The ISME Journal (2013) 7, 1962–1973

www.nature.com/ismej

ORIGINAL ARTICLE

Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling

This article has been corrected since Advance Online Publication and an erratum is also printed in this issue

Brett J Baker[1], Cody S Sheik[1], Chris A Taylor[2,4], Sunit Jain[1], Ashwini Bhasi[2], James D Cavalcoli[2] and Gregory J Dick[1,2,3]

[1]*Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, USA;* [2]*Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA and* [3]*Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA*

The deep ocean is an important component of global biogeochemical cycles because it contains one of the largest pools of reactive carbon and nitrogen on earth. However, the microbial communities that drive deep-sea geochemistry are vastly unexplored. Metatranscriptomics offers new windows into these communities, but it has been hampered by reliance on genome databases for interpretation. We reconstructed the transcriptomes of microbial populations from Guaymas Basin, in the deep Gulf of California, through shotgun sequencing and *de novo* assembly of total community RNA. Many of the resulting messenger RNA (mRNA) contiguous sequences contain multiple genes, reflecting co-transcription of operons, including those from dominant members. Also prevalent were transcripts with only limited representation (2.8 times coverage) in a corresponding metagenome, including a considerable portion (1.2 Mb total assembled mRNA sequence) with similarity (96%) to a marine heterotroph, *Alteromonas macleodii*. This *Alteromonas* and euryarchaeal marine group II populations displayed abundant transcripts from amino-acid transporters, suggesting recycling of organic carbon and nitrogen from amino acids. Also among the most abundant mRNAs were catalytic subunits of the nitrite oxidoreductase complex and electron transfer components involved in nitrite oxidation. These and other novel genes are related to novel *Nitrospirae* and have limited representation in accompanying metagenomic data. High throughput sequencing of 16S ribosomal RNA (rRNA) genes and rRNA read counts confirmed that *Nitrospirae* are minor yet widespread members of deep-sea communities. These results implicate a novel bacterial group in deep-sea nitrite oxidation, the second step of nitrification. This study highlights metatranscriptomic assembly as a valuable approach to study microbial communities.

The ISME Journal (2013) **7,** 1962–1973; doi:10.1038/ismej.2013.85; published online 23 May 2013
Subject Category: Integrated genomics and post-genomics approaches in microbial ecology
Keywords: Archaea; deep sea; transcriptomics; nitrification; *Alteromonas*; *Nitrospirae*

Introduction

Microorganisms mediate the marine carbon and nitrogen cycles, and thus control nutrient bioavailability, primary productivity, and production and consumption of greenhouse gases such as N_2O and CO_2 in the oceans (Ward *et al.*, 2007; Füssel *et al.*, 2012; Karl *et al.*, 2012). The deep ocean represents the largest active reservoir of carbon on the planet, containing ~50 times more inorganic carbon than the atmosphere (Raven and Falkowski, 1999). Thus, understanding the primary agents of carbon cycling in the deep sea is of considerable interest. The 'biological pump' has been considered a driving force for sequestration of carbon to the ocean interior (Raven and Falkowski, 1999) and the 'microbial carbon pump', in which heterotrophic bacteria generate recalcitrant dissolved organic carbon, represents a more recently recognized form of carbon sequestration (Jiao and Zheng, 2011). Therefore, elucidating the key microbial players that mediate interconversions between dissolved inorganic carbon, particulate organic carbon and dissolved organic carbon is crucial to understand the carbon cycle as it pertains to global change.

As nitrogen is often the co-limiting nutrient for productivity in the oceans, the carbon cycle is intimately linked to biogeochemical transformations of nitrogen (Zehr and Kudela, 2011). Recent

Correspondence: BJ Baker or GJ Dick, Department of Earth and Environmental Sciences, University of Michigan, CC Little Building, 1100 North University Avenue, Ann Arbor, MI 48109, USA.
E-mail: acidophile@gmail.com or gdick@umich.edu
[4]Current address: Compendia Bioscience, 110 Miller Avenue, Ann Arbor, MI 48104, USA.
Received 25 August 2012; revised 11 April 2013; accepted 22 April 2013; published online 23 May 2013

transformative advances in environmental DNA sequencing have revealed the pathways, organisms and genes involved in the nitrogen cycle including anaerobic ammonia oxidation (anammox; Strous *et al.*, 2006), denitrification (Ward *et al.*, 2007), N_2O production (Santoro *et al.*, 2011) and ammonia oxidation (Könneke *et al.*, 2005). Ammonia-oxidizing Archaea (AOA) are now recognized as major contributors to oceanic nitrification (Wuchter *et al.*, 2006) by catalyzing the first step, oxidation of ammonia to nitrite (Könneke *et al.*, 2005). These AOA are numerically abundant, especially in the deep sea, where they account for up to 40% of total cells (Karner *et al.*, 2001); thus, they have been estimated to be among the most abundant Archaea on earth (Pester *et al.*, 2011). Despite these new insights, fundamental questions about the marine nitrogen cycle remain open. For example, because nitrite produced by AOA typically does not accumulate in the environment (Dore and Karl, 1996), nitrite oxidation must be equally prevalent as AOA in nitrification (Ward *et al.*, 2007). Correlation between nitrite-oxidizing bacteria (NOB) and ammonia-oxidizing Archaea populations suggests metabolic coupling between these groups (Mincer *et al.*, 2007; Santoro *et al.*, 2010); yet NOB are observed at much lower abundance than their ammonia-oxidizing counterparts (Koops and Pommerening-Roser, 2001; Mincer *et al.*, 2007; Santoro *et al.*, 2010). This high AOA:NOB ratio is unexplained even when the greater free energy available from ammonia oxidation is taken into account; thus, the mechanisms and organisms responsible for nitrite removal remain unresolved (Ward *et al.,* 2007; Zehr and Kudela, 2011). Previously unrecognized nitrite reduction by AOA has recently been highlighted as another potential sink for nitrite (Santoro *et al.*, 2010; Baker *et al.*, 2012).

Metatranscriptomics is emerging as a valuable tool for tracking the metabolic activity of microbial communities as they occur in nature. Although the relationship between the abundance of RNA and protein is not simple, thus complicating efforts to use transcript abundance as a direct proxy for metabolic activity, metatranscriptomics still provides highly informative views of the interactions between microbes and their environments (Moran *et al.*, 2012). This approach offers the ability to sequence and quantify messenger RNA (mRNA) of specific genes and populations within an entire community, potentially including those that have not been previously identified. To date, analysis of metatranscriptomic sequence data has primarily relied on mapping of complementary DNA (cDNA) reads to genomic data sets derived from either public databases (Frias-Lopez *et al.*, 2008; Shi *et al.*, 2009; Stewart *et al.*, 2011) or from accompanying metagenomic sequencing (Shi *et al.*, 2011; Lesniewski *et al.*, 2012). These approaches are limited by reference data sets that lack the full diversity inherent to natural communities and by public databases that are biased toward readily cultured representatives. Thus, a large fraction of metatranscriptomic data is typically unclassified (Frias-Lopez *et al.*, 2008). Here, we attempt to resolve the metabolic activity of novel and minor community members through *de novo* assembly of metatranscriptomic sequence reads from a hydrothermal plume in Guaymas Basin, Gulf of California, where enhanced primary production is fueled by ammonia oxidation, methanotrophy and sulfur oxidation (Lesniewski *et al.*, 2012). Reconstruction of transcriptomes of deep-sea community members enabled identification of abundant transcripts involved in nitrite oxidation and carbon cycling from organisms with limited representation in metagenomic data sets.

Materials and methods

Sample collection and processing

Samples were collected in 10-l bottles by CTD-rosette (Sea-Bird, Bellevue, WA, USA) aboard the *R/V New Horizon* (Table 1), as described previously (Dick and Tebo, 2010). Briefly, samples were collected by 'tow-yo' of the CTD-rosette then immediately filtered onto 0.2 μm polycarbonate membranes with N_2 gas once on deck and preserved in RNAlater (Ambion, Grand Island, NY, USA). Although potential changes in the RNA pool during collection are a concern, as discussed previously (Lesniewski *et al.*, 2012), these changes are minimized by the fact that samples are kept under *in situ* conditions (cold, dark) throughout collection and immediately filtered and preserved once onboard. RNA was extracted from filters using the MirVana miRNA Isolation kit (Ambion) and treated with *DNAase* I, and concentrated and re-purified using RNeasy MinElut Kit (Qiagen, Valencia, CA, USA). RNA amplification by random priming and cDNA synthesis was performed as described previously (Shi *et al.*, 2009; Stewart *et al.*, 2011). Sequencing was performed on an Illumina HiSeq2000 instrument (San Diego, CA, USA) at the University of Michigan Sequencing Core.

Ribosomal RNA (rRNA)-based taxonomy abundance assessment

In order to assess the diversity of organisms present in the RNA data set, we mapped cDNA reads from the plume community to the SILVA small subunit rRNA gene database (Pruesse *et al.*, 2007) using riboPicker software (Schmieder *et al.*, 2012).

cDNA sequencing and analyses

cDNA reads were dereplicated by removing duplicated reads (100% match, identical length), then quality trimmed using Sickle (http://www.github.com/najoshi/sickle). For all read mapping we used trimmed and dereplicated data sets. Dereplication

Table 1 Summary of samples characteristics and sequencing results

Sample name	*Depth (m)*	*Temperature (°C)*	*O_2 μM*	*RNA ng l^{-1}* [a]	*Number of cDNA reads*	*Number of DNA reads*	*Average mRNA contig*
Plume-4 (GD-6)	1950	3.0	27.7	203/319	206,157,516		243 bp
Plume-1	1996	3.0	27.4	NA		576 187[b]	
Plume-2	1775	3.0	27	NA		563 818[b]	
Plume-3	1963	2.9	26.1	209	664 240[b]		
Bkgrd-1 (GD-7)	1600	3.0	28.5	64/69	244 519 176	358 335[b]	196 bp
Bkgrd-2	1600	2.6	46.5	67	406 533[b]	406 533[b]	

The 'Plume-4' sample was acquired at 27°30.360 111°20.818 (on 16 July 2004) and sample 'Bkgrd-1' was at 27°29.174 111°21.844 (on 13 July 2004).
Materials from all of the above samples were also utilized in this study.
[a]Total RNA concentrations shown are results of two independent extractions for each sample that were pooled.
[b]These samples were sequenced as part of a prior study (Lesniewski *et al.*, 2012).

reduced the number of reads from 206 to 45 million in the plume and 245 to 130 million in the background sample. We also mapped reads before dereplication to confirm that the general trends seen were not artifacts of preprocessing of the sequences. Reads were assembled with Velvet (1.2.01, https://github.com/dzerbino/velvet) and subsequently processed using the transcriptomic assembler Oases (0.2.04; Schulz *et al.*, 2012). Abundance of cDNA reads was determined by mapping all of the cDNA reads to the assembled transcripts fragments. Mapping was done using Burrows-Wheeler Aligner (Li and Durbin, 2009) with default settings (maximum mismatch = 4%). We manually checked the mRNA transcripts discussed in depth here for chimeras by viewing the read mapping in integrated genome viewer. The trends reported for the NOB and other low-abundance members were observed in that analysis as well. Assembled transcript contiguous sequences (contigs) were searched for functions using DOE JGI IMG/MER annotation pipeline (Markowitz *et al.*, 2012). The cDNA reads are available at NCBI SRA under accession numbers SRX134769 (plume) and SRX134768 (background). The assembled and annotated plume transcript library is available via IMG under taxon object ID 236347000. All comparisons of cDNA assemblies with metagenomic data was done with previously described data (Lesniewski *et al.* 2012), which was a co-assembly of reads from the same sample (Bkgrd-1), and additional ones (Plume-1 and -2, and Bkgrd-2).

Phylogenetic analyses

All phylogenetic trees were generated using maximum likelihood (RaxML) with ARB software (Ludwig *et al.*, 2004). rRNA-containing transcripts were identified using riboPicker package (Schmieder *et al.*, 2012). 16S rRNA sequences were aligned in Greengenes (DeSantis *et al.*, 2006). Alignments of mRNA sequences were done using CLUSTALW with manual refinement. In order to identify all of the 16S and 23S rRNA sequences in the transcript assembly, we first searched the plume assembly with *Candidatus* Nitrospira defluvii 16S and 23S rRNA genes. Matches were then imported and aligned to the Greengenes 16S rRNA and the Silva 23S rRNA databases (DeSantis *et al.,* 2006). We then generated large neighbor joining trees with thousands of reference sequences. Only contigs >350 bp were used in the 16S tree and only those >500 were used in the 23S rRNA tree. Only those sequences that fell within the *Nitrospirae* were kept. The 23S rRNA phylogeny was generated using 1909 characters. Group names in the 16S rRNA tree are based on those characterized by Lebedeva *et al.* (2011).

Identification of NOB and anammox transcripts

We searched all annotated genes on the Guaymas mRNA transcripts using all of the *Ca*. N. defluvii genes. We then compared these hits with the nonredundant NCBI protein database. Only those that had top hits to *Ca*. N. defluvii and *Leptospirillum* sp. were then considered to belong to the *Nitrospirae*. We searched the metatranscriptomic assembly for transcripts of key anammox genes, hydrazine oxidoreductase from the genome of *Ca.* Kuenenia stuttgartiensis and hydrazine hydrolase from *Ca.* Scalindua sp. (FM163627). Supplementary Figure S7 is based on the number of reads that match with e-value $<1E^{-5}$ (BLASTx). One gene transcript, 2236391221, had a match of 57% (bitscore of 56.6, e-value $7E^{-10}$). However, comparison of this transcript to Genbank revealed that it is most similar to several sequences obtained from microbes not thought to be capable of anammox, including *Shewanella woodyi* (79% similarity) and SUP05 (68% similarity).

Analyses of transcript sequence variants

cDNA reads were mapped to assembled contigs using Burrows-Wheeler Aligner mapping software (Li and Durbin, 2009). SNPs were identified by visually comparing reads mapped using integrated genome viewer.

16S rRNA gene pyrosequencing
DNA was extracted from one-fourth of a filter with the MoBio PowerSoil DNA isolation kit (MO-BIO, Carlsbad, CA, USA). In addition to bead beating, filters were incubated at 65 °C for 20 min to facilitate cellular lysis. Bead beating was performed using the MP-Bio FastPrep-24 (MP Biomedicals, Santa Ana, CA, USA) for 45 s at setting 6.5. 16S rRNA genes were amplified in triplicate 25 μl reactions containing the following (final concentration): 12.5 μl 5 Prime HotMasterMix (5 PRIME, Gaithersburg, MD, USA), 2 μl (15 μM) each forward and reverse primers, 1 μl community DNA. PCR thermocycler conditions were as follows: initial denaturation at 95 °C for 4 min, followed by 30 rounds of 95 °C for 30 s, 50 °C for 1 min, 72 °C for 1 min and final elongation of 72 °C for 10 min. Triplicate PCRs were combined and cleaned using a MoBio UltraClean PCR Clean-up kit. DNA was quantified using PicoGreen (Invitrogen, Carlsbad, CA, USA). Previously described 16S rRNA gene primers targeting the V4 region (515F/806R; Bates *et al.*, 2010) were used with reverse primers containing a 12-base barcode. Individual barcoded samples were combined into a single sample at equivalent concentrations, and then sent to Engencore (http://www.engencore.sc.edu) for pyrosequencing using Titanium chemistry (Engencore, University of South Carolina, Columbia, SC, USA). Sequences were error-corrected with Pyronoise (Quince *et al.*, 2009) implemented in Mothur (v 1.25.0; Schloss *et al.*, 2009). Species level operational taxonomic units (OTUs) were binned at 97% similarity and chimera checked using the OTU pipe (http://www.drive5.com/otupipe) command within QIIME (Caporaso *et al.*, 2010). Default parameters were used with the exception of less abundant OTUs being kept for downstream analysis. OTUs were taxonomically classified with BLASTn (Altschul *et al.*, 1990; ver 2.2.22, e-values cutoff 10^{-8}) using Greengenes taxonomy (available at http://www.qiime.wordpress.com) and customized to include NOB 16S rRNA sequences recovered from Guaymas Basin transcriptomic libraries.

Results and discussion

De novo assembly of transcripts
Random shotgun metatranscriptomic sequencing was conducted on a sample from the Guaymas Basin hydrothermal plume (1950 m water depth) and from a location just above the plume (1600 m), referred to as 'background', for comparison. *De novo* assembly of metatranscriptomic reads yielded 78 250 assembled contigs containing 81 452 predicted genes. 18 501 (23%) of these were putative protein-coding genes (non-transfer RNA or rRNA) of which 12 605 (68%) were assigned putative functions. The large number of non-protein coding transcripts can be explained by a high level of fragmentation (due to fine-scale variability in highly similar sequences) of rRNA genes that is commonly seen in short-read sequencing data (Miller *et al.,* 2011). Several of the mRNA contigs have homology to multiple genes of related function, reflecting assembly of co-transcribed genes from operons (Supplementary Figure S1). Among the most abundant were transcripts involved in oxidation of sulfur, ammonia and methane from dominant community members. Also, highly expressed were genes encoding ribosomal proteins (Supplementary Figure S2) from dominant groups, including AOA (Baker *et al.*, 2012), sulfur-oxidizing SUP05 *Gammaproteobacteria* and methanotrophs (Lesniewski *et al.*, 2012). These results are consistent with previous analyses of the same samples based on genome databases (Lesniewski *et al.,* 2012). The majority (53%) of the 16S rRNA containing reads belong to members of the *Gammaproteobacteria* (including methanotrophs and the sulfur-oxidizer, SUP05) (Figure 1). This is consistent with previous findings (Dick and Tebo, 2010; Lesniewski *et al.,* 2012) and their high coverage in genomic assemblies.

Heterotrophy
Metatranscriptomic assembly revealed abundant transcripts from community members that were not well represented in corresponding metagenomic data sets (Lesniewski *et al.,* 2012). The most abundant ribosomal protein-coding transcripts were highly similar (up to 99% DNA similarity) to a deep-sea heterotroph, *Alteromonas macleodii* (Ivars-Martínez *et al.*, 2008) (Supplementary Figure S2), which has limited representation in corresponding metagenomic data (averaging 2.8 times coverage) and 16S rRNA gene surveys (Dick and Tebo, 2010). The longest of these assembled transcripts is nearly 5 kb and contains an operon of 11 co-transcribed ribosomal protein

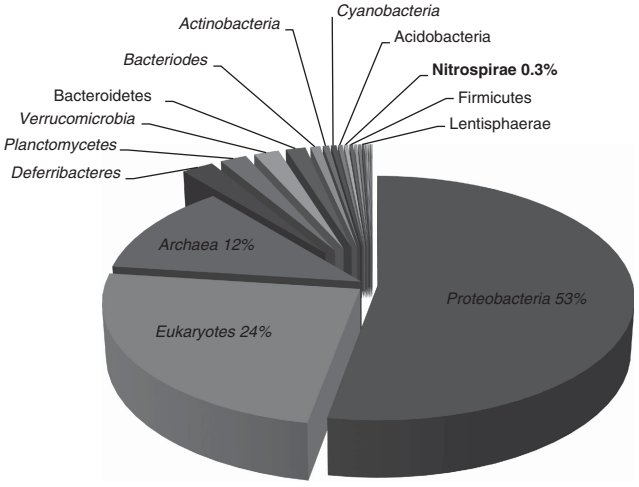

Figure 1 Abundance of major phyla based on classification of rRNA transcript reads. All 16S rRNA reads (total of 14 571 562) were mapped (>75% over half the read length cutoff) to a 16S rRNA gene database (SILVA).

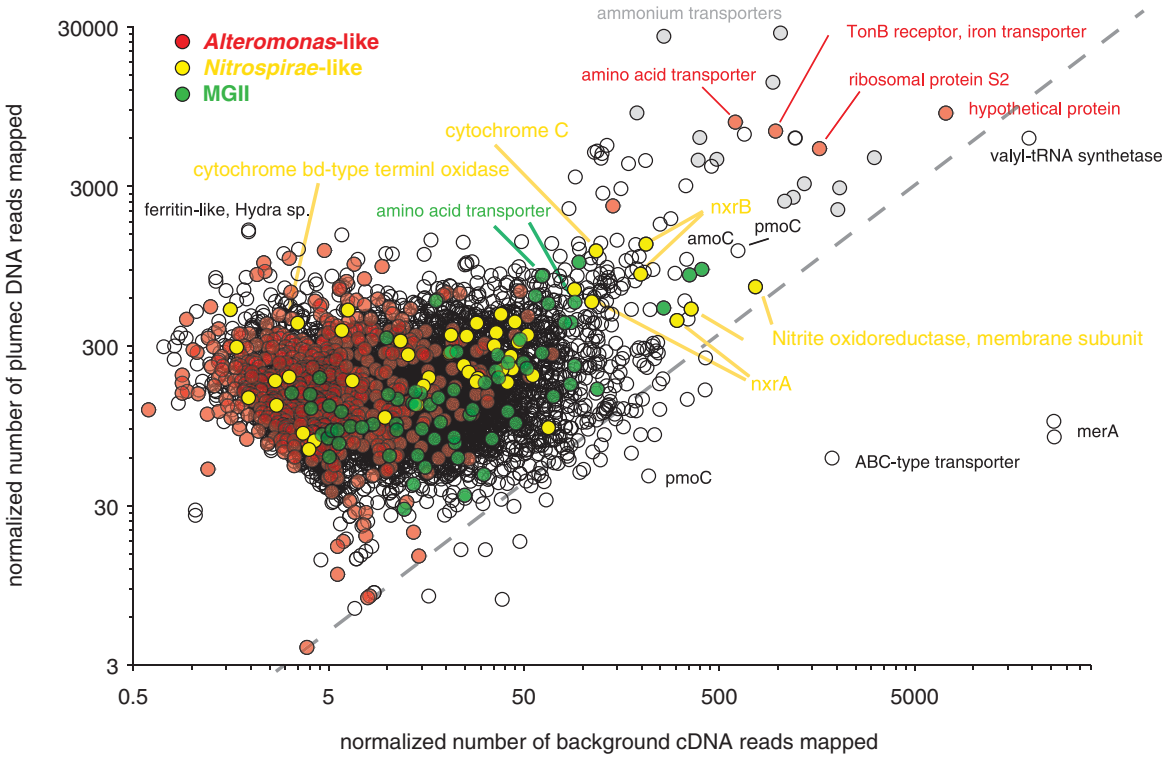

Figure 2 Abundance of gene transcripts in plume and background based on mapping transcripts to the plume *de novo* metatranscriptomic assembly. Red filled circles are mRNAs that have high similarity to *Altermonoas* sp., yellow are those related to *Nitrospirae* and green are MGII. Gray filled circles are highly transcribed ammonium transporters, most of these belong to AOA, consistent with previous findings (Baker *et al.*, 2012). The dotted line indicates equal representation of transcripts in plume and background.

genes. Overall, 1968 mRNA contigs were identified totaling just over 1.2 Mb of consensus sequence (Figure 2), with an average similarity of 96% to *A. macleodii*. These transcripts are generally less abundant in the background compared with plume (Figure 2). The four most abundant *Alteromonas*-like transcripts are for TonB, an amino-acid transporter, ribosomal protein S2 and a hypothetical protein (Figure 2). TonB is a membrane-bound receptor that is commonly involved in iron uptake systems in a variety of bacteria. However, it has been shown that this protein family is also involved in transport of other metals and various carbohydrates (Schauer *et al.*, 2008).

Given that ATP-binding cassette transporters are an essential component of hetrotrophy and uptake of dissolved organic carbon in the oceans (Jiao and Zheng, 2011), we compared transcriptional activity among putative ATP-binding cassette amino-acid transporters present in the metatranscriptome. Interestingly, 5 of the 10 most abundantly represented amino-acid transporters in the plume metatranscriptome have high similarity to Euryarchaea Marine Group II (MGII), suggesting this group utilizes exogenous amino acids as a carbon and/or nitrogen source. Given the low coverage of MGII in the metagenome (Figure 3 and Supplementary Figure S3), we searched the transcript assembly using a recently obtained MGII genome (Iverson *et al.*, 2012). A total of 112 transcript contigs (nearly 72 kb total) were identified with an average similarity of 91% to the MGII genome. Putative functions could be assigned to only 37 of these assembled contigs; the vast majority was annotated as 'hypothetical proteins', underscoring the lack of knowledge of this group. MGII are ubiquitous in marine environments (Martin-Cuadrado *et al.*, 2008), yet their physiology and function has remained enigmatic until their recent implication in heterotrophy (Iverson *et al.*, 2012).

MGII have proteorhodopsin genes for energy generation in the photic zone (Frigaard *et al.*, 2006). Upon searching for proteorhodopsin genes in the deep Guaymas metatranscriptome none were idnentified, as expected for a dark environment and consistent with previous 454-based results (Lesniewski *et al.*, 2012). We did, however, identify expression of a V-type H^+-translocating inorganic pyrophosphatase gene, which are implicated in energy generation in symbionts (Kleiner *et al.*, 2012). These results hint that deep-sea MGII Archaea utilize H^+-translocating inorganic pyrophosphatase as an important mechanism of energy conservation and ATP generation. Also prominent among MGII transcripts were several RNA processing genes including multiple RNA-binding *Rrp4* and *RNase PH* genes (Supplementary Figure S4).

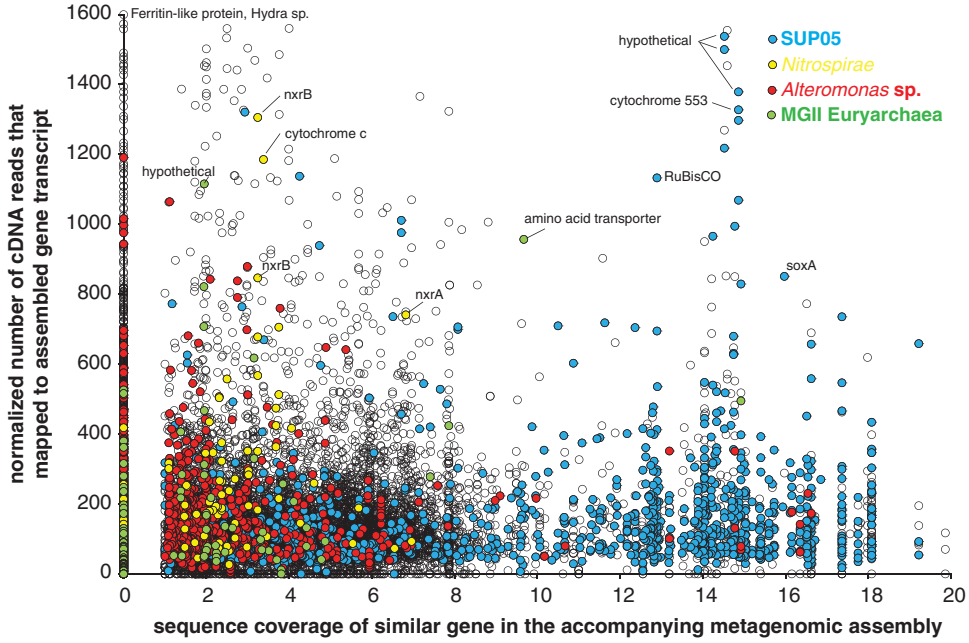

Figure 3 Plot of gene transcript abundance vs coverage in the metagenomic assembly from Lesniewski *et al.* (2012). Abundance is the number of cDNA reads mapped to the transcript, normalized to the length of the gene. Top matches in the genomic DNA library assembly are greater than e-value of $1E^{-10}$.

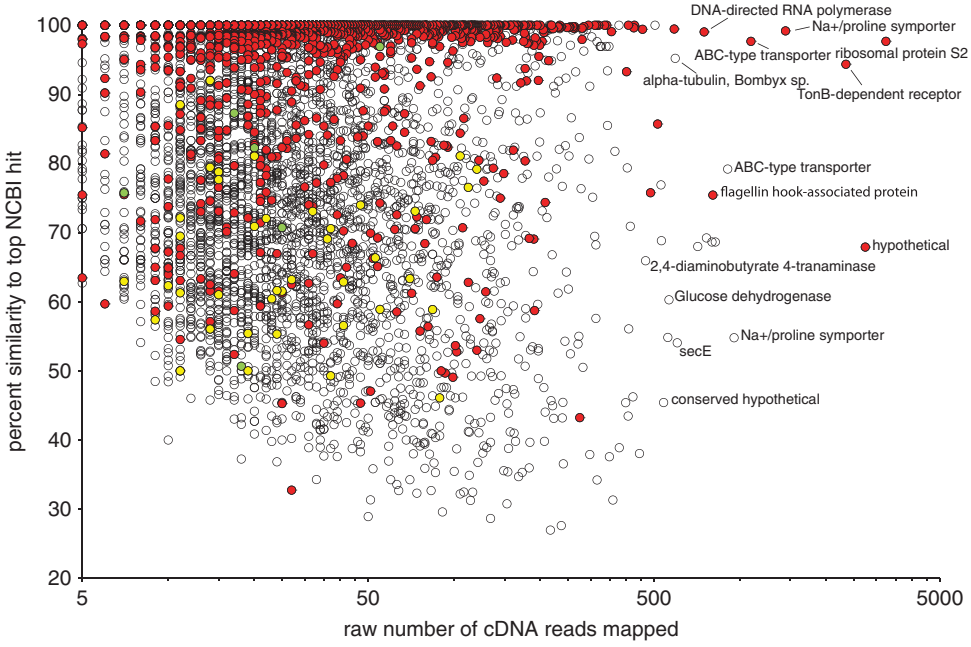

Figure 4 Transcripts not present in accompanying metagenomic data but with similarity to sequences in public databases. Each circle represents an assembled mRNA contig. Plotted is percent similarity to NCBI sequence vs the number of plume cDNA reads recruited. Coloring is consistent with Figures 2 and 3; red are *Alteromonas*, yellow are *Nitrospirae* and green are MGII.

Transcripts absent from metagenomic data

To further assess the extent of sequences present in the metatranscriptome but absent from the metagenome, we compared the transcripts with a prior metagenomic assembly derived from the same samples (Lesniewski *et al.*, 2012). Eight thousand three hundred sixty metatranscriptome-specific mRNAs were found, totaling over 3.4 Mb of consensus sequence. We were unable to assign potential function to 3419 (41%) of these genes, and 2447 did not have confident matches to sequences in public databases. Many of the most active genes present in this category are of unknown function (Supplementary Figure S5). Overall, 16% (1378 of 8360) of the metatranscriptome-specific genes are closely related to genes from *Alteromonas*, including the abundant TonB receptor and ribosomal proteins (Figure 4).

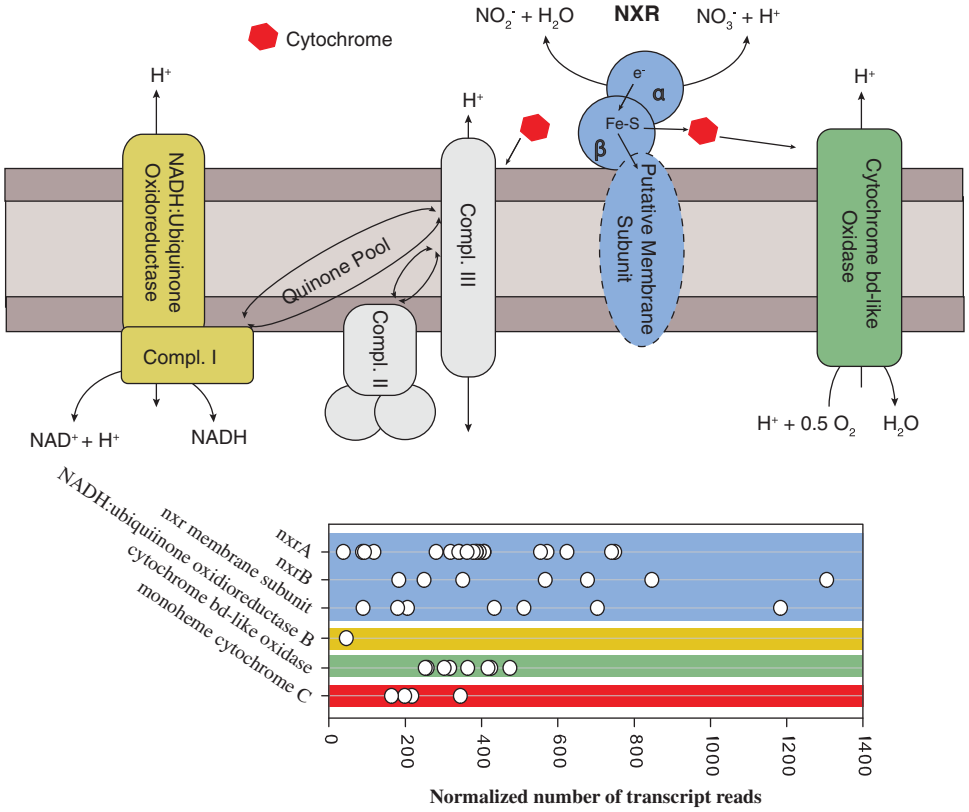

Figure 5 Schematic model and abundance of transcripts in the plume for proteins involved in nitrite oxidation and associated electron transfer. Colored proteins were detected in the plume cDNA libraries. Complexes in gray were not identified but are included in the model of electron transport for reference. Arrows show movement of electrons and protons. For transcript abundance, multiple circles for each gene represent multiple closely related gene sequence variants. Normalization is calculated as the number of cDNA reads mapped divided by lengths of the genes and multiplied by 1000.

Identification of nitrite oxidation transcripts

Some of the most abundant transcripts in the community are from genes for nitrite oxidation and associated energy metabolism (Figure 2). These highly transcribed genes encode the key enzyme for nitrite oxidation, nitrite oxidoreductase (NxrA, NxrB and the membrane subunit), as well a *c*-type cytochrome and cytochrome *bd*-type terminal oxidase for reduction of O_2 (Lücker *et al.*, 2010). Except for *nxrA*, all of these genes are most similar to *Ca.* Nitrospira defluvii (Supplementary Figure S6 and Supplementary Table S1), a nitrite-oxidizing member of the phylum *Nitrospirae* (Lücker *et al.*, 2010). These *nxr* genes are phylogenetically distinct from those recently discovered in *Chloroflexi* sp. (Sorokin *et al.*, 2012). Many of the components proposed to oxidize nitrite and reduce O_2 in *Ca.* N. defluvii are present and highly transcribed in the Guaymas Basin metatranscriptome (Figure 5).

Assignment of *nxr* genes to aerobic nitrite oxidation by *Nitrospirae* is complicated because at least two members of the phylum *Planctomycetes*, *Ca.* K. stuttgartiensis and *Ca.* Scalindua profundus, also contain *nxrA* and *nxrB*-like genes, which are thought to be involved in nitrite oxidation during anammox (Strous *et al.*, 2006). Although the prevalence of anammox seems unlikely in oxic waters of the deep Guaymas Basin ($\sim$28 μM O_2), it could take place in particle-associated, anoxic microenvironments (Wright *et al.*, 2012). Thus, to evaluate the possibility that the *nxr*-like transcripts we observed are from anammox microorganisms, attempts to identify additional anammox-related gene transcripts were made. No appreciable abundance of transcripts of key anammox genes, hydroxylamine oxidoreductase (*hao*) or hydrazine hydrolase (*hzh*), were identified in the metatranscriptomic assembly (Supplementary Figure S7). Further, the concurrent abundant expression of the cytochrome bd terminal oxidase most closely related to a *Ca.* N. defluvii homolog indicates aerobic metabolism, ruling out a role for NXR in anammox or H_2-linked denitrification, which has been suggested for *Nitrospira moscoviensis* (Ehrlich *et al.*, 1995). Therefore, we conclude that the abundant transcripts encoding a novel NXR and associated electron transport chain are involved in aerobic nitrite oxidation.

Guaymas Basin plumes are enriched in ammonium and hydrocarbons (Bazylinski *et al.*, 1989), thus, may be more representative of areas of intense nitrogen and carbon cycling (for example, oxygen minimum zones (Wright *et al.*, 2012)) than the typical deep ocean. However, the high abundance

of transcripts from *nxr* and associated electron transport genes in the non-plume background sample shows that their prominence is not restricted to ammonium-rich hydrothermal plumes (Figure 2 and Supplementary Figure S8).

Recovery and characterization of Nitrospirae 16S rRNA and rRNA genes

As the novel *nxr* transcripts are not directly linked to conserved phylogenetic markers (that is, do not co-occur on a single assembled contig), it is not possible to definitively assign these genes to taxa present in our data set. To probe this question further, we searched the transcript library for known NOB rRNA gene sequences. No 16S rRNA genes from common NOB genera (for example, *Nitrospina*, *Nitrobacter*, *Nitrococcus, Nitrospira*) were present; however, two phylotypes that fall within phylum *Nitrospirae* were identified. Phylogenetic analyses of the *Nitrospirae* 16S rRNA and 23S rRNA genes indicated two distinct clusters (81% and 82% 16S rRNA gene similarity to *Ca*. N. defluvii) that represent novel members of the *Nitrospirae* phylum (Figure 6). One of these phylotypes (referred to hereafter as deep ocean *Nitrospirae* nitrifier, 'DONN') recruited four times more rRNA transcripts than the other in the plume and was the only transcriptionally active *Nitrospirae*-like phylotype in the background sample. The closest match to DONN in public databases shares only 88% sequence identity to uncultured *Nitrospirae* clones (Figure 6), highlighting the novelty of this group. Comparison of all cDNA reads to a comprehensive 16S rRNA gene database revealed that only 0.3% of the rRNA reads matched most closely to *Nitrospirae* (Figure 1). In addition, $<0.001\%$ of all rRNA gene-containing reads from the genomic library were identified as *Nitrospirae*.

Prevalence of Nitrospirae metabolic gene transcripts

Further support for the assignment of *nxr* genes to *Nitrospirae* comes from the prevalence of additional abundant mRNA transcripts with high similarity to *Nitrospirae*. In total, we identified 160 *Nitrospirae*-like genes (including several species/strain variants) on 142 assembled mRNA fragments (Figure 7). Interestingly, 115 of these have similarity to contigs in the accompanying metagenomic data set (Figure 3) but are present at low coverage (2.4 times). To confirm this, we searched all the previously published 454 data sets (Lesniewski *et al.*, 2012) and found the same trend of a high (5:1) cDNA:DNA ratio in total community *nxr* genes. Taken together, these results reveal the low abundance yet high

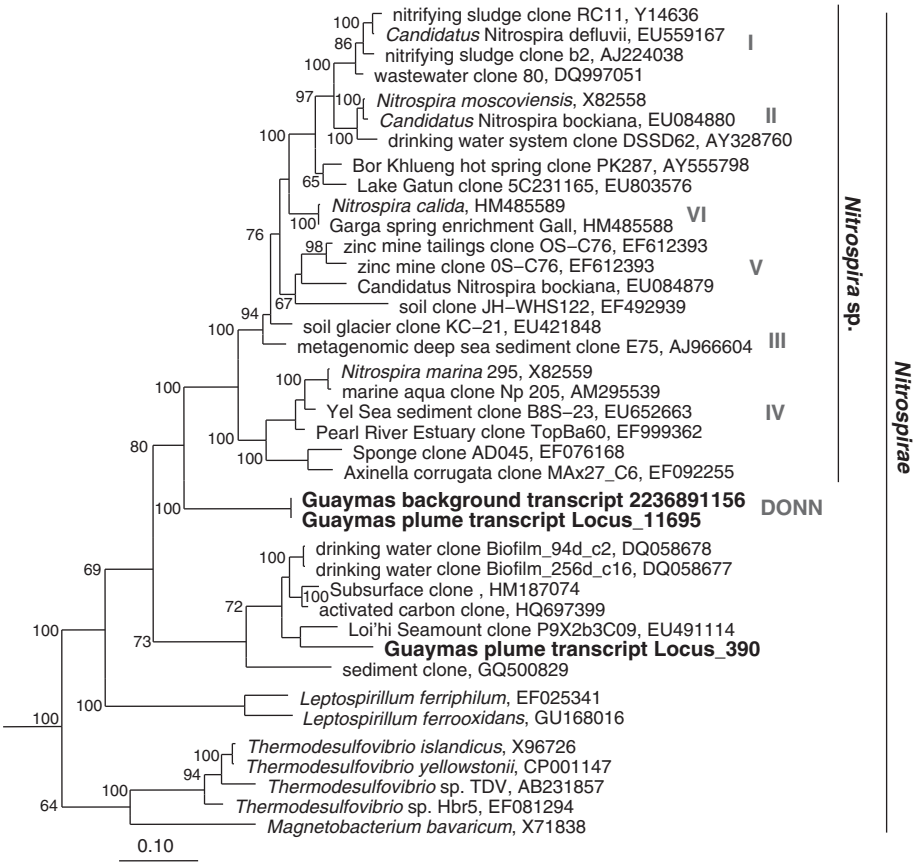

Figure 6 Phylogeny of *Nitrospira*-like 16S rRNA genes from assembled transcripts. Trees were generated using the maximum likelihood method and *Planctomycetes brasiliensis* as the outgroup.

Figure 7 Abundance of assembled transcripts most closely related to *Nitrospirae* from the plume transcript assembly. Each circle represents a distinct gene sequence, with assigned functions listed on the left. Thus, multiple data points for each gene represent sequence variants present in the community. Abundance is based on the number of reads that mapped to the assembled transcript. Normalization is calculated as the number of cDNA reads mapped divided by lengths of the genes and multiplied by 1000.

transcriptional activity of *Nitrospirae* in the deep Guaymas Basin (Supplementary Table S1).

NOB are thought to be primarily autotrophic, but there is evidence for enhanced growth of *Nitrospira* sp. when supplied with simple organic carbon sources such as pyruvate (Boon and Laudelout, 1962; Ehrlich *et al.,* 1995; Lücker *et al.,* 2010). All described *Nitrospirae* utilize the reductive tricarboxylic acid cycle. Among the abundant *Nitrospirae*-like transcripts in the Guaymas Basin metatranscriptome were those from genes integral to carbon metabolism via the TCA cycle. Transcripts of several strain variants of all three subunits (α, β and γ) of 2-oxoglutarate:ferredoxin oxidoreductase and pyruvate:ferredoxin oxidoreductase genes were identified (Figure 7). However, genes encoding the ATP-citrate lyase (indicative of CO_2 fixation via reductive TCA) were not recovered; thus, we are unable to verify whether the reductive TCA cycle operates for CO_2 fixation in these DONN populations as it does in *Ca.* N. defluvii (Lücker *et al.*, 2010).

Abundance and distribution of low-abundance yet transcriptionally active microbial groups

Given that a large proportion of transcripts originate from minor community members (MGII, *Alteromonas* and DONN groups), we sought to further assess the abundance and distribution of these groups in Guaymas Basin as well as in hydrothermal plumes of the Eastern Lau Spreading Center, which is located in the southwestern Pacific and hosts geochemically diverse hydrothermal vents. Analysis of high throughput 16S rRNA gene pyrosequencing libraries taken from various depths of the water column yielded OTUs corresponding to all three groups, and confirmed their presence across geographically disparate microbial communities. At Guaymas, three dominant 16S rRNA gene phylotypes of *Alteromonas* sp. were present, which collectively represent 1.01–4.04% of the total Guaymas Basin community at depths of 1300–1900 m. These phylotypes were not detected in near surface samples (12.5 m) but increased to 1.0 and 1.76% near the oxygen minimum zone (356 and 554 m, respectively). At Lau Basin, only two of the three Guaymas Basin *Alteromonas* phylotypes were detected. Two dominant MGII phylotypes were present as minor community members at Guaymas, comprising only 0–0.67% of the total community. MGII were not detected in the two near surface Guaymas Basin samples (12.5 and 356 m). At Lau only one of the two Guaymas MGII phylotypes were detected, and ranged in abundance from 0–1.3%.

We found that both *Nitrospirae* phylotypes are present as low-abundance community members at Guaymas Basin, both in previously obtained clone libraries from Dick and Tebo (2010) and in new pyrosequencing data (Supplementary Figure S9). The DONN group is most abundant in the deep basin, but even there it only accounts for ~0.25% of the community (Supplementary Figure S10). Similar *Nitrospirae* phylotpes were also identified as minor members of Lau Basin communities (Supplementary Figure S11), further suggesting that the novel *Nitrospirae* phylotypes reported here are widespread and consistently less abundant members of deep-sea microbial communities. It is also important to note that commonly used probes for the *Nitrospira* (Füssel *et al.*, 2012) have two nucleotide mismatches to the DONN group, so these organisms may have been missed by previous studies. Thus, more work is needed to assess the distribution of DONN in diverse marine environments where nitrification is prevalent.

The stark contrast in NXR abundance between transcript (high abundance) and metagenomic (low abundance) libraries calls attention to the concept that keystone ecological functions can be performed by low-abundance species of the biosphere. In the case of NOB, low abundance despite high metabolic activity may be inherent to their physiology. Cultured NOB grow slowly (Watson *et al.*, 1986), presumably owing to low free energy yield from nitrite oxidation (Boon and Laudelout, 1962), which likely constrains the abundance of *in situ* NOB populations. Further, the disparity in population size between NOB (low abundance) and AOA (high abundance) at Guaymas Basin implies that cell-specific nitrite oxidation rates must be large relative to those of ammonia oxidation (assuming quantitative conversion to nitrate by the NOB). In the common terrestrial NOB *Nitrobacter winogradskyi*, enzyme saturation is evident under micromolar concentrations of nitrite (Watson *et al.*, 1986), and it is estimated that the NXR enzyme may comprise 10–30% of total cell protein (Bock *et al.*, 1991). Increased transcription of *nxr* genes but not rRNA genes has also been observed in *Ca.* N. defluvii enrichments (Lücker *et al.*, 2010) and actively fertilized soils where *Nitrobacter*-like *nxr* expression was elevated (Wertz *et al.*, 2011). Our data suggests a similar scenario occurs in the deep sea, where NOB highly transcribe *nxr* genes to maximize nitrite oxidation that provides only modest energy and growth yield. Another possible explanation is that the disparity between DNA and RNA abundance of the NOB reflects a recent transcriptional response to nitrite in the environment, and the subsequent increase in DNA (cell division) would soon follow.

Intrapopulation variability

The high sequence coverage produced by *de novo* metatranscriptomic assembly provides opportunities to investigate gene sequence variation and ecological dynamics of strains within natural populations. In many cases, multiple sequence variants of each gene involved in nitrite oxidation were recovered, indicating the presence of several closely related strains or multiple gene copies within a genome (Figure 7). The most highly expressed transcripts tended to have the greatest number of variants within the data set, likely as a result of greater coverage of those regions. The NXR variants cluster into two divergent groups (82–84% similar at the DNA level), likely representing the two different *Nitrospirae* groups, but it is impossible to rule out the alternative interpretation of gene duplicates within a single genome as in *Ca.* N. defluvii (Lücker *et al.*, 2010). The most highly expressed *nxrAB* type in the plume is also the dominant type in the background, suggesting that the same strain is dominant in both communities. Many of the sequence substitutions between these transcripts are synonymous. For example, the two most abundant *nxrAB* operon variants (GBPt_c08738 and GBPt_c08738) have eight nucleotide polymorphisms within a 72-bp region in the *nxrB* transcripts, yet they have identical amino-acid sequences.

However, we also identified minor variants that were only present in the plume and that have seven distinct nucleotide substitutions clustered solely within the metal-coordinating [Fe-S] center of NxrB (Supplementary Figures S12 and S13). This site is homologous to a region of nitrate reductase of *Escherichia coli* that mediates intramolecular electron transfer. The high frequency of nucleotide polymorphisms around this region suggests that selective pressures (perhaps substrate concentrations) maintain such variation.

Conclusions

Despite the recent explosion of microbial genome sequencing, environmental shotgun sequencing continues to reveal vast genetic novelty, which presents fundamental challenges to our ability to fully characterize natural microbial communities. Our findings demonstrate that *de novo* metatranscriptomic assembly offers the ability to assess transcriptionally active populations of diverse and novel microbial communities at high resolution (to the strain level). More importantly, it enables the reconstruction and functional characterization of transcripts that would have otherwise been overlooked by mapping to reference genomic databases. In the deep Gulf of California, this approach revealed the functional importance of low-abundance populations of *Alteromonas* and archaeal MGII in heterotrophy and novel *Nitrospirae* in nitrite oxidation. The high RNA:DNA ratio and novelty of genes implicated in nitrite oxidation suggest explanations for why these *Nitrospirae* have eluded detection and are under-represented relative to their ammonia-oxidizing counterparts. These new insights into novel NOB indicate that the

distribution and role of this functional group should be reconsidered, as we seek to understand the fate of nitrite in terms of nutrient cycling and production of greenhouse gases in the oceans.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This research is funded by the Gordon and Betty Moore Foundation through Grant GBMF2609 to Dr Gregory Dick and National Science Foundation (OCE 1029242). We thank Drs Anders Andersson and Meng Li for their helpful discussions. We especially thank Donald Zak for revisions to the manuscript. We also thank Karthik Anantharaman for cDNA library preparations and Prashanna Balaji for assistance with read mapping.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Baker BJ, Lesniewski R, Dick GJ. (2012). Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. *ISME J* **6**: 2269–2279.

Bates ST, Berg-Lyons D, Caporaso JG, Walters WA, Knight R, Fierer N. (2010). Examining the global distribution of dominant archaeal populations in soil. *ISME J* **5**: 908–917.

Bazylinski DA, Wirsen CO, Jannasch HW. (1989). Microbial utilization of naturally occurring hydrocarbons at the Guaymas Basin hydrothermal vent site. *Appl Environ Microbiol* **55**: 2832–2836.

Bock E, Koops HP, Harms H, Ahlers B. (1991). In: Shively JM, Barton LL (eds) *Variations in Autotrophic Life*. Academic Press: London, pp 171–200.

Boon B, Laudelout H. (1962). Kinetics of nitrite oxidation by *Nitrobacter winogradskyi*. *Biochem J* **85**: 440–447.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Costello EK, Fierer N *et al.* (2010). QIIME allows analyses of high-throughput community sequencing data. *Nat Meth* **7**: 335–336.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Dick GJ, Tebo BM. (2010). Microbial diversity and biogeochemistry of the Guaymas Basin hydrothermal plume. *Environ Microbiol* **12**: 1334–1347.

Dore JE, Karl DM. (1996). Nitrite distributions and dynamics at station ALOHA. *Deep-Sea Res II* **43**: 385–402.

Ehrlich S, Behrens D, Lebedeva E, Ludwig W, Bock E. (1995). A new obligately chemolithoautotrophic, nitrite-oxidizing bacterium, *Nitrospira moscoviensis* sp. nov. and its phylogenetic relationship. *Arch Microbiol* **164**: 16–23.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al.* (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* **105**: 3805–3810.

Frigaard NU, Martinez A, Mincer TJ, DeLong EF. (2006). Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**: 847–850.

Füssel J, Lam P, Lavik G, Jensen MM, Holtappels M, Gunter M *et al.* (2012). Nitrite oxidation in the Namibian oxygen minimum zone. *ISME J* **6**: 1200–1209.

Ivars-Martínez E, Martin-Cuadrado AB, D'Auria G, Mira A, Ferriera S, Johnson J *et al.* (2008). Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISME J* **2**: 1194–1212.

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* **33**: 587–590.

Jiao N, Zheng Q. (2011). The microbial carbon pump: from genes to ecosystems. *Appl Environ Microbiol* **77**: 7439–7444.

Karl DM, Church MJ, Dore JE, Letelier RM, Mahaffey C. (2012). Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proc Natl Acad Sci USA* **109**: 1842–1849.

Karner MB, DeLong EF, Karl DM. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**: 507–510.

Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J *et al.* (2012). Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc Natl Acad Sci USA* **109**: E1173–E1182. early edition.

Koops H-P, Pommerening-Roser A. (2001). Distribution and ecophysiology of the nitrifying bacteria emphasizing cultured species. *FEMS Microbiol Ecol* **37**: 1–9.

Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**: 543–546.

Lebedeva EV, Off S, Zumbrägel S, Kruse M, Shagzhina A, Lücker S *et al.* (2011). Isolation and characterization of a moderately thermophilic nitrite-oxidizing bacterium from a geothermal spring. *FEMS Microbiol Ecol* **75**: 195–204.

Lesniewski R, Jain S, Anantharaman K, Schloss PD, Dick GJ. (2012). The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME J* **6**: 2257–2268.

Li H, Durbin R. (2009). Fast and accurate short read alignment with Burrow-Wheeler Transform. *Bioinformatics* **25**: 1754–1760.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nuc Acids Res* **32**: 1363–1371.

Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, Vacherie B *et al.* (2010). A *Nitrospira* metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci USA* **107**: 13479–13484.

Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Grechkin Y *et al.* (2012). IMG/M: the integrated metagenome data management and comparative analysis system. *Nucl Acids Res* **40**: D123–D129.

Martin-Cuadrado A-B, Rodriquez-Valera F, Moreira D, Alba D, Ivars-Martínez E, Henn MR *et al.* (2008). Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J* **2**: 865–886.

Miller CS, Baker BJ, Thomas BC, Singer S, Banfield JF. (2011). EMRIGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Gen Biol* **12**: R44.

Mincer TJ, Church MJ, Taylor LT, Preston C, Karl DM, DeLong EF. (2007). Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre. *Environ Microbiol* **9**: 1162–1175.

Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan L-K *et al.* (2012). Sizing up metatranscriptomics. *ISME J* **7**: 237–243.

Pester M, Schleper C, Wagner M. (2011). The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Curr Opin Microbiol* **14**: 300–306.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al.* (2007). SILVA: a comprehensive online resourse for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nuc Acids Res* **35**: 7188–7196.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM *et al.* (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Meth* **6**: 639–641.

Raven JA, Falkowski PG. (1999). Oceanic sinks for atmospheric CO_2. *Plant Cell Eviron* **22**: 741–755.

Santoro AE, Buchwald C, McIlvin MR, Casciotti KL. (2011). Isotopic signature of N2O produced by marine ammonia-oxidizing Archaea. *Science* **333**: 1282–1285.

Santoro AE, Casciotti KL, Francis CA. (2010). Activity, abundance and diversity of nitrifying archaea and bacteria in the central California current. *Environ Microbiol* **12**: 1989–2006.

Schauer K, Rodionov DA, de Reuse H. (2008). New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem Sci* **33**: 330–338.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

Schmieder R, Lim YW, Edwards R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* **28**: 433–435.

Schulz MH, Zerbino DR, Vingron M, Birney E. (2012). *Oases*: Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**: 1086–1092.

Shi YM, Tyson GW, DeLong EF. (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**: 266–269.

Shi YM, Tyson GW, Eppley JM, DeLong EF. (2011). Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J* **5**: 999–1013.

Sorokin DY, Lucker S, Vejmelkova D, Kostrikina NA, Kleerebezem R, Rijpstr WI *et al.* (2012). Nitrification expanded: discovery, physiology and genomics of a nitrite-oxidizing bacterium from the phylum Chloroflexi. *ISME J* **6**: 2245–2256.

Stewart FJ, Ulloa O, DeLong EF. (2011). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol* **14**: 23–40.

Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW *et al.* (2006). Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440**: 790–794.

Ward BB, Capone DG, Zehr JP. (2007). What's new in the nitrogen cycle? *Oceanography* **20**: 101–109.

Watson SW, Bock E, Valois FW, Waterbury JB, Schlosser U. (1986). *Nitrospira marina* gen. nov., sp. nov.: a chemolithotrophic nitrite-oxidizing bacterium. *Arch Microbiol* **144**: 1–7.

Wertz S, Leigh AKK, Grayston SJ. (2011). Effects of long-term fertilization of forest soils on potential nitrification and on the abundance and community structure of ammonia oxidizers and nitrite oxidizers. *FEMS Microbiol Ecol* **79**: 142–154.

Wright J, Konwar KM, Hallam SJ. (2012). Microbial ecology of expanding oxygen minimum zones. *Nature* **10**: 381–394.

Wuchter C, Abbas B, Coolen MJL, Herfort L, van Bleijswijk J, Timmers P *et al.* (2006). Archaeal nitrification in the ocean. *Proc Natl Acad Sci USA* **103**: 12317–12322.

Zehr JP, Kudela RM. (2011). Nitrogen cycle of the open ocean: from genes to ecosystems. *Ann Rev Mar Sci* **3**: 197–225.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)