NADH oxidoreductase (*ifoAB*) (Fig. 3b)<sup>23</sup>. Within the methanogenic reactor where it was initially described, HMTAb91-A is postulated to accomplish thermodynamically unfavorable amino-acid degradation supporting methanogenesis<sup>23</sup>. HMTAb91-A/B clades appear restricted to methanogenic ecosystems as no metagenomic or metatranscriptomic sequences were recruited from non-methanogenic locations.

## Discussion

Co-metabolic functions encoded and expressed within globally distributed Marinimicrobia clades would fill several hitherto unassigned niches in the nitrogen and sulfur cycles and support recent modeling efforts integrating biogeochemical and multi-omic sequence information in the Saanich Inlet water column<sup>24,32,33</sup> (Fig. 5). The N<sub>2</sub>O reductase expressed on a global basis by ZA3312c-A and SHBH1141 clades has the potential to act as a biological filter for N<sub>2</sub>O produced by the ubiquitous marine processes of ammonia oxidation (e.g., *Thaumarchaeota*)<sup>27</sup> and partial denitrification (e.g., SUP05)<sup>12,31</sup>. In contrast, nitrate reduction to NO<sub>2</sub><sup>-</sup> by other Marinimicrobia clades (i.e., Arctic96B-7-A and SHAN400) has potential to provide NO<sub>2</sub><sup>-</sup> to anaerobic ammonium-oxidizing (*Planctomycetes*) and nitrite-oxidizing (*Nitrospina*) bacteria in dysoxic, suboxic, and anoxic waters. The polysulfide reductase expressed by multiple Marinimicrobia clades (e.g., Arctic96B-7, SHAN400, and SHBH1141) has potential to provide an energy storage mechanism via accumulation of polyS that can be reduced or oxidized under changing water column redox conditions and support both

cooperative and dynamic interactions including cryptic sulfur cycling and dark carbon fixation<sup>34</sup>.

The application of eco-thermodynamics principles to microbial ecology provides perspective on how thermodynamic constraints serve to shape microbial community structure and the nature of co-metabolic interactions along energy gradients. Indeed, phylogenetic branching patterns often coincided with energy yields of redox pairs for identified clade energy metabolism, with deeper branching clades near the base of the electron tower where lower energy yields would increase potential for metabolic coupling. Additionally, many Marinimicrobia clades encoded enzyme systems tied to both nitrogen- and sulfur-cycling, suggesting extensive specialization for metabolic cooperation bridging within and between biogeochemical cycles. Such dependencies likely confound isolation efforts within the phylum and point to an ancestral state primed for co-existence. The extent to which this reflects the diversification of other phyla, particularly MDM across the Tree of Life is an interesting area of research with implications for understanding and directing the evolution of metabolic networks driving Earth's biogeochemical cycles.

## Methods

**SAG collection, sequencing, assembly, and decontamination.** SAGs from Gulf of Maine, HOT station ALOHA, South Atlantic Gyre, the Terephthalate degrading bioreactor and Etoliko Lagoon Sediment were included in Rinke et al.<sup>17</sup>, and collection, assembly and decontamination follows accordingly. See Supplementary Data 1 for details on SAG genomics. SAGs from Northeast subarctic Pacific (NESAP) and Saanich Inlet followed the following protocol. Replicate 1-ml aliquots of sea water collected for single-cell analyses were cryopreserved with 6% glycine betaine (Sigma-Aldrich), frozen on dry ice and stored at -80 °C. Single-cell sorting, whole-genome amplification, real-time PCR screens, and PCR product sequence analyses were performed at the Bigelow Laboratory for Ocean Sciences Single Cell



Genomics Center ([www.bigelow.org/scgc](http://www.bigelow.org/scgc)), as described by Stepanauskas and Sieracki<sup>35</sup>. SAGs from the NESAP were generated at the DOE Joint Genome Institute (JGI) using the Illumina platform as described in Rinke et al.<sup>17</sup>. SAGs from Saanich Inlet were sequenced at the Genome Sciences Centre, Vancouver BC, Canada, as described in Roux et al.<sup>36</sup>. All SAGs were assembled at JGI as described in Rinke et al.<sup>17,36</sup>.

The following steps were performed for SAG assembly: (1) filtered Illumina reads were assembled using Velvet version 1.1.04<sup>37</sup> using the VelvetOptimiser script (version 2.1.7) with parameters: (--v --s 51 --e 71 --i 4 --t 1 --o --ins\_length 250 --min\_contig\_lgth 500") 2) wgsim (-e 0 -l 100 -r 0 -R 0 -X 0) 3) Allpaths-LG (prepareAllpathsParams: PHRED\_64 = 1 PLOIDY = 1 FRAG\_COVERAGE = 125 JUMP\_COVERAGE = 25 LONG\_JUMP\_COV = 50, runAllpathsParams: THREADS = 8 RUN = std\_pairs TARGETS = standard VAPI\_WARN\_ONLY = True OVERWRITE = True). SAG prediction analysis and functional annotation was performed within the Integrated Microbial Genomes (IMG) platform<sup>38</sup> (<http://img.jgi.doe.gov>) developed by the Joint Genome Institute, Walnut Creek, CA, USA.

**Phylogenomic analysis of SAGs.** The PhyloPhlAn pipeline was used to determine relationships among Marinimicrobia SAGs<sup>39</sup> (Supplementary Fig. 3) as well as the phylogenetic placement of Marinimicrobia within the bacterial domain (Supplementary Fig. 2). In both cases, fasta files for the 25 SAGs and related genomes were passed to PhyloPhlAn and resulting trees were visualized and drawn using GraPhlAn. The 25 Marinimicrobia SAGs and related genomes were inserted into the already built PhyloPhlAn microbial Tree of Life containing 3737 genomes using the “insert” functionality, and a de novo phylogenetic tree was created using the “user” functionality based solely on the 25 Marinimicrobia SAG and related genome fasta files. Default parameters were used in each case with the exception of a custom annotation file used in GraPhlAn to colour the leaves based on phylum in the microbial Tree of Life, and subgroup in the de novo phylogenetic tree.

**Metagenome fragment recruitment.** The proportion of Marinimicrobia represented in the 594 globally distributed metagenomes (Supplementary Fig. 3A) was determined by SAG nucleotide sequence alignment to individual metagenomes using FAST<sup>40</sup>. Parameters of 70% nucleotide identity cutoff over 70% of the contig length (or 454-read, where applicable) were employed to encompass the Marinimicrobia phylum<sup>41</sup>. The small subunit ribosomal RNA (SSU rRNA) gene was removed from SAG sequences before alignment searches to prevent cross-recruitment to non-Maritimicrobia sequences. The total length of contigs passing the cutoff for a given metagenome was summed and divided by the total contig length for that metagenome to calculate percentage of Marinimicrobia. Where data on O<sub>2</sub> concentration was available, for Saanich Inlet, NESAP, ETSP<sup>15</sup>, and Peruvian upwelling<sup>42</sup>, O<sub>2</sub> status of the sample was used as indicated. Data on O<sub>2</sub> concentration were unavailable for Marine-Misc. and terrestrial samples.

Biogeography of Marinimicrobia SAG-affiliated clades was similarly determined using alignment parameters of 95% identity cutoff and >200 base pairs (bp) alignment length to ensure only contigs with high sequence similarity while maintaining clade resolution. Metagenomic contigs mapping to more than one Marinimicrobia clade were assigned to the clade with greatest percent identity and in the event of a tie were assigned to the clade with the greatest alignment length. Overall abundance was calculated for each metagenome by summing the total lengths of all contigs with hits to a given Marinimicrobia clade divided by the total size of the SAG and the total size of the assembled metagenome in base pairs. Results by metagenome (Supplementary Datas 2 and 3) and clade were then summed in Fig. 1c and itemized in Supplementary Data 4. Global relative abundance of Marinimicrobia clades shown in Supplementary Fig. 3B was calculated similarly by summing the total lengths of all contigs with hits to a given Marinimicrobia clade divided by the total size of the SAG and the total size of the assembled metagenome in base pairs and then summing for all hits to a given clade.

**Saanich Inlet and NESAP metagenomes and metatranscriptomes.** Saanich Inlet metagenomes and metatranscriptomes were collected, sequenced, and assembled as described in Hawley et al.<sup>24</sup> and cognate chemical and physical measurements can be found in and Torres-Beltran et al.<sup>32</sup>. Briefly, Saanich Inlet samples for metagenomic and metatranscriptomic sequencing were collected by Niskin or Go-Flow on line with CTD. Samples for metatranscriptomics, 2l, were filtered by peristaltic pump with in-line 2.7 μM prefilter onto a sterivex filter with 1.8 ml RNALater added and frozen on dry ice within 20 min of bottle on-deck. Metagenomic samples, 20l, were filtered within 8 h of collection by peristaltic pump with in-line 2.7 μM prefilter onto a sterivex filter with 1.8 ml lysis buffer added and frozen at -80 °C. Metagenomic and metatranscriptomic samples were processed, sequenced, and assembled according to Hawley et al.<sup>24</sup> at the JGI using the Illumina HiSeq platform.

Sampling in the NESAP was conducted via multiple hydrocasts using a Conductivity, Temperature, Depth (CTD) rosette water sampler aboard the CCGS *John P. Tully* during three Line P cruises: 2009-09 [June 2009, major stations P4 (48°39.0 N, 126°4.0 W, 7 June), P12 (48°58.2 N, 130°40.0 W, 9 June), and P26 (50°N, 145°W, 14 June), 2009-10 [August 2009, major stations P4 (21 August), P12

(23 August) and P26 (27 August)], and 2010-01 [February 2010, major stations P4 (4 February) and P12 (11 February)]. At these stations, large volume (20 l) samples for DNA isolation were collected from the surface (10 m), while 120 l samples were taken from three depths spanning the OMZ core and upper and deep oxyclines (500, 1000, 1300 m at station P4; 500, 1000, 2000 m at station P12). Sequencing and assembly was carried out as described above for Saanich Inlet and accession numbers are available in Supplementary Data 2.

**Construction and validation of population genome bins.** Marinimicrobia population genome bins were constructed by identifying metagenomic contigs from Saanich Inlet, and NESAP metagenomes mapping to specific SAG(s) using a supervised binning method based in part on methodologies developed by Dodsworth et al.<sup>43</sup> in the construction of OP9 population genome bins. Initially, determination of membership of individual SAGs to SAG-clusters making up a given phylogenetic clade was conducted. SAG tetranucleotide frequencies were then calculated and converted to z-scores with TETRA (<http://www.megx.net/tetra>)<sup>44,45</sup>. Z-scores were reduced to three dimensions with principal component analysis (PCA) using PRIMER v6.1.13<sup>46</sup> and hierarchical cluster analysis of the z-score PCA with Euclidian distance (also performed in PRIMER) was carried out to generate SAG-clusters. These SAG-clusters reflected phylogenetic placement of the SAGs by SSU rRNA gene analysis. For construction of population genome bins, metagenomic contigs from NESAP and SI data sets were aligned to SAG contigs with >95% nucleotide identity using BLAST<sup>47</sup> and a minimum of 5 kilobase pairs alignment length, Tetranucleotide frequencies of all metagenomic contigs passing this identity and length threshold were calculated and converted to z-scores. SAG-supervised binning as described in Dodsworth et al. using linear discriminant analysis was carried out using all z-scores with the SAG-bins as training data to classify the metagenomic contigs as making up a given population-genome bin.

Individual SAGs and population genome bins were analyzed for completeness and strain heterogeneity using CheckM v1.0.5<sup>54</sup>. Specifically, the lineage\_wf workflow was used with default parameters. The lineage\_wf workflow includes determination of the probable phylogenetic lineage based on detected marker genes. The determined lineage then dictates the sets of marker genes that is most relevant for estimating a given genome's completeness and other statistics. The strain heterogeneity metric is highly informative for population genome bins as it is essentially the average amino-acid identity for pairwise comparisons of the (lineage appropriate) redundant single-copy marker genes within a population genome bin (Supplementary Data 5). For population genome bins the higher the strain heterogeneity value, the more similar the amino acid identity of the redundant marker genes indicating the sequences in the bin originate from a closely related, if not identical, phylogenetic source.

**Maritimicrobia genome streamlining.** Gene-coding bases and COG-based gene redundancy shown in Supplementary Fig. 1A, B were calculated using cluster of orthologous group (COG)-based genome redundancy as described in Rinke et al.<sup>17</sup>. Each gene's COG category was predicted through the JGI IMG pipeline. COG redundancy was calculated by averaging the occurrence of each COG in the genome. The percentage of gene-coding bases was calculated by dividing the number of bases contributing to protein- and RNA-coding genes by the total genome size. For SAGs, the length of the assembled genome was used rather than the estimated genome size.

**Annotation and identification of metabolic genes of interest.** Genes of interest were identified in the SAGs and in IMG/M (<https://img.jgi.doe.gov/cgi-bin/m/main.cgi>)<sup>48</sup> for the metagenomic contigs which made up the population genome bins. Contigs making up Marinimicrobia population genome bins were run through MetaPathways 2.5<sup>49,50</sup> to annotate open reading frames (ORFs) and reconstruct metabolic pathways. As the population genome bins were constructed from multiple metagenomes they contained redundant sequence information, BLASTp<sup>47</sup> (amino-acid identity cutoff >75%) was used to identify all copies of a given gene of interest in each population genome bin, which was then used in gene model validation and expression mapping.

**Gene expression mapping.** Metatranscriptomes from three time points in Saanich Inlet time series<sup>24</sup> were used to investigate changes in gene expression along water column redox gradients over time for selected ORFs involved in energy metabolism and electron shuttling. Quality controlled reads from metatranscriptomes were mapped to identified ORFs of interest using bwa -mem<sup>51</sup> and reads per kilobase per million mapped (RPKM) per ORF was calculated using RPKM calculation in MetaPathways 2.5<sup>52</sup>. For each population genome bin RPKM values for a given sample were summed for ORFs with the same functional annotation to yield an RPKM for a given functional gene. For other taxonomic groups in Saanich Inlet shown in Supplementary Fig. 6B, genes were identified by sequence alignment searches of Saanich Inlet metatranscriptomes (bioSample indicated above) assembled and conceptually translated using BLASTp against selected nitrogen and sulfur cycling genes from Hawley et al.<sup>12</sup> and RPKM values calculated as described above.



**Global distribution and expression of *nosZ*.** Further analysis was carried out to determine the global distribution of Marinimicrobia *nosZ* in 594 metagenomes. The *nosZ* nucleotide sequences from SHBH1141 and ZA3312c, which exhibited a 65% nucleotide identity to each other by BLAST, were clustered at 95% identity using the USEARCH cluster fast algorithm<sup>53</sup>, resulting in three clusters, two SHBH1141 and one ZA3312c. Nucleotide sequence alignment was carried out using FAST<sup>40</sup>, with parameters of >80% nucleotide identity and >60 bp alignment length against 594 metagenomes. For Saanich Inlet and NESAP data sets, abundance of *nosZ* in a given metagenome or metatranscriptome was determined by summing the RPKM value for ORF hits to either SHBH1141 or ZA3312c for a given metagenome or metatranscriptome. For 454 sequenced<sup>15,42</sup> metagenomes and metatranscriptomes (Peru<sup>42</sup> and ETSP<sup>15</sup>), the number of reads which hit to either SHBH1141 or ZA3312c were summed for a given metagenome. For the TARA Oceans data set, the number of genes identified in an assembled metagenome was summed. Metatranscriptomic data for Tara was unavailable at this time.

**Data availability.** Single-cell amplified genomes and associated assemblies generated for this study from Saanich Inlet and the northeastern subarctic Pacific Ocean are available in JGI IMG with Taxon OIDs: 2537562244, 2537562243, 2537562242, 2537562237, 2537562241, 2537562240, 2537562239, 2537562238, and 2537562245. Metagenomes from Saanich Inlet and the northeastern subarctic Pacific Ocean are available at NCBI with BioSample accession codes: SAMN0324878 to SAMN0324887, SAMN0324895 to SAMN0324900, SAMN0324919, SAMN0324920, SAMN0324964 to SAMN0324982, and SAMN0324987 to SAMN0324991. Metatranscriptomes used for expression analysis are available at NCBI with BioSample accession codes: SAMN05223291 to SAMN05223293, SAMN05224498 to SAMN05224507, SAMN05224510, SAMN05224511, SAMN05224516, SAMN05224517, and SAMN05236416.

Received: 28 November 2016 Accepted: 11 September 2017

Published online: 15 November 2017

## References

- Georgescu-Roegen, N. *The Entropy Law and the Economic Process* (Harvard University Press, Cambridge, MA, 1971).
- Ayres, R. U. Eco-thermodynamics: economics and the second law. *Ecol. Econ.* **26**, 1282–1285 (1997).
- Falkowski, P. G., Fenchel, T. & DeLong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
- Reed, D. C., Algar, C. K., Huber, J. A. & Dick, G. J. Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proc. Natl Acad. Sci. USA* **111**, 1879–1884 (2014).
- Louca, S. et al. Integrating biogeochemistry with multi-omic sequence information in a model oxygen minimum zone. *Proc. Natl Acad. Sci. USA* **113**, E5925–E5933 (2016).
- Hug, L. A. et al. Critical bihid with bacteria from new phyla and little studied lineages. *Environ. Microbiol.* **18**, 159–173 (2016).
- Tripp, H. J. et al. SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452**, 741–744 (2008).
- DeLong, E. F. Life on the thermodynamic edge. *Science* **317**, 327–328 (2007).
- Morris, B. E., Henneberger, R., Huber, H. & Moissl-Eichinger, C. Microbial syntrophy: interaction for the common good. *FEMS Microbiol. Rev.* **37**, 384–406 (2013).
- Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).
- Aylwarda, F. O. et al. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc. Natl Acad. Sci. USA* **112**, 5443–5448 (2015).
- Hawley, A. K., Brewer, H. M., Norbeck, A. D., Paša-Tolic, L. & Hallam, S. J. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proc. Natl Acad. Sci. USA* **111**, 11395–11400 (2014).
- Wright, J. J., Konwar, K. M. & Hallam, S. J. Microbial ecology of expanding oxygen minimum zones. *Nat. Rev. Microbiol.* **10**, 381–394 (2012).
- Walsh, D. A. et al. Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* **326**, 578–582 (2009).
- Stewart, F. J., Ulloa, O. & DeLong, E. F. Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ. Microbiol.* **14**, 23–40 (2012).
- Tsementzi, D. et al. SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature* **536**, 179–183 (2016).
- Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
- Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 6048 (2016).
- Gies, E. A., Konwar, K. M., Beatty, J. T. & Hallam, S. J. Illuminating microbial dark matter in meromictic Sakinaw Lake. *Appl. Environ. Microbiol.* **80**, 6807–6818 (2014).
- Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
- Allers, E. et al. Diversity and population structure of Marine Group A bacteria in the Northeast subarctic Pacific Ocean. *ISME J.* **7**, 256–268 (2013).
- Wright, J. J. et al. Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. *ISME J.* **8**, 455–468 (2014).
- Nobu, M. K. et al. Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *ISME J.* **9**, 1710–1722 (2015).
- Hawley, A. K. et al. A compendium of multi-omic sequence information from the Saanich Inlet water column. *Sci. Data* **4**, 170160 (2017).
- Hallam, S. J., Torres-Beltran, M. & Hawley, A. K., Monitoring microbial responses to ocean deoxygenation in a model oxygen minimum zone. *Sci. Data* **4**, 170158 (2017).
- Béjà, O. et al. Bacterial Rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906 (2000).
- Santorio, A. E., Buchwald, C., McIlvin, M. R. & Casciotti, K. L. Isotopic signature of N<sub>2</sub>O produced by marine ammonia-oxidizing archaea. *Science* **333**, 1282–1285 (2011).
- Steinberg, S. M. & Badal, J. L. Oxalic, glyoxalic and pyruvic acids in eastern Pacific Ocean waters. *J. Mar. Res.* **42**, 697–708 (1984).
- Anantharam, V., Allison, M. J. & Maloney, P. C. Oxalate: formate exchange. *J. Biol. Chem.* **264**, 7244–7250 (1989).
- Greening, C. et al. Genomic and metagenomic surveys of hydrogenase distribution indicate H<sub>2</sub> is a widely utilised energy source for microbial growth and survival. *ISME J.* **10**, 761–777 (2016).
- Shah, V., Chang, B. X. & Morris, R. M. Cultivation of a chemoautotroph from the SUP05 clade of marine bacteria that produces nitrite and consumes ammonium. *ISME J.* **11**, 263–271 (2017).
- Torres-Beltrán, M. et al. A compendium of geochemical information from the Saanich Inlet water column. *Sci. Data* **4**, 170159 (2017).
- Capelle, D. W., Hawley, A. K., Hallam, S. J. & Tortell, P. D. A multi-year time-series of N<sub>2</sub>O dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Limnol. Oceanogr.* <http://dx.doi.org/10.1002/lno.10645> (2017).
- Canfield, D. E. et al. A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. *Science* **330**, 1375–1378 (2010).
- Stepanuskas, 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).
- Roux, S. et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3**, e03125 (2014).
- Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
- Markowitz, V. M. et al. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, D115–D122 (2012).
- Segata, N., Bornigen, D., Morgan, X. C. & Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **4**, 2304 (2013).
- Kim, D., Hahn, A. S., Hanson, N. W., Konwar, K. M. & Hallam, S. J. FAST: fast annotation with synchronized threads. in *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology* pp 1–8 (IEEE, Chiang Mai, 2016).
- Varghese, N. J. et al. Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* **43**, 6761–6771 (2015).
- Schunck, H. et al. Giant hydrogen sulfide plume in the oxygen minimum zone off Peru supports chemolithoautotrophy. *PLoS ONE* **8**, e68661 (2013).
- Dodsworth, J. A. et al. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat. Commun.* **4**, 1854 (2013).
- Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glockner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938–947 (2004).
- Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. & Glockner, F. O. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* **5**, 163 (2004).
- Clarke, K. R. & Gorley, R. N. PRIMER v6: User Manual/Tutorial. (PRIMER-E, Plymouth, 2006).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **216**, 403–410 (1990).
- Markowitz, V. M. et al. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**, 2271–2278 (2009).

49. Konwar, K. M., Hanson, N. W., Page, A. P. & Hallam, S. J. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* **14**, 202 (2013).
50. Hanson, N. W. et al. Metabolic pathways for the whole community. *BMC Genomics* **15**, 619 (2014).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. Konwar, K. M. et al. MetaPathwaysv2.5: quantitative functional, taxonomic and usability improvements. *Bioinformatics* **31**, 3345–3347 (2015).
53. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
54. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 7 (2015).

## Acknowledgements

We thank the Joint Genome Institute (JGI), including Susannah Tringe, Stephanie Malfatti, and Tijana Glavina del Rio, for technical and project management assistance. We thank Captain Ken Brown and his crew for all their support aboard The RSV Strickland, as well as our sea-going technicians at UBC, Chris Payne and Laura Pakhomova. We thank the scientists and crew aboard CCGS *John P. Tully*, in particular Marie Robert, as well as Fisheries and Oceans Canada for logistical support. We thank the officers and crew of the RV *Ka'imikai-O-Kanaloa* and the HOT team for sample collection at station ALOHA, and Jane Heywood and Michael Sieracki for South Atlantic field sample collection. We thank the many technicians and undergraduate helpers in the Hallam lab for support. This work was performed under the auspices of the US Department of Energy (DOE) JGI supported by the Office of Science of US DOE Contract DE-AC02-05CH11231, by National Science Foundation Grants OCE-1232982 (to R.S. and B.K.S.), the G. Unger Vetlesen and Ambrose Monell Foundations, the Tula Foundation-funded Centre for Microbial Diversity and Evolution, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Foundation for Innovation, the Canadian Institute for Advanced Research through grants awarded to S.J.H., and the US National Science Foundation grant OCE-1232982 to R.S. and B.S. J.J.W. was supported by NSERC and the Tula Foundation. M.T.-B. was supported by Consejo Nacional de Ciencia y Tecnología (CONACyT) and the Tula Foundation. A.K.H. was supported by the Tula Foundation.

## Author contributions

A.K.H. carried out biogeography analysis, expression analysis, denitrification pathway analysis, prepared most figures, and aided in energy metabolism analysis. M.K.N. aided

in writing and carried out energy metabolism and genome streamlining analysis and associated figures. J.J.W. assisted in composition and carried out phylogenetic analysis and associated figure production and conception of project. W.E.D. carried out population genome bin construction and regression analyses. C.M.-L. carried out CheckM analysis and read-mapping. B.S. carried out PhyloPhlAn analysis and aided biogeography analysis. P.S. aided in S.A.G. assembly and biogeography. B.K.S. aided in SAG acquisition and biogeography analysis. C.R. aided in biogeography analysis. M.T.-B. aided in *nosZ* expression analysis. K.M. aided in production of Fig. 1b. M.-T.L. aided in energy metabolism and genome streamlining analysis. R.S. carried out S.A.G. sorting and provided feedback on analyses. T.W. aided in S.A.G. decontamination. S.J.H. designed the research, aided in data analysis, and interpretation and supervised the group. A.K.H. and S.J.H. wrote the paper with input from co-authors.

## Additional information

**Supplementary Information** accompanies this paper at doi:10.1038/s41467-017-01376-9.

**Competing interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017