SAR11 bacteria linked to ocean anoxia and nitrogen loss

Despina Tsementzi1, Jieying Wu2, Samuel Deutsch3, Sangeeta Nath3, Luis M. Rodriguez-R3, Andrew S. Burns2, Piyush Ranjan2, Neha Sarode3, Rex R. Malmstrom4, Cory C. Padilla4, Benjamin K. Stone4, Laura A. Bristow5, Morten Larsen6, Jennifer B. Glass7, Bo Thamdrup6, Tanja Woyke7, Konstantinos T. Konstantinidis1,2 & Frank J. Stewart2

Bacteria of the SAR11 clade constitute up to one half of all microbial cells in the oxygen–rich surface ocean. SAR11 bacteria are also abundant in oxygen minimum zones (OMZs), where oxygen falls below detection and anaerobic microbes have vital roles in converting bioavailable nitrogen to N2 gas. Anaerobic metabolism has not yet been observed in SAR11, and it remains unknown how these bacteria contribute to OMZ biogeochemical cycling. Here, genomic analysis of single cells from the world’s largest OMZ revealed previously uncharacterized SAR11 lineages with adaptations for life without oxygen, including genes for respiratory nitrate reductases (Nar). SAR11 nar genes were experimentally verified to encode proteins catalysing the nitrite–producing first step of denitrification and constituted ~40% of OMZ nar transcripts, with transcription peaking in the anoxic zone of maximum nitrate reduction activity. These results link SAR11 to pathways of ocean nitrogen loss, redefining the ecological niche of Earth’s most abundant organismal group.

Alphaproteobacteria of the SAR11 clade form one of the most ecologically dominant organism groups on the planet, representing up to half of the total microbial community in the oxygen–rich surface ocean1–5. All characterized SAR11 isolates, including the globally ubiquitous Candidatus Pelagibacter genus, are aerobic heterotrophs adapted for scavenging dissolved organic carbon and nutrients under the oligotrophic conditions of the open ocean6–9. Gene-based surveys have also revealed diverse SAR11 lineages at high abundance in the deep waters of the meso- and bathypelagic realms10–13. However, the functional properties that distinguish SAR11 bacteria living in distinct ocean regions remain unclear. All known SAR11 genomes are small (typically less than 1.5 megabase pairs (Mb)), with genomic streamlining as a potential adaptation to the nutrient-limiting conditions of the open ocean11. It has been hypothesized that adaptations in SAR11 do not involve large variations in gene content6,8, suggesting that the contribution of SAR11 to ocean biogeochemistry is primarily through its role in aerobic oxidation of organic carbon.

Although genetic or biochemical evidence of anaerobic metabolism has not been reported for SAR11, high abundances of SAR11 have also revealed diverse SAR11 lineages at high abundance in the deserts of the meso- and bathypelagic realms6–9. However, the functional properties that distinguish SAR11 bacteria living in distinct ocean regions remain unclear. All known SAR11 genomes are small (typically less than 1.5 megabase pairs (Mb)), with genomic streamlining as a potential adaptation to the nutrient-limiting conditions of the open ocean11. It has been hypothesized that adaptations in SAR11 do not involve large variations in gene content6,8, suggesting that the contribution of SAR11 to ocean biogeochemistry is primarily through its role in aerobic oxidation of organic carbon.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Diverse SAR11 SAGs from anoxic waters

Samples for SAG analysis were obtained from two depths in the anoxic zone: at 125 m at the NO$_3^-$ maximum (6 μM), and at 300 m in the core of the NO$_3^-$ reduction zone. Single prokaryotic cells were isolated by fluorescence-activated cell sorting, subjected to genome amplification, and screened by 16S rRNA gene fragment (470 bp) polymerase chain reaction (PCR) and Sanger sequencing. From this screen, 23% and 32% of SAGs from 125 m and 300 m, respectively, were confidently assigned to the SAR11 family Pelagibacteraeaceae (Fig. 1b), thus confirming the substantial numerical abundance of SAR11 in the OMZ. From this SAR11 subset, 10 SAGs from 125 m and 12 SAGs from 300 m were randomly selected for shotgun sequencing (Illumina), along with 5 technical control SAR11 SAGs from the oxic surface waters of the Gulf of Mexico (GoM). After sequencing, quality filtering and assembly, a total of 19 SAGs were used for analysis: 15 OMZ SAGs (5 from 125 m, 10 from 300 m) and 4 GoM control SAGs (Supplementary Table 1). These genomes exhibited varying levels of completeness (~2–90%; average 30%) and no detectable contamination (Extended Data Fig. 1), as assessed by the presence of single-copy housekeeping genes, 16S rRNA gene identities, and the taxonomic assignment of SAG contigs (Supplementary Tables 1, 2 and Supplementary Discussion).

The identified SAGs represented a diverse and novel SAR11 community in the OMZ. Phylogenetic reconstructions based on either 16S rRNA genes or single-copy housekeeping proteins placed the 19 SAGs in 5 subclades of SAR11 (Fig. 2a). Average amino acid identity (AAI) comparisons among all available SAR11 genomes (Supplementary Table 3) further corroborated this classification, placing: (1) seven OMZ SAGs within the previously characterized deep-branching monophyletic group of subclade IIa (hereafter designated subclade IIa.A), distinct (>5% 16S divergence) from SAG H1M058 from the tropical North Pacific (hereafter designated subclade IIa.B); (2) three OMZ SAGs within the deep-branching subclade IIb; (3) two OMZ SAGs within subclade Ic, which includes recently described SAGs from the bathypelagic ocean; (4) four OMZ and all four GoM surface SAGs within subclade Ib, which thus far lacks genome representatives; and (5) OMZ SAG A7 most closely related to H1M59, a member of the divergent SAR11 subclade V*.8,30,31 Note that the exact placement of subclade V in the SAR11 phylogeny is unstable depending on the marker gene and outgroup used.8,32,33 The average estimated genome size of OMZ SAGs was 1.33 Mb (Supplementary Table 1), consistent with prior reports of genome streamlining in SAR11.

OMZ SAR11 abundance peaks under oxygen depletion

To estimate the in situ abundance and activity of OMZ SAR11, metagenome and metatranscriptome reads from OMZ sites and from diverse oxic ocean regions (Supplementary Table 4) were recruited to 39 available SAR11 genomes (Supplementary Table 1). Metagenomic read recruitment, performed essentially as described previously, showed that each OMZ SAR11 subclade represents a sequence-discrete (and hence tractable) population (Supplementary Discussion), but with each population encompassing substantial intra-population variation (~92–100% average nucleotide identity between members of the population versus <90% between populations), as well as gene content variability (Extended Data Fig. 2). We therefore estimated SAR11 abundance at the subclade level, based on the average coverage of 507 genes shared between genomes from all SAR11 subclades. On the basis of this analysis, SAR11 subclades Ic, IIa.A and IIb together comprised about 10–30% of the bacterial community in ETNP and ETSP metagenomes and metatranscriptomes from depths with undetectable O$_2$ (Fig. 2b, c), consistent with the high abundance of SAR11 in the pool of cells sorted for SAG analysis (Fig. 1b). Subclade IIa.A, composed exclusively of seven SAGs from this study, was particularly abundant, making up to 15% of the community in anoxic samples. All OMZ subclades were absent from or much less abundant (<5%) in metagenomes from oxic sites, including those from above the ETNP OMZ (Fig. 2b). Together, these results identify newly described SAR11 subclades whose distribution is linked to an oxygen-depleted niche.

Metabolic adaptations to low oxygen in SAR11 genomes

OMZ and GoM SAGs were then analysed for evidence of microaerobic or anaerobic metabolism. Surprisingly, in 8 of the 15 OMZ SAGs, belonging to SAR11 subclades Ic, IIa.A, IIb and V, protein family-based classification detected genes encoding the respiratory Nar of the DMSO reductase superfamily (Fig. 2a). Evidence of a complete canonical nar operon (narGHJI)—encoding the α subunit that catalyses NO$_3^-$ reduction to NO$_2^-$ (NarG), the iron–sulfur-containing β subunit (NarH) that transfers electrons to the molybdenum cofactor of NarG, the transmembrane cytochrome b-like γ subunit (NarI) involved in electron transfer from membrane quinols to NarH, and the NarJ chaperone involved in enzyme formation—was found within a single assembled contig in four SAGs (A6, E4, D9, A7), while partial narG and narH fragments were identified in another four SAGs (Extended Data Fig. 3). In all SAR11 SAGs containing nar on a contig, we identified other genes upstream or downstream on the same contig taxonomically assigned to SAR11 reference genomes (Supplementary Table 5 and Supplementary Discussion), further confirming the association of nar with SAR11. Genes encoding the NO$_3^-$ / NO$_2^-$ transporter NarK and proteins for biosynthesis of the essential molybdenum cofactor (moeA, mobA) were also identified in eight and five of the SAGs, respectively (Supplementary Table 1). In only four of the fifteen OMZ SAGs were nar or cofactor synthesis genes not detected, presumably due to sequencing gaps (completeness of these SAGs: 4–20%; Supplementary Table 1). In contrast, these genes were not detected in any of the four control SAGs from the oxic GoM, despite high completeness of those
genomes (average 61%). Genes encoding for downstream steps of denitrification or other dissimilatory anaerobic metabolisms were not found in any of the SAGs. However, in contrast to all previously analysed SAR11 genomes, three of the OMZ SAGs, all from subclade Ia.A, also contained genes encoding high-affinity O₂- using bd-type terminal oxidases (Supplementary Table 1). Compared with the coxl-type oxidases present in all known SAR11 genomes, including the OMZ SAGs analysed here, bd-type oxidases have a much higher affinity for O₂ (3–8 mM; Supplementary Discussion), suggesting a potential for microaerobic respiration by OMZ SAR11. These results provide the first indication of adaptation to low oxygen in SAR11 and the ability to respire NO₃⁻ to NO₂⁻ in the absence of oxygen, consistent with the distribution of these bacteria in the OMZ water column.

**Multiple divergent Nar proteins in OMZ SAR11**

Phylogenetic placement of all identified narG and narH genes and partial fragments revealed two divergent nar variants in OMZ SAGs (Fig. 3a and Extended Data Fig. 3): (1) an ‘OP1 type’ in which all four nar genes and an upstream cytochrome c protein were most similar (56–78% amino acid identity) to homologues from ‘Candidatus Acetothermus autotrophicum’ (Supplementary Table 5), a putative anaerobic acetogen of the candidate bacterial phylum OP1 (ref. 35); and (2) a ‘Gamma-type’ variant most similar (51–78% identity) to Nar from a denitrifying Gammaproteobacteria endosymbiont (Ca. Vesicomyosocius okutanii strain HA)36. At least two of the OMZ SAR11 SAGs from subclade Ia.A, as well as SAG A7 from subclade V, encoded both OP1- and Gamma-type nar variants, suggesting that divergent nar copies (~42% amino acid identity) co-occur in the same genome (Supplementary Discussion). Multiple nar operons per genome have been reported for diverse bacteria and are hypothesized to be related to adaptation to different oxygen conditions, with one variant constitutively expressed at low baseline levels37–39. For both OP1- and Gamma-type variants, the sequence divergence among recovered sequences was consistent with the phylogenetic placement of the SAGs. For example, OP1-type narG fragments represented three distinct 97% amino acid identity clusters (Fig. 2a). Sequences from clade Ia.A SAGs fell within the same cluster, sharing ~96.5% identity with sequences of the closely related Ic and Ib subclades, and ~90% with sequences from the more distant A7 SAG (Extended Data Fig. 3). This pattern suggests diversification of nar operons in parallel with its genomic background, and also confirms that these sequences are not a systemic contaminant (Supplementary Discussion).

**Biochemical characterization of SAR11 Nar**

We sought to characterize further the biochemical function of SAR11 nar genes. Phylogenetic reconstruction based on 392 proteins of the diverse DMSO superfamily revealed that both OP1- and Gamma-type Nar fall within the clade of membrane-bound cytoplasm-orientated Nar and NO₂⁻ oxidoreductases (Nxr), and were most closely related to Nar from known NO₃⁻-reducing bacteria (Fig. 3a)40. The lack of a TAT peptide motif at the N terminus corroborated the probable cytoplasmic orientation of the NarG active site41, similar to experimentally verified Nar in *Escherichia coli*42. Additionally, the identified NarG sequences contain diagnostic functional domains found in NarG but not in other oxidoreductases of the DMSO reductase superfamily (Extended Data Fig. 4)40.

To verify NO₃⁻ reduction potential in SAR11, we introduced full-length SAR11 nar operons into a NO₃⁻ reductase-deficient *E. coli* mutant and tested for enzyme activity. The Gamma-type nar operon was successfully expressed in *E. coli*, yielding Nar proteins of the predicted size range and enabling growth of the mutant under anaerobic conditions in the presence of NO₃⁻, coupled with simultaneous NO₃⁻ reduction to NO₂⁻ (Extended Data Fig. 5), thereby providing direct evidence for the function of this enzyme in vivo. The OP1-type operon did not reverse the *E. coli* mutant phenotype, presumably due to the much greater divergence of this variant from the *E. coli* nar operon. Given the high similarity of Nar and Nxr protein sequences43–45, and the reversibility of the NO₃⁻ reduction reaction, it is possible that either or both OP1- and Gamma-type proteins could also function in situ to oxidize NO₂⁻ aerobically. Although it is enticing, this possibility is
reads were then classified within a reference phylogeny containing \( \text{narG} \) using a similarity search-trained model that discriminates \( \text{NO}_3^- \) within the OMZ to evaluate the contribution of SAR11 cells to \( \text{NO}_3^- \). SAR11 genes encode functional \( \text{narG} \) suggesting that the identified SAR11 operons are found in SAR11 genomes within the OMZ. First, while our SAG collection captured only a fraction of total \( \text{nar} \) diversity, additional \( \text{nar} \) operons were identified in metagenomic contigs classified as SAR11 (Extended Data Figs 3, 7 and Supplementary Table 6). Second, the majority of the metagenomic \( \text{narG} \) reads showed \( >95\% \) nucleotide identity with the \( \text{narG} \) genes encoded by the SAGs, suggesting that SAR11 cells are among the major contributors of Nar enzymes in the OMZ (Fig. 2b).

**Metatranscriptome sequencing**

SAR11-affiliated \( \text{nar} \) genes are transcribed in the OMZ. The abundance of both OP1- and Gamma-type variants in ETNP metatranscriptomes increased steadily from the lower oxycline (85 m) to the OMZ core (300 m), directly paralleling the abundance of the respective genes and the depth trend in \( \text{NO}_3^- \) reduction rates (Fig. 1c). Notably, within the ETNP OMZ, an average of 39% of all \( \text{narG} \) transcripts shared \( >95\% \) nucleotide identity with the OP1- or Gamma-type sequences detected in SAR11 SAGs (Fig. 3c), a conservative lower-bound estimate of the contribution of SAR11 bacteria to the total \( \text{nar} \) transcripts within the OMZ. Accordingly, within the anoxic OMZ depths, \( \text{nar} \) genes are among the most transcriptionally active genes in the SAG genomes (Extended Data Fig. 8). The high transcriptional activity of SAR11 \( \text{nar} \) operons, interpreted alongside their distribution relative to \( \text{NO}_3^- \) reduction rates, suggests that SAR11 bacteria contribute substantially to community \( \text{NO}_3^- \) respiration.

**Conclusions**

Collectively, our findings identify diverse and abundant SAR11 lineages whose genome content and environmental distribution reflect adaptation to an anoxic niche, unlike all other SAR11 bacteria characterized to date. The experimentally verified \( \text{NO}_3^- \) reductase activity in the Gamma-type SAR11 \( \text{nar} \) variant, along with the high expression levels of divergent SAR11 \( \text{nar} \) genes in the functionally anoxic core of the OMZ, suggest that persistence in this niche is linked to \( \text{NO}_3^- \) respiration, consistent with the fundamental importance of this process in OMZs. Nitrate respiration in OMZs constitutes the primary mode for organic carbon mineralization and the main production route of \( \text{NO}_2^- \), a critical substrate for the major nitrogen loss processes of anammox and denitrification. The presence and activity of \( \text{nar} \) operons in SAR11, as well as the high abundance of \( \text{nar} \)-associated SAR11 clades in the OMZ, implicate these versatile organisms as major contributors to the initiation of OMZ nitrogen loss. Together, these findings redefine the ecological niche of one of the planet’s most dominant groups of organisms, providing a set of genomic references to establish SAR11 as a model for studies of nitrogen and carbon cycling in OMZs.
METODS

Collection of ETNP and GoM samples for SAG analysis. No statistical methods were used to predetermine sample size. Selection of the SAR11 SAGs was randomized. The investigators were not blinded to allocation during experiments and outcome assessment. Samples were collected from the ETNP OMZ during the Oxygen Minimum Zone Microbial Biogeochemistry Expedition (OMZoMBiE) cruise (RV Horizon, 13–28 June 2013). Sea water for single-cell sorting and single-cell transcriptomics were collected during the same cruise. Analyses were performed on five stations (Supplementary Table 1) and 27 SAR11 SAGs were recovered. Libraries were prepared and sequenced in two lanes on an Illumina HiSeq (150 bp paired reads).

ETNP OMZ rate measurements, and oxygen and nutrient analysis. Samples for oxygen and nutrient measurements were collected on the same date and casts as those for single-cell sorting described above. Samples for rate measurements and metagenomics/transcriptomics (below) were collected a few hours later on the same day. Detailed collection and analysis procedures for those samples have been previously described. Briefly, oxygen concentrations were determined using a rosette-mounted sensors, including a SBE43 dissolved oxygen sensor for microsensor sensitivity and a high-resolution switchable trace amount oxygen (STOX) sensor for nanomolar-level measurements. CTD and oxygen measurements (SBE43) from three casts spanning this sampling period revealed no detectable movement in the oxycline, indicating stability in water column conditions.

Metagenome and metatranscriptome samples. Metadata, sequencing statistics, and accession numbers of all analysed metagenome and metatranscriptome data sets are in Supplementary Table 4. Here, we summarize the OMZ and GoM data sets at the core of our analysis. ETNP OMZ metatranscriptomic and metagenomic data sets were generated via MiSeq Illumina sequencing as described in ref. 19 and ref. 47, respectively, for 5 depths at station 6: the upper oxycline (30 m), lower oxycline (85 m), secondary chlorophyll maximum (100 m), secondary nitrite maximum (125 m) and OMZ core (300 m) (Supplementary Table 4). Metagenome data sets from the ETSP were generated by Roche 454 pyrosequencing as previously described for 4 depths at an OMZ site (20° 05′S, 70° 48′W) off the coast of Iquique, Chile: the suboxic (<10 m) upper OMZ just below the oxycline (70 m), the anoxic OMZ core (110 m, 200 m), and the oxic zone below the OMZ (1000 m). The ETNP and ETSP data sets analysed here reflect the 0.2–1.6 μm biomass size fraction; this fraction was shown to contain the vast majority of bacterioplankton and SAR11 cells. We also included two additional metagenomes, sampled on 5 May 2014 from the same site (station 6) in the ETNP, in order to obtain full-length metagenome sequence data for the purpose of screen sampling. These metagenomes were obtained from depths of 68 m within the oxycline and 120 m within the OMZ. For the 9 GoM metagenomes released with this study, samples were collected from Niskin bottles (60 l per depth), and filtered on board using the same filtration systems as for the ETNP and ETSP metagenomes (0.2–1.6 μm fraction). DNA was extracted with the same protocol as for the OMZ samples and libraries were prepared and sequenced in two lanes on an Illumina HiSeq (150 bp paired reads).

All metagenomic and metatranscriptomic data sets were quality trimmed as described below for the SAG samples. The metatranscriptomic data sets were further filtered to remove rRNA transcripts using the SortMeRNA algorithm. Four-hundred and fifty-four metagenomic data sets were filtered to remove duplicate sequences. The quality trimmed reads from the OMZ metagenomes (ETNP and ETSP), were assembled with IDBA49 and genes were predicted on scaffolds longer than 500 bp with MetaGeneMark.hmm50 and 16S rRNA gene sequences were detected in only one SAG, SAG A2 from the GoM, as multiplicity of divergent and nearly full-length marker genes. This SAG was excluded from further analysis. For the final data set of 19 SAGs, coding sequences were predicted on scaffolds longer than 500 bp with MetaGenomeMark.hmm50 and 16S rRNA gene sequences were identified using RNAmmer51. 16S rRNA sequences were identified in the assemblies (4/4 GoM SAGs, and 8/15 OMZ SAGs) were compared to the 470 bp 16S fragment obtained during the initial SAG sequencing and confirmed to be identical. As an additional quality control step, all predicted genes from the 19 SAGs were taxonomically annotated using MyTaxa48 and the taxonomic distributions of adjacent genes in the concatenated assembly (10 gene windows) were inspected for possible contamination. As discussed in Supplementary Discussion, a contaminant genome in the assembled contigs can be visualized in the MyTaxa scan plots (Extended Data Fig. 1).

Predicted genes were functionally annotated using the blast2go pipeline for assignment to metabolic pathways, and screened manually for evidence of anaerobic energy metabolism. Detected genes of anaerobic metabolism, including nitrate reductase (nar) genes, as well as terminal oxidase genes and the single-copy marker gene rpoS, were further verified using HMMER3 (http://hmmer.janelia.org/) with default settings and recommended cutoffs for a match against available Pfam models. Statistics of SAG quality control, assemblies, and contamination testing are in Supplementary Table 1 and 2.

Phylogenetic placement of SAGs. The evolutionary relatedness of SAR11 SAGs was assessed using the identified full or nearly full-length 16S RNA gene sequences from the assembled SAGs. For the SAGs from which no full-length 16S RNA fragments were assembled, the shorter fragments obtained during screening were used in pairwise comparisons with full-length sequence references (Supplementary Table 3, 16S matrix). The 16S RNA sequences from publicly available SAR11 genomes, as well as previously published 16S sequences13,14 from subclades with no genome representatives, were included in the alignment to aid in the classification of the SAR11 subclades. Additionally, genome representatives of divergent alphaproteobacteria classes, as well as the beta- and gammaproteobacteria class were included to facilitate the rooting of the tree. Maximum likelihood phylogenetic reconstruction was performed with RAXML with 1,000 bootstraps and the GTR model for nucleotides.

Additionally, hidden Markov models (HMMs) of 106 housekeeping genes found in single copy in bacterial genomes were used to identify marker genes in available SAGs and reference genomes using HMMER3 (http://hmmer.janelia.org/). The identified marker genes (Supplementary Table 1) were aligned using Clustal Omega and the protein alignments concatenated using Aln.cat.rb from the enve-omics collection (http://enve-omics.ce.gatech.edu/) to remove invariable sites and maintain protein coordinates. The concatenated alignment was used to build a maximum likelihood phylogeny with RAxML, using 1,000 bootstraps, and the PROTGAMMAUTO function, which identifies the best amino acid substitution model for each protein. SAGs where assigned to SAR11 subclades based on...
on the consensus categorization of both 16S rRNA and marker gene phylogenies, in accordance with previously published subclade identification sequences6-13. OMZ-sourced SAR11 SAGs from the SAR11 IIa lineage were further categorized as subclade IIa.A, to differentiate them from the currently available reference SAR11 IIa representative (HMB058), classified here as subclade IIa.B. Average amino acid identities (AAs) were estimated as described previously68.

**narG functional gene validation and phylogeny.** Reference nitrate reductase and nitrite oxidoreductase protein sequences (n = 697) representing divergent bacterial and archaeal phyotypes was downloaded from UniProtSwiss-Prot69. Together with representatives of other DMSO family oxidoreductases (n = 71), as using a guide the reference tree from ref. 64. From this 697-sequence set, 321 full-length NarG/NxrA sequences were selected to represent all the clades, along with the 71 additional non-NarG/NxrA proteins. The NarG/NxrA subset included the closest relatives to the SAG OP1 and Gamma-type Nar variants, as determined by BLAST. All protein sequences (n = 392), including the full-length NarG identified in the SAGs, were aligned with Clustal Omega, and a maximum likelihood phylogeny was reconstructed with RAxML with 1,000 bootstrap replicas and the PROGRAMMAAUTO model. Partial fragments of the NarG protein were then aligned to the alignment using MAFFT’s ‘addfragments’70, and the evolutionary placement algorithm (EPA) implemented in RAxML was used to place them within the reference tree66. The same procedure was followed for the phylogenetic reconstruction and placement of identified NarH protein sequences.

Quantification of **narG**-encoding reads from the metagenomes and metatranscriptomes was done using BLAST searches against a manually curated NarG database and the software ROCker (L. H. Orellana, L. M. Rodriguez-R and K. T. Konstantinidis, manuscript submitted). Using receiver-operator curve (ROC) analysis, ROCker identifies the most discriminant BLAST bit-score per position in a reference alignment (NarG database) given a certain read length by simulating in silico metagenomic data sets that include the reference genes. This strategy permits the accurate estimation of abundance of target genes in short-read data sets, minimizing false negatives and positives derived from closely related proteins or conserved domains, a critical challenge in the detection of NarG due to the ubiquity of other closely related DMSO oxidoreductases. The NarG database was manually curated and confirmed by the phylogenetic reconstruction of all available nitrate reductase and nitrite oxidoreductase sequences and visual inspection of the multi-sequence alignment for conservation of known functional domains and motifs. The final NarG database consisted of 697 nitrate reductases/nitrite oxidoreductases (positive set) and 71 representative non-NarG/NxrA DMSO family proteins (negative set for identification of false positive BLAST matches). All data sets, as well as the ROCker models built for narG quantifications in metagenomes with different read lengths, are available at http://emvice-omics.cs.gatech.edu/rocker/.

Additionally, the model for the identification of rpcB fragments in metagenomes was used to estimate coverage of rpcB in metagenomes.

The abundance of **narG** sequences in met-omic data sets was estimated as genomic equivalents for each sample, by normalizing the coverage of **narG** for the gene length (reads per nucleotide of **narG**) and dividing the normalized value by the rpcB-normalized coverage (reads per nucleotide of rpcB) as shown in Supplementary Table 4. To quantify the abundance of the **narG** variants (OP1, Gamma-type), protein fragments were predicted in all identified (from ROCker) **narG** reads using FragGeneScan77 and placed in the reference DMSO tree using RAxML-EPA. The abundances of the OP1-type or Gamma-type variants were estimated based on the number of reads that were placed in the terminal or internal nodes of the aforementioned clades on the reference tree, using JPlace.to_iToL.rb and the recruitment plots were shown in Extended Data Fig. 8b. BLASTP was used to map the reads against the reference NarG sequences, and the recruitment plots were constructed using BlastTab (casbi) pl and BlastTab.rec.plot.R scripts from the emvice-omics collection.

Thus, the reported abundances of OP1 and Gamma-type **narG** in metagenomes/metatranscriptomes are based on phylogenetic assignment of **nar** reads, rather than a strict sequence similarity cutoff. To estimate a lower limit for the abundance of **narG** sequences presumably encoded by SAR11 genomes, the number of reads with more than 95% nucleotide identity to the reference **narG** sequences found in the SAGs was estimated, and shown in Extended Data Fig. 6b, c. The figure shows abundance estimates for reads that are phylogenetically assigned to OP1 and Gamma nodes, with partitioning of the data into reads that share less than and more than 95% nucleotide identity with the SAR11 OP1 and Gamma-type references.

**NarG divergence in reference closed genomes.** Identification of **narG** in all closed genomes available from GOLD (27,461 bacterial and 685 archaeal genomes)68 was performed using HMMER3 with default settings. The results were further refined by a competitive BLAST search69 against the custom-made **narG** reference database (used for ROCker), which included DMSO family oxidoreductase enzyme reference sequences. Matches with best hit against **narG** sequences and a bit score higher than 900 were annotated as nitrate reductases or nitrite oxidoreductases. When found in multiple copies (up to 6), a reciprocal BLASTP search was performed to estimate sequence divergence, measured as amino acid identity. Quantification of **SAR11** clades in metagenomes and metatranscriptomes. For each metagenome/metatranscriptome, reads potentially derived from **SAR11** genomes were identified by a competitive BLAST best-match approach. Each read was aligned against the entire *SAR11* genome collection (NarG+*, ncBi-fip (2638 bacterial, 165 archaeal) and 39 genome representatives of the SAR11 lineage, including 20 published isolate or SAG sequences and the 19 SAG sequences produced in this study (Supplementary Table 1). Metagenomic and metatranscriptomic reads (predicted ORFs with FragGeneScan) were then compared against the database using BLASTP, and the subset of reads with a best match against any of the **SAR11** genomes and an e value < 0.001 was classified as **SAR11** reads (Supplementary Table 4). To quantify the relative abundance of distinct **SAR11** subclades, the **SAR11** reads were further classified as follows. We used the coverage of marker genes that could be found in all the subclades to more accurately estimate the abundance of distinct subclades and overcome both the biased representation of **SAR11** subclades in the available genomes, and the partial nature of **SAR** genomes. For all 39 available **SAR11** genomes, 5,707 orthologous genes (OGs) were identified by reciprocal best match and Markov clustering with inflation l.5 using ogs.mcl.rb from the emvice-omics collection. From the identified OGs, 507 were represented at least once in each of the 8 subclades. All metagenomic and metatranscriptomic reads (**SAR11** subsets) were mapped against the database containing all protein sequences from the 507 OGs (which were tagged according to subclade of origin) using the BLASTX option from Diamond69 and only the best matches for each read were kept. The coverage of each OG for each subclade was estimated based on that competitive best match result, normalized for the gene length (reads per bp of each OG), and the average coverage of all 507 OGs was used to estimate the abundance of subclades. Additionally, the number of rpcB reads for each metagenome was identified (for either the total data set or the subset of the **SAR11** reads), and the coverage of rpcB was used as a normalization factor to estimate the abundance of **SAR11** subclades over the total bacterial community.

**Functional characterization of **SAR11** nar operons.** A previously constructed NO3−-reductase deficient *Escherichia coli* strain71 was used as the genetic system for heterologous expression of **SAR11** nar genes. We used whole-genome sequencing (Illumina MiSeq) to confirm that this strain lacked all three NO3−-reductases (**ΔnarG1ΔnarZΔnapAB narZ::Δ**; Extended Data Fig. 5). The phenotype of this strain, hereafter referred to as the triple mutant, was verified by a lack of NO3−-production and an absence of growth with NO3− under anaerobic conditions, compared to the wild-type MC1400 *E. coli* strain (Extended Data Fig. 5).

Complete sequences from one OP1-type and one Gamma-type **nar** operon, containing upstream and downstream sequences, were identified from the ETNP 300 m and ETNP 120 m metagenomes (see above). These sequences were confirmed to be identical to the operons in SAG A7 (which was lacking part of the N terminus of the **narG** gene; Extended Data Fig. 3). Purified DNA from the ETNP 300 m and ETNP 120 m metagenomes was used as template for PCR amplification. In addition, we used genomic DNA from *E. coli* strain K12 MG1655 as a positive control. Because metagenomic samples are usually fragmented and the entire **nar** operon is 6.9 kb, primers were designed to amplify the OP1-type, Gamma-type and *E. coli* wild-type operon in two blocks. The first block spanned from the native **NarG** ribosome binding site to the end of the **narG** gene, and the second block included the end of the **narG** gene to the *narl* stop codon. The resulting PCR products were gel purified, assembled and cloned into pBBA1 K, a low-copy vector including the IPTG-inducible pTrc promoter72 by In-Fusion cloning (Clontech, Mountain View, CA). The cloning reactions were transformed into TOP10 cells, cultured in LB plates containing pBPA (Promega, Madison, WI) and sequenced by Pachio sequencing services (Menlo Park, CA). The final **nar** operons were identical (OP1 operon, and **NarG** proteins of Gamma operon) or nearly identical with silent substitutions (99% and 98% AAI for the Gamma-type **NarG** and **H** proteins) compared to the sequences from SAG A7 (GenBank accession KX275213, KX275214). Correct clones were isolated for each operon type, and purified plasmid was used to electroporate the triple mutant *E. coli* strains described above to generate recombinant strains expressing the heterologous **nar** operons for functional characterization.

For anaerobic cultures performing NO3− respiration, strains were first induced in LB medium with 0.5 mM IPTG for 5 h, and 20 μl of inoculum was subsequently transferred to the NaNO3-feeding chamber in a tight tube over a nitrogen atmosphere. The medium was prepared as previously described73, from composed potassium phosphate buffer (100 mM, pH 7.4), 15 mM (NH4)2SO4, 9 mM NaCl, 2 mM MgSO4, 5 μM Na2MoO4, 10 μM MoH+ salt, 100 μM CaCl2, 0.5% casaminoacids and 0.01% thiamine. Glyceral (40 mM) was used as the sole carbon source, and NO3− was added at 30 mM. IPTG (0.5 mM), kanamycin (30 μg/ml) and streptomycin (30 μg/ml) were used with the...
recombinant strains. Samples for NO$_3^-$ and NO$_2^-$ concentrations were obtained at regular time intervals during incubations, filtered through 0.2 μm porosity filters and injected into a Dionex DX ion chromatography unit with the Dionex IonPac AS14A analytical column. Growth in incubations was assessed as optical density (OD$_{600nm}$). Growth curve data from replicated cultures (triplicate) were fitted to a logistic model with variables $r$ (specific growth rate), $P_0$ (initial population) and $K$ (carrying capacity), using nonlinear least-squares estimates and prediction of OD per time point with confidence intervals as implemented in env.growthcurve from the geo-omics collection (http://enve-omics.ce.gatech.edu/).

Nitrite reductase activity was further verified in cell lysates from cells grown anaerobically for 12 days. Cells resuspended in 100 mM sodium phosphate buffer (pH 7.2) containing 0.02% Tween 80 were lysed by sonication in a Bioruptor UCD-200 (Diagonode). Protein concentration of the cell lysate was determined using a Qubit 2.0 fluorometer (Thermo Fisher Scientific) and 100 μg of protein was added to a reaction containing 100 mM NaNO$_3$ and benzyl viologen as electron donor. The reaction was bubbled with N$_2$ for 2 min before initiation with the addition of 50 μl of 30 mM sodium dithionite in 10 mM NaOH (final volume: 500 μl). Aliquots (50 μl) were removed at 20 min intervals and NO$_2^-$ concentration determined colorimetrically after the addition of 50 μl Griess reagent (prepared with equal volumes of 0.1% N-1-naphthylethenediamine dihydrochloride in water and 1% sulfinamide in 5% phosphoric acid). All assays were performed in triplicate. Finally, NO$_2^-$ production from NO$_3^-$ was further confirmed using whole cell assays with 8 replicate clones (per recombinant strain) grown aerobically on 96-well plates in 70 μl Luria–Bertani (LB) broth supplemented with 30 mM NO$_3^-$ and various IPTG concentrations. Nitrite production was identified via the Griess reaction as described above.

**Quantitative PCR of 16S rRNA genes and SAR11 nar variants.** Quantitative PCR (qPCR) was used to count OP1- and Gamma-type narG and total bacterial 16S rRNA gene copies. Seawater samples for qPCR were collected in 2014 from three sites in the ETNP, including station 6 from which the SAG samples were obtained.

Primers for narG PCR were designed based on alignments of narG sequences recovered from OMZ SAGs, targeting sites inclusive of all OMZ SAR11-1 affiliated nar variants and exclusive of narG from the closest database reference sequences. Primer selection resulted in the following: GammaF, 5′-GGCTGAAATTTTCTTCTCATACGGA-3′; and GammaR, 5′-AGTTCAATCCGTCACTTCTCTACAT-3′ amplifying a 401-nucleotide fragment of the Gamma-type narG; and OP1F, 5′-ACACATTAAAGAAGAAGAATTAGG-3′; and OP1R, 5′-TGGATTCGTTTTTCAACTACATTTC-3′ amplifying a 288-nucleotide fragment of the OP1-type narG. PCR reactions were performed with DNA template from the OMZ 300 m sample (station 6) and the oxic Gulf of Mexico as a negative control with the following conditions: incubation at 50 °C for 2 min, 95 °C for 10 min, followed by 40 cycles of denaturation at 95 °C (15 s) and annealing at 53°C (for OP1) and 54°C (for Gamma) (1 min each). Amplicons with the expected length were observed only in the OMZ sample and were purified and concentrated using the QIAquick PCR purification kit (Qiagen). Clone libraries were prepared with the TOPO TA cloning kit (Life Technologies) following the manufacturer’s protocol. Plasmid DNA was isolated using the PureLink Quick Plasmid Miniprep Kit (Life Technologies). E. coli strain DH5α was transformed with each clone set and pooled on LB plates with X-Gal and Kan. Ampicillin-resistant colonies were isolated with the PureLink Quick Plasmid Miniprep Kit (Life Technologies). Inserts were purified using the QIAquick Purification kit and sequenced on an Applied Biosystems 3730xl DNA Analyzer using BigDye Terminator v3.1 cycle Sanger sequencing (Life Technologies). Sequencing recovered 14 sequences generated using OP1 primers and 12 generated using Gamma primers. All OP1-like sequences were most closely related (via BLASTX against the NCBI-nr database) to narG of an uncultured *Acetotherma* bacterium OP1 (dbj|BALS57372.1), whereas all Gamma-like sequences were most closely related to the gammaproteobacterial endosymbiont of *Calyptogena okutani* (*Ca. Vescicomyosoccus okutani*; reclassification of OP1), consistent with the phylogenetic classification of the recovered sequences within each clone set shared on average 96% (OP1 set) and 93% (Gamma set) nucleotide identity, raising the possibility that our primer sets may not amplify all OP1 and Gamma-type narG variants in the community. We therefore consider our abundance estimates to be lower bounds. The OP1 and Gamma primer sets, along with universal bacterial 16S rRNA gene primers 1053F and 1392R, were used for SYBR Green-based qPCR. Tenfold serial dilutions of DNA from a plasmid carrying narG amplicons (described above) and a single copy of the 16S rRNA gene (from *Dehalococcoides mccartyi*) were included on each qPCR plate and used to generate standard curves, with a detection limit of ~30 and 10–15 gene copies/ml for 16S rRNA and narG variants, respectively. Assays were run on a 7500 Fast PCR System and a StepOnePlus Real-Time PCR System (Applied Biosystems). All samples were run in triplicate with conditions as follows: 2 min incubation at 50°C, followed by 10 min at 95°C followed by 40 cycles of denaturation at 95°C (15 s) and annealing at 60°C (1 min).
Extended Data Figure 1 | Evaluation of contamination based on MyTaxa taxonomic affiliations. a, Representative MyTaxa plots to test for contamination based on taxonomic affiliations of predicted genes. The MyTaxa algorithm\(^5\) predicts the taxonomic affiliation on the basis of a weighted classification scheme that takes into account the phylogenetic signal of each protein family. Each gene is assigned to the deepest taxonomic resolution (out of phylum, genus and species) for which a high-confidence value can be obtained (score 0.5). Each MyTaxa scan represents taxonomic distributions of all the predicted genes for one genome, given in windows of 10 genes, and sorted based on their position in the concatenated assembly of the genome (when a partial genome is used). a, b, White space in the histograms represents genes that could not be assigned to a given taxon due to (1) lack of BLASTP hits against the reference database (a collection of closed and draft genomes) or (2) lack of high confidence scores. Notice that for the representative OMZ SAG E5, more than 80% of the genes can be classified as Candidatus Pelagibacter (SAR11), with an additional 10% assigned to Proteobacteria. Note there are no genome representatives for this taxon (that is, SAR11 subclade IIa.A) in the database upon which MyTaxa is based. Similar results are obtained for the bathytype SAR11 SAG\(^6\), as this genome also lacks representatives. The closed genome from a coastal isolate HTCC1002 is shown for comparison to demonstrate a typical pattern for cases when close relatives of the query genome are available in the reference database, as is the case for this isolate. b, Taxonomic classifications of genes from the 19 SAGs analysed here. Each distribution was obtained from the MyTaxa scans performed for each SAG. The percentage of the total genes that could be taxonomically classified with MyTaxa was on average ~60%, and varied depending on the completeness of the genome (that is, partial genes are less likely to be assigned taxonomy with high confidence). These values are also reported in Supplementary Table 1. Of the genes that could be classified, the majority (>90%) were classified to SAR11 taxa.
Extended Data Figure 2 | Microdiversity within the SAR11 populations. a, Recruitment plot of metagenomic reads from the ETNP OMZ 300 m sample, against scaffolds from SAG E4. Notice that the recruited reads vary in identities from 100% down to 85%, indicating the presence of closely affiliated clades, as well as extensive microdiversity within the same clade (that is, reads sharing >95% identity). b, Phylogenetic reconstruction of reference RpoB protein sequences from SAR11 genomes, and placement of identified RpoB metagenomic sequences (denoted with the cross symbols). The alignment length was 1,406 columns with 5.9% gaps or undetermined sites. The presence of multiple divergent rpoB reads within the same subclade (predominantly for subclades IIa.A and Ic) suggests high abundance but also extensive microdiversity within those populations (rather than clonal populations).
Extended Data Figure 3 | nar genes encoded by SAR11 populations of OMZs. a, nar operon and adjacent genes identified in SAR11 SAGs from the ETNP OMZ, and in assemblies from the 85 m and 300 m ETNP OMZ metagenomes. narG sequences with at least 97% amino acid similarity are represented with the same colour. b, c, Representative maximum likelihood phylogeny to show sequence variation among full-length or near full-length narG (b) and narH (c) amino acid sequences identified in the SAGs. A subset of cytoplasm-oriented Nar and Nxr enzymes from publicly available genomes is also included. A comprehensive phylogeny showing the placement of SAR11 nar sequences relative to enzymes (n = 392) of the DMSO family is in Fig. 2a. Coloured pies represent the placement of shorter narG/narH gene fragments identified in the SAGs. Bootstrap values over 50 are shown. Outgroups (arrows) are E. coli dmsA (b) and dmsB (c). Note that the Gamma-type nar-containing contig recovered in E4 (Fig. 2a) contains narHII, but not narG; E4 Gamma-type is therefore not represented in Fig. 3b. All genes co-localized in the nar-containing contigs are listed in Supplementary Table 5. The p-numbers are gene identifiers given by the gene prediction software, consistent with those in Supplementary Table 5.
Extended Data Figure 4 | See next page for caption.
Extended Data Figure 4 | Identified NarG in SAR11 SAGs are members of the DMSO superfamily of oxidoreductases. a, Phylogenetic reconstruction of NarG and DMSO enzymes. The tree shown in Fig. 2 is presented here but has been expanded to include diverse DMSO oxidoreductases for direct comparison with the NarG/NxrA enzymes. Notice that both OP1 (green, blue, grey) and Gamma-type (red, orange) variants cluster within the cytoplasmically oriented Nar and Nxr enzymes. Six-hundred and ninety-seven NarG/NxrA proteins were identified from UniRef63, and from those, 321 full-length sequences were selected to represent all the diverse clades. An additional 71 non-NarG/NxrA proteins, representative of the diverse enzymes of the DMSO superfamily were also included in the collection. The full-length amino acid sequences were aligned with Clustal Omega64 and the phylogenetic tree was constructed by maximum likelihood and 1,000 bootstraps using RAxML60. The alignment length was 1,803 columns, out of which 31.2% were gaps or undetermined. Partial NarG sequences identified in the SAGs were placed on the tree using the EPA algorithm from RAxML66. The same collection of proteins was used to train the Rocker models and quantify the narG metagenomic fragments, and can be found in the enve-omics website (http://enve-omics.cc.gatech.edu/rocker/models). b, Alignment of NarG sequences from OMZ SAR11 with representative sequences from the DMSO superfamily of oxidoreductases. The protein motifs in the second and third panels are present in all functional Nar enzymes (NarG) and Nxr enzymes (NxrA) but not in closely related enzymes of the DMSO superfamily. The first panel shows the presence/absence of the TAT signal peptide (SRRSFLK), whose presence typically denotes a protein excreted to the outer membrane40,41. SAR11 NarG is instead oriented towards the cytoplasm (lack of TAT). The second panel shows the cysteine-rich motif typically found in the N terminus of the type-II DMSO superfamily oxidoreductases75 and believed to enable the formation of a [4Fe–4S] cluster in these proteins76. The Asn in position 158 of the alignment is typically found in catalytic subunits of nitrite reductases and DMSO oxidoreductases (DmsA) but not in other DMSO family enzymes. The third panel shows the Gln(Q) and Thr(T) in positions 398 and 399 within the putative substrate entry channel of the protein, which differentiate the Nar proteins from all other oxidoreductases of the DMSO family40.
Extended Data Figure 5 | Functional characterization of the SAR11 nar operons in the E. coli heterologous expression system. a, Genotype of the E. coli triple mutant confirmed by whole-genome sequencing. The triple mutant lacks complete functional operons of all three NO₃⁻ reductase enzymes, and thus is incapable of NO₃⁻ reduction. b, Anaerobic growth of triple-mutant clones, complemented with the SAR11 nar operons. For each strain three independent clones were monitored, and data from the replicate growth curves were fitted into a logistic model. Shaded areas represent the 95% confidence intervals of optical density readings (OD₆₀₀nm) in the fitted logistic growth models. NO₃⁻ and NO₂⁻ were measured in parallel with ion chromatography. Note that the Gamma-type SAR11 operon complements the triple-mutant phenotype, growing anaerobically by reducing NO₃⁻ to NO₂⁻. E. coli encodes functional nitrite reductases, thus the accumulated NO₂⁻ can be further reduced to ammonia, accounting for the non-stoichiometric NO₂⁻ production. c, Whole-cell NO₂⁻ production assays under aerobic conditions. Eight independent clones (columns A–H) of each type (C1–C5) were inoculated in Luria–Bertani (LB) broth supplemented with 30 mM NO₃⁻ and different isopropyl-β-D-thiogalactoside (IPTG) concentrations, and the well plate was incubated for 2 days at room temperature. Griess reagent was added, and development of pink colour indicated NO₂⁻ production.
Extended Data Figure 6 | Relative abundance of narG variants in ETNP OMZ metagenomes and metatranscriptomes and various other ocean metagenomes. a, Relative abundance and diversity of NarG/NxrA enzymes as revealed by phylogenetic placement of identified narG metagenomic reads (coloured pies). All identified short metagenomic narG reads from various oceanic metagenomes were placed within a reconstructed reference NarG tree to estimate the abundance of the different narG variants. The results of the placement are presented in five separate trees, based on the origin of the analysed metagenomic reads (ETSP metagenomes, ETNP metagenomes and metatranscriptomes,oxic bathypelagic and oxic surface metagenomes) for clarity. In each of the five trees, the coloured pies represent the abundance (normalized for data set size) of the short metagenomic reads clustering in the respective node. Specifically, the pie radius reflects read abundance as a percentage of the total narG genome equivalents identified (that is, number of narG reads compared to number of rpoB reads, normalized for gene length and total number of reads in each metagenome), with the size of grey pies representing the highest and lowest relative abundance, respectively.

The reference tree is the same as in Fig. 3a. Scale bars represent substitutions per amino acid. Notice that the two narG variants affiliated with the SAR11 SAGs (highlighted in orange for the OP1 type and blue for the Gamma type) are only abundant in the metagenomes and metatranscriptomes from the OMZ, where they comprise more than 70% of the total narG read pool, as can also be observed in Fig. 3b and c. The number of narG reads of the OP1 or Gamma type are also given in Supplementary Table 1. b, qPCR-based abundance of SAR11-affiliated narG genes in the ETNP OMZ relative to NO₂⁻, NO₃⁻ and O₂ concentrations and qPCR-based counts of 16S rRNA. Counts of total bacterial 16S rRNA, OP1-type narG, and Gamma-type narG genes at three stations (map on legend) west of Manzanillo, Mexico in May 2014. Map was created with Ocean Data View (http://odv.awi.de). All assays were performed in triplicates, and the bars represent s.e.m. Note that counts of OP1- and Gamma-type narG variants are probably underestimates given the observed microdiversity in the community (Extended Data Figs 2 and 7), and therefore there is a possibility that our primers did not match all OP1- and Gamma-type variants.
Extended Data Figure 7 | Diversity of OP1 and Gamma-type narG amino acid sequences in the ETNP OMZ metagenome. a, Phylogenies showing all full-length narG sequences recovered in the ETNP OMZ metagenomes (85, 100, 125, 300 m), as well as those from the SAR11 SAGs and corresponding narG reference sequences, with the left tree showing OP1-type variants and the right tree showing Gamma-type variants. NarG sequences are colour-coded based on the taxonomic classification of adjacent genes in the same metagenomic scaffolds, as show in Supplementary Table 6. b, Recruitment of metagenomic reads (predicted open reading frames) from the OMZ 300 m sample, against OP1- (left) or Gamma- (right) type narG sequences from the SAR11 SAGs. The metagenomic reads used for recruitment were identified as ‘narG’ using the ROCker pipeline, and their identity was further confirmed by phylogenetic placement within the narG clade on a reference DMSO superfamily protein tree, to minimize non-specific recruitments in conserved protein regions. Note that based on this analysis, the OP1-type narG variants are highly diverse in the OMZ metagenome.
Extended Data Figure 8 | Transcriptional profile of predicted genes from the SAR11 OMZ SAG-D9. Transcriptomic reads with >99% identity matches were counted for each gene, and the counts were normalized for the data set size. Note that the nar operon genes are among the most actively transcribed in the ETNP 300 m OMZ sample.