Gene expression in the deep biosphere

William D. Orsi[1*], Virginia P. Edgcomb[1], Glenn D. Christman[2], and Jennifer F. Biddle[2]

Affiliations:

[1] Department of Geology and Geophysics, Woods Hole Oceanographic Institution

[2] College of Earth, Ocean, and Environment, University of Delaware

Keywords: marine subsurface, mRNA, metatranscriptomics, deep biosphere, Illumina sequencing, marine sediment, subseafloor, ODP Leg 201, gene expression

Running title: Gene expression in the deep biosphere.

*To whom correspondence should be addressed

Scientific ocean drilling has revealed a deep biosphere of widespread microbial life in sub-seafloor sediment. Microbial metabolism in the marine subsurface likely plays an important role in global biogeochemical cycles[1-3] but deep biosphere activities are not well understood[1]. Here, we describe and analyze the first subseafloor metatranscriptomes from anaerobic Peru Margin sediment up to 159 meters below seafloor (mbsf) represented by over 1 billion cDNA sequence reads. Anaerobic metabolism of amino acids, carbohydrates, and lipids appear to be dominant metabolic processes, and profiles of dissimilatory sulfite reductase (*Dsr*) transcripts are consistent with porewater sulfate concentration profiles[1]. Moreover, transcripts involved in cell division increase as a function of microbial cell concentration, indicating that increases in subseafloor microbial abundance are a function

of cell division across all three domains of life. These data support calculations[1] and models[4] of subseafloor microbial metabolism and represent the first holistic picture of deep biosphere activities.

Abundant microbial cells[5, 6] exist in sub-seafloor (>1.5 mbsf) sediment and represent a significant portion of Earth's biomass[7, 8]. Marine sediment contains Earth's largest pool of organic carbon, which may be the primary energy source for subsurface microbes[1, 2, 9, 10, 11]. A model recently suggested biomass turnover rates on the order of thousands of years in the marine subsurface that are hypothesized to have an impact on global biogeochemical cycling over geological timescales[4]. Logistical sampling constraints, the complex sediment matrix composed of organic material and minerals, and low metabolic rates[3,4], have hindered directed testing of microbial activities at the molecular level in this environment. A better understanding of deep biosphere activities will help define the deep biosphere's role in global biogeochemical cycles[12].

We optimized an mRNA extraction and amplification protocol for subseafloor sediment, and combined this with high-throughput sequencing to report the first dataset on microbial gene expression in the marine subsurface, demonstrating that despite the extremely low metabolic rates[1, 4], mRNA-based investigations of the deep biosphere are possible and informative. We use the gene expression data to reconstruct active community metabolism and results support calculations[1] and models[4] of sub-seafloor microbial activities. The Peru Margin (Ocean Drilling Program Leg 201, Site 1229D) was analyzed because a wealth of biogeochemical data exists for this site [e.g. 1, 4,, 6, 9, 10] that exhibits peaks of cell abundance, and profiles of sulfate and methane suggestive of microbial activity[1] (Figure 1).

Picogram quantities of total RNA were extracted from 25 grams of Peru Margin sediment from six depths (5, 30, 50, 70, 91, 159 mbsf), consistent with basal levels of microbial activity predicted for this environment[3, 4]. Illumina® sequencing of total cDNA produced over 1 billion reads, with 50% to 85% of reads mapping to open reading frames that were assigned a functional annotation (Table S1).

The dominance of transcripts from Firmicutes, Actinobacteria, Alphaproteobacteria, and Gammaproteobacteria (Fig. S1) is consistent with previous cultivation-based, metagenomic, and phylogenetic surveys from Peru Margin subsurface sediment[1, 5, 13, 14], and suggests these to be some of the most active microbial groups. The abundance of gammaproteobacterial transcripts (Fig. S1) suggests that they are likely the most active microbial group in the deeper, anoxic, subseafloor sediment at this site. Fungal transcripts were also present in every sample ranging in representation from 3% at 70 mbsf to 20% at 5 mbsf. Archaea and Chloroflexi are present in noticeably low abundance, despite their previous detection at this site[6, 13, 15], suggesting that our approach might miss organisms with lower mRNA expression levels. As such, interpretations of relative abundances should be treated cautiously[16]. Changes in pressure and temperature may have altered gene expression during sampling. However, low representation of heat shock proteins (a proxy for physiological stress response[17]) in protein coding reads ($< 10^{-5}$ %) suggests the physiological state of most microbes was not significantly altered during sample retrieval and storage.

Dissimilatory sulfate reduction may represent a major form of microbial metabolism and energy production in the sub-seafloor[1, 2, 18] and is indicated by porewater sulfate concentrations at Site 1229[1] (Fig. 1). Representation of *Dsr* transcripts was highest in sediment with sulfate profiles suggestive of biogenic sulfate reduction (Fig. 1) and supports biogeochemical evidence for sulfate reduction at this site[1, 4]. Surprisingly, transcripts coding for dissimilatory nitrate reductases (*Nar*) were represented throughout the sediment column, despite no measureable nitrate. The origin of nitrate as a substrate in this sediment is unknown, but could potentially be produced as a by-product of anaerobic ammonium oxidation. Once produced, nitrate would likely not accumulate to measurable concentrations given the higher free energy yield of nitrate as electron acceptor compared to the dominant electron acceptors in this environment, sulfate and iron. Nitrate reduction appears to be performed predominantly by Alphaproteobacteria and Betatproteobacteria at most depths (Fig. 1) and the resulting nitrite is likely reduced by Fungi, Gammaproteobacteria, and Firmicutes (Fig. S3). In contrast, Deltaproteobacteria and Firmicutes are the dominant groups expressing *Dsr* transcripts at 5 and 30 mbsf, and Gammaproteobacteria were the only group with detectable *Dsr* transcripts at deeper depths (Fig. 1). Expression of *Dsr* transcripts from a methanogenic lineage (Fig 1) in the deep biosphere supports the evidence that anaerobic oxidation of methane (AOM) may not be an obligate syntrophic process[19].

Gene expression from methanogenic lineages was found, including Methanosarcinales, which contain the anaerobic methane-oxidizing group ANME-2[20] (Fig. S4). However, we did not detect any transcripts coding for methyl-coenzyme reductase M (*mcrA*), arguably the best diagnostic enzyme for AOM and methanogenesis. This could be explained by low levels of archaeal mRNA expression and a masking of *mcrA* gene expression by archaeal housekeeping genes. As a DNA-based study detected *mcrA* genes from this site[21], this explanation seems likely. Consistent with DNA-based observations from other sites[20], gene expression from methanogens was detected in the sulfate reduction zones (Fig. S4). Methylotrophic methanogenesis has been documented in shallow sediment sulfate reduction zones that contain non-competitive substrates such as trimethylamine[22, 23]. Our detection of trimethylamine methyltransferase transcripts from Methanosarcinales and Methanobacteriales (Fig S4) suggests that this process occurs in the deep subseafloor and support previous suggestions of biogenic methane at this site[1]. While Crenarchaeota have been suggested to be dominant at this site[6, 13, 15], they are a minority contribution to the metatranscriptome (Fig S1) even with incorporating new, partially completed, single cell genomes from shallow sediments[24] (Table S2). One explanation is that Crenarchaeota may have relatively low levels of mRNA expression in the deep biosphere.

A model suggests turnover of microbial biomass in this environment[4], but at the extremely low metabolic rates proposed it is unknown whether growth yield leads to cell division or to biomass turnover without division[4, 25]. Representation of transcripts involved in cell division (Table S3) increases at sulfate methane transition zones (SMTZs) where cell abundances increase by an order of magnitude ($p = 0.03$, Figs 1, S5). Our data suggest that the portion of the vegetative population that is actively dividing is largest in the SMTZs, and that observed peaks in cell counts at SMTZs are a result of *in situ* cell division. Cell division transcripts from all three domains of life strongly indicate a diversity of actively dividing cells in deeply buried sediment, including Fungi. The dominance of transcripts involved in amino acid metabolism (Fig. 2) and coding for peptidases (Fig. S6) support a recent model of amino acid turnover in the deep biosphere[4] and evidence for peptidase activity in shallow marine sediments[24].

Microbial motility has been proposed for deep sediment[5], however, calculations of mean metabolic rates suggest that flagellar motility may not be possible in the deep biosphere[26]. We detected expressed ORFs involved in in flagellar, gliding, and twitching based motility (Table S3) up to 159 mbsf (Fig. 3) and the abundance of these categories decreases with decreasing sediment porosity ($p = 0.01$, Fig. 3), indicating that microbial motility is related to the space available for movement. The evidence for motility presented here implies that metabolic rates are not equal across all cells in the deep biosphere and that some cells may be significantly more metabolically active than others. The offset in taxonomic assignment of motility reads (Fig. S7) relative to total mRNA reads (Fig. S1) is suggestive of such differences.

DNA repair may represent a mechanism by which microbes in the deep biosphere are able to cope with the slow degradation of DNA over geological timescales due to spontaneous chemical or radiolytic reactions in the subseafloor[25, 26]. The representation of DNA repair transcripts involved in nucleotide excision and mismatch repair (Table S3) increases linearly with sediment depth ($p = 0.004$, Fig. 3). This suggests DNA repair is a survival mechanism for microbial populations in ancient sediment and supports the suggestion that dormancy may not be a feasible survival strategy for the deep biosphere, because it does not completely arrest the slow degradation of DNA[25, 26].

Fungal metabolic transcripts confirm previous suggestions of living fungi in the subseafloor[9, 13, 27], and are the first direct evidence for active fungal metabolism in the deep biosphere. Five percent of transcripts involved in carbohydrate, amino acid, and lipid metabolism were assigned to Fungi, suggesting that Fungi play an overlooked role in organic carbon turnover in sub-seafloor sediment (Fig. 2). Fungal expression of transcripts coding for hydrolases involved in protein, carbohydrate, and lipid degradation (Fig. S6) indicates they degrade a variety of organic carbon substrates in deep subseafloor sediment.

Microbial expression of antibiotic defense mechanisms, polyketide synthases, and non-ribosomal proteins was detected (Fig. S8). Polyketide synthases and non-ribosomal proteins are involved in the biosynthesis of natural products (*e.g.* antibiotics, immunosuppressants, antifungals) of clinical and industrial importance. These findings warrant further investigation into potentially novel secondary metabolites produced by the deep biosphere, and support the hypothesis that the deep biosphere may represent a “seed bank” of novel biotechnological and biomedical innovation[28].

A comparison of the metatranscriptomic data to existing metagenomic datasets from this site[13, 29] reveals an increased representation of key metabolic and cell cycle functional genes in the metatranscriptome including those involved in DNA repair, DNA replication, transcription, amino acid biosynthesis, and lipid biosynthesis (Fig. 4). The significant difference between mRNA and metagenome samples with similar biogeochemical profiles (upper SMTZ and 50 mbsf: 5/12 samples) suggests these to be some of the more active processes. Although not a primary group in the overall annotations, activity of Archaea in the deep biosphere is highlighted by archaeal ATPase and DNA polymerase transcripts that are overrepresented in the metatranscriptomes relative to metagenomes ($p < 0.0005$). An analysis of similarity test (ANOSIM) indicates that the gene expression approach captures a significantly different picture of microbial activities compared to DNA based data ($p=0.001$, Fig. S9). As deep biosphere studies move forward, joint investigation of both nucleic acid pools are needed for full interpretation of metabolic activity and potential.

Metatranscriptomic analysis enables a refined view of deep biosphere activities. Microbial activity in deeply buried marine sediment is important because the collective activities of subsurface microbiota directly influences whether elements such as carbon are sequestered for millions of years in sediment or returned to the ocean, impacting food webs and climate[12]. Our data suggest the latter is mediated by diverse metabolic activities across all three domains of life in the sub-seafloor.

METHODS SUMMARY

Sample collection. Subsurface sediment samples from the continental shelf of Peru, Ocean Drilling Program (ODP) Site 1229D (77º 57.4590’ W, 10º 58.5721’ S), were obtained during ODP Leg 201 on March 6th, 2002.

RNA extraction, purification, and amplification. RNA was extracted from 25 g subseafloor sediment according to the protocol described by Orsi *et al* [26] using the FastRNA Pro Soil-Direct Kit® (MP Biomedicals, Solon, OH). In addition to the manufacturers instructions, physical and chemical adjustments to the sample were used to increase RNA yield and purity (see Supplemental Methods). DNA was removed using the Turbo DNA-free® kit (Life Technologies, Grand Island, NY), increasing the incubation time to 1 hour to ensure rigorous DNA removal. The MEGA-Clear® RNA Purification Kit (Life Technologies, Grand Island, NY) was used to further purify the RNA. Removal of contaminating DNA in RNA extracts was confirmed by the absence of visible amplification of SSU rRNA genes after 35 cycles of PCR using the RNA extracts as template. Total RNA was used as template for cDNA amplification using the Ovation RNA-Seq v2 System® (NuGEN technologies).

Bioinformatic analyses. Quality control was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Read assembly and mapping were performed in CLC Genomics Workbench 5.0 (CLC Bio Inc.). The Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline (RAMMCAP) available through CAMERA (http://camera.calit2.net/) was used to annotate contigs against COG and Pfam databases. Heatmaps and statistical tests were performed in R (http://www.r-project.org/) using the vegan (http://vegan.r-forge.r-project.org/) and matR (metagenomics.anl.gov) packages. Taxonomic assignments of contigs were performed using PhymmBL[30] with addition of fungal genomes available in the NCBI RefSeq and JGI databases and four partial single cell archaeal genomes from a shallow sediment site[24].

REFERENCES

1. D'Hondt, S. *et al.* Distributions of microbial activities in deep subseafloor sediments. *Science* **306**, 2216-2221 (2004).
2. Schrenk, M. O., Huber, J. A., & Edwards, K. J. Microbial provinces in the subseafloor. *Ann Rev Mar Sci* **2**, 279-304 (2010).
3. Jorgensen, B. B. & D'Hondt, S. A starving majority deep beneath the seafloor. *Science* **314**, 932-934 (2006).
4. Lomstein, B. A., Langerhuus, A. T., D'Hondt, S., Jorgensen, B. B., & Spivack, A. J. Endospore abundance, microbial growth and necromass turnover in deep sub-seafloor sediment. *Nature* **484**, 101-104 (2012).
5. Parkes, J., Cragg, B., & Wellsbury, P. Recent studies on bacterial populations and processes in subseafloor sediments: A review. *Hydrogeol J* **8**, 11-28 (2000).
6. Biddle, J. F. *et al.* Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. *Proc Natl Acad Sci USA* **103**, 3846-3851 (2006).
7. Kallmeyer, J., Pockalny, R., Adhikari, R., Smith, D. C., & D'Hondt, S. Global distributions of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci U S A* **109**, 16213-16216 (2012).
8. Whitman, W. B., Coleman, D. C., & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc Natl Acad Sci U S A* **95**, 6578-6583 (1998).

9. Biddle, J. F., House, C. H., & J. E. Brenchley. Microbial stratification in deeply buried marine sediment reflects changes in sulfate/methane profiles. *Geobiology* **3**, 287-295 (2005).

10. D'Hondt, S. *et al.* Subseafloor sedimentary life in the South Pacific Gyre. *Proc Natl Acad Sci U S A* **106**, 11651-11656 (2009).

11. D'Hondt, S., Rutherford, S., & Spivack, A. J. Metabolic activity of subsurface life in deep-sea sediments. *Science* **295**, 2067-2070 (2002).

12. Hinrichs, K. U. & Inagaki, F. Downsizing the deep biosphere. *Science* **338**, 204-205 (2012).

13. Biddle, J. F., Fitz-Gibbon, S., Schuster, S. C., Brenchley, J. E., & House, C. H. Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proc Natl Acad Sci USA* **105**, 10583-10588 (2008).

14. Teske, A. In: *Proceedings of the Ocean Drilling Program, Volume 201, Scientific Results.* B. B. Jørgensen *et al*., Eds. (ODP, College Station, TX, 2006), Chapter 2, pp. 1–19 (http://www-odp.tamu.edu/publications/201_SR/120/120.htm).

15. Lipp, J. *et al*. Significant contribution of Archaea to extant biomass in marine subsurface sediments. *Nature* **454**, 991-994 (2008).

16. Moran, M. A. *et al.* Sizing up metatranscriptomics. *ISME J* **7**, 237-243 (2012).

17. Gao, H. *et al*. Global transcriptome analysis of the heat schock response of Shewanella oneidensis. *J Bacteriol* **22**:7766-7803 (2004).

18. Jørgensen, B. B., D'Hondt, S., & Miller, D. J. In: *Proceedings of the Ocean Drilling Program, Volume 201, Scientific Results,* (ODP, College Station, TX, 2006), pp. 1–45 (www-odp.tamu.edu/publications/201_SR/201sr.htm).

19. Milucka, J. *et al*. Zero valent sulphur is a key intermediate in marine methane oxidation. *Nature* **491**, 541-546 (2012).

20. Lever, M. Functional gene surveys from ocean drilling expeditions - a review and perspective. *FEMS Microbiol Eco* **84**, 1-23 (2013).

21. Webster, G., Parkes, R. J., Cragg, B. A., Newberry, C.J., Weightman, A. J., Fry, J. C. Prokaryotic community composition and biogeochemical processes in deep subseafloor sediments from the Peru Margin. *FEMS Microbiol Ecol* **58**, 65-85 (2006).

22. Oremland, R. S. & Polcin, S. Methanogenesis and Sulfate Reduction: Competitive and noncompetitive substrates in estuarine sediments. *Appl Environ Microbiol* **44**, 1270-1276 (1982).

23. Valentine, D. L. Emerging topics in marine methane biogeochemistry. *Annu Rev Mar Sci* **3**:147-171 (2011).

24. Lloyd, K. *et al*. Predominant archaea in marine sediments degrade detrital proteins. *Nature* **496**, 215-218 (2013).

25. Jorgensen, B. B. Deep subseafloor microbial cells on physiological standby. *Proc Natl Acad Sci USA* **108**, 18193-18194 (2011).

26. Hoehler, T. M & Jorgensen, B. B. Microbial life under extreme energy limitation. *Nat Rev Microbiol* **11**, 83-94 (2013).

27. Orsi, W., Biddle, J., Edgcomb. Deep sequencing of subseafloor eukaryotic rRNA reveals active Fungi across multiple subsurface provinces. *PLoS ONE* **8**, e56335 (2013).

28. Parkes, R. J. & Wellsbury P. in *Microbial Diversity and Bioprospecting.* (ed Bull, A.T.) 120-129 (ASM Press, 2004).

29. Martino, A. J. *et al*. Novel degenerate PCR method for whole-genome amplification

applied to Peru Margin (ODP Leg 201) subsurface samples. *Front Microbiol* **3**, 17 (2012).

30. Brady, A. & Salzberg, S. L. Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated markov models. *Nature Methods* **6**, 673-676 (2009).

-

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was fostered by a Center for Dark Energy Biosphere Investigations (CDEBI) grant OCE-0939564 to WO and a NSF IOS grant 1238801 to JFB. We thank Christopher House (Penn State U.) and Andreas Teske (UNC) for graciously providing samples. We also thank Mitchell Sogin and Richard Fox at the Josephine Bay Paul Center (Marine Biological Laboratory, Woods Hole, MA) for kindly providing access to computing resources. Edward Leadbetter (WHOI) and Steven Hallam (UBC) provided helpful comments on the manuscript, and we also thank Steven D'Hondt (URI) for helpful discussions on the deep biosphere. This is CDEBI contribution 137.

Author Contributions W.O. performed experiments, analyzed data, and wrote the paper; W.O., J.B., and V.E. designed experiments and developed ideas. W.O. and G.C. developed analytical tools. All authors participated in data interpretation and provided editorial comments on the manuscript.

Author Information Data has been deposited in the NCBI Short Read Archive under accession number SRA058813 and in MG RAST (metagenomics.anl.gov) under accession numbers 4515478.3, 4515477.3, 4515476.3, 4510337.3, 4510336.3, and 4510335.3. Reprints and permission information is available at www.nature.com/reprints. The authors declare no competing interests. Correspondence and requests for materials should be addressed to W.O. (william.orsi@gmail.com).

Figure and Table legends

Figure 1: Biogeochemical and gene expression profiles of the deep biosphere from Peru Margin sediments, IODP site 1229D. **(a)** Cell abundance, sulfate concentrations, and methane concentrations, dotted lines indicate the SMTZs. Values were taken from the Ocean Drilling Program Janus Database (http://www-odp.tamu.edu/database/). **(b)** Proportion of cell division transcripts within the cluster of orthologous genes (COG) class D (cell cycle control/cell division/chromosome partitioning, n = 30.22 million reads), see Table S3 for description of cell division proteins. The proportion of **(c)** *Dsr* and **(d)** *Nar* transcripts relative to total transcripts involved in energy production (COG class C, n = 92.33 million reads). See Figure S2 for number of sequences and ORFs used in each comparison, and *E*-values for hits in the COG database.

Figure 2: The proportion of reads mapping to ORFs assigned to amino acid, lipid, and carbohydrate metabolism (eleven most dominant taxa shown). Note the relative abundance of amino acid metabolism (both anabolic and catabolic) relative to lipid and carbohydrate metabolism across all depths. See Figure S2 for the number of sequences and ORFs used in each comparison, and *E*-values for hits in the COG database.

Figure 3: Transcripts involved in cell motility and DNA repair. **(a)** The percentage of reads mapping to ORFs coding for proteins involved in different modes of cellular motility (see Table S3 for descriptions). **(b)** A correlation of cell motility transcripts versus sediment porosity (R^2 = 0.8, p = 0.01) and 95% prediction interval (red dotted lines). **(c)** The percentage of reads mapping to ORFs involved in DNA repair (only eleven most dominant taxa are shown, see Table S3 for descriptions). **(d)** A correlation of DNA repair transcripts versus sediment depth (R^2 = 0.9, p = 0.004) and 95% prediction interval (red dotted lines). See Figure S2 for the number of sequences and ORFs used in each comparison and *E*-values for ORF hits in COG database.

Figure 4: A comparison of gene expression data to existing metagenomic studies[13, 29] from IODP site 1229. Functional genes significantly (Kruskal-Wallis test, $p < 0.0005$) overrepresented in the metatranscriptomic samples relative to metagenomic data include DNA repair and replication transcripts, RNA polymerase, and archaeal ATPase and DNA polymerase transcripts. The dendrogram represents a UPGMA hierarchical clustering analysis (Manhattan distance) of significantly overrepresented mRNA transcripts, note the complete separation of mRNA samples from DNA samples.

Methods

Sample collection and storage Subsurface sediment samples from the continental shelf of Peru, Ocean Drilling Program (ODP) Site 1229D (77º 57.4590’ W, 10º 58.5721’ S), were obtained during ODP Leg 201 on March 6th, 2002. Careful precautions were taken during sampling to avoid contamination during the sampling process. For IODP cores, contamination tests were performed using Perfluorocarbon tracers and fluorescent microspheres (for more information see http://www-odp.tamu.edu/publications/201_IR). Sediment samples were immediately frozen at -80 ºC after sampling and stored at -80 ºC until used for mRNA extractions in this study (10 year storage time at -80 ºC).

RNA extraction and purification Extraction of subseafloor RNA was performed according to the protocol of Orsi *et al* [26]. To summarize, RNA was extracted from 25 grams of sediment using the FastRNA Pro Soil-Direct Kit® (MP Biomedicals, Solon, OH). It was necessary to scale up the volume of sediment that is typically extracted with the kit (~0.5 grams) due to the low biomass inherent to marine subsurface samples. All tubes, tips, and disposables used were certified RNAse free and all extraction procedures were performed in a laminar flow hood to reduce aerosol contamination by bacterial and fungal cells/spores. Five 15ml Lysing Matrix E® tubes (MP Biomedicals, Solon, OH) were filled with 5 g sediment and 5 ml of Soil Lysis Solution® (MP Biomedicals, Solon, OH). Tubes were vortexed to suspend the sediment and Soil Lysis Solution® was added to the tube leaving 1 ml of headspace. Tubes were then homogenized for 60 seconds on the FastPrep-24 homogenizer® (MP Biomedicals, Solon, OH) with a setting of 4.5. Contents were pooled into two 50ml tubes and centrifuged for 30 minutes at 4,000 RPM (3220 x g) at room temperature (RT). Supernatants were combined in a new 50ml tube and 1/10 volume of 2M Sodium Acetate (pH 4.0) was added. An equal volume of phenol-chloroform (pH 6.5) was added and vortexed for 30 seconds, incubated for 5 minutes at room temperature, and spun at 4000 RPM (3220 x g) for 20 minutes at 4 ºC. The aqueous phase was transferred to a new 50ml tube. Nucleic acids were precipitated by adding 2.5 and 1/10 volumes 100% ethanol and 3M Sodium Acetate, respectively, and incubating overnight at -80 ºC. The next day, tubes were spun at 4000 RPM (3220 x g) for 60 minutes at 4 ºC and the supernatant removed. Pellets were washed with 70% ethanol, spun for 15 minutes at 4 ºC, and air-dried. Dried pellets were resuspended with 0.25 ml RNAse-free sterile water and combined into a new 1.5ml tube. 1/10 volume of 2M Sodium Acetate (pH 4.0) and an equal volume of phenol:chloroform (pH 6.5) were added, vortexed for 1 minute, and incubated for 5 minutes at RT. This was necessary to remove residual organic material (*i.e.* humic acids) resulting from the rather large pellet/precipitate. After centrifuging at 14,000 RPM (20,817 x g) for 10 minutes at 4 ºC, the top phase was removed into a new 1.5ml tube. 0.7 volumes of 100% isopropanol was added and incubated for 1 hour at -20 ºC (to precipitate nucleic acids). Tubes were then centrifuged for 20 minutes at 14,000 (20,817 x g) RPM at 4 ºC and the supernatant removed. Pellets were washed with 70% ethanol and centrifuged at 14,000 RPM (20,817 x g) for 5 minutes at 4 ºC. After removing ethanol and air-drying, pellets were resuspended in 0.2 ml of RNAse free sterile water. DNA was removed using the Turbo DNA-free® kit (Life Technologies, Grand Island, NY), increasing the incubation time to 1 hour to ensure rigorous DNA removal. After this step, samples were taken through the protocol supplied with the FastRNA Pro Soil-Direct kit® to the end (starting at the RNA Matrix® and RNA Slurry® addition step), including the column purification step to remove residual humic acids (see FastRNA Pro Soil-Direct Kit® manual). Extraction blanks were performed (adding sterile water instead of sample) to ensure that aerosolized contaminants did not enter sample and reagent tubes during the extraction process. Absence of DNA and RNA contamination was confirmed by no visible amplification of small subunit (SSU) rRNA and rRNA genes from extraction blanks after 35 cycles of PCR and RT-PCR.

After RNA extraction, used the MEGA-Clear® RNA Purification Kit (Life Technologies, Grand Island, NY) to purify the RNA. This kit removes short RNA fragments (mostly produced during the extraction protocol) and residual inhibitors (*i.e*. humics). We followed the protocol all the way through the optional precipitation/concentration step, resuspending the RNA pellet in 10 microliters of RNAse free sterile water. Prior to cDNA amplification, the removal of contaminating DNA in RNA extracts was confirmed by the absence of visible amplification of SSU rRNA genes after 35 cycles of PCR using the RNA extracts as template.

cDNA amplification and Illumina sequencing Five microliters of purified RNA was used as template for whole cDNA amplification using the Ovation RNA-Seq v2 System® (NuGEN technologies, http://www.nugeninc.com/nugen/index.cfm/products/cs/ngs/rna-seq-v2/). We followed the manufacturers instructions for cDNA amplification, and the resulting quantity of cDNA was checked on a Nanodrop (Thermo Scientific) and Fluorometer (Qubit 2.0, Life Technologies). Quality of the amplified cDNA was checked on a Bioanalyzer (Agilent Biotechnologies) prior to Illumina® sequencing. Illumina® library preparation and paired-end sequencing was performed at the University of Delaware Sequencing and Genotyping Center (Delaware Biotechnology Institute, Newark DE).

Quality control and assembly Quality control of the dataset was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), with a quality score cutoff of 28. Approximately 1 billion paired-end reads that passed quality control were imported into CLC Genomics Workbench 5.0® (CLC Bio Inc.) and assembled using the paired-end Illumina assembler. Contigs were assembled over a range of kmer sizes (20, 50, 60, 64) with a minimum contig size cutoff of 300 nucleotides. The kmer size of 50 resulted in the highest number of contigs and these contigs were chosen for use in downstream analyses. To reduce the formation of chimeric assemblies, we used a paired-end sequencing approach and performed assemblies without scaffolding. Reads were mapped onto the contigs using the read mapping option in CLC Genomics Workbench to retain information on relative abundance of contigs.

Functional annotation of contigs Contigs were submitted to CAMERA (Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis, http://camera.calit2.net/) and assigned to clusters of orthologous gene (COG) families, gene ontologies (GO), and protein families (Pfam), using the Rapid Analysis of Multiple Metagenomes with a Clustering and Annotation Pipeline (RAMMCAP) using the 6 reading frame translation option for open reading frame (ORF) prediction and BLASTn for rRNA identifications. The cutoff criterion *E*-value of 10^{-5} was used for BLASTx searches against the COG, Pfam, and TIGRfam databases. For identification of bacterial and archaeal ORFs, the RAMMCAP analyses were performed using the bacterial and archaeal genetic code (-t 11 in advanced options). For identification of fungal ORFs, additional RAMMCAP analyses were performed using the standard genetic code for eukaryotes and the alternative yeast genetic code (-t 1 and –t 12 in advanced options). For comparative analysis of the metatranscriptomes to existing metagenomes from ODP Site 1229D we submitted the metatranscriptomes to MG-RAST (metagenomics.anl.gov), which were annotated according to the standard bioinformatics pipeline (http://blog.metagenomics.anl.gov/mg-rast-for-the-impatient-readme-1st/).

Taxonomic annotation of contigs Contigs were assigned to high-level taxonomic groups (Class level and above) using PhymmBL[30]. In addition to the default interpolated markov model (IMM) database (that contains only bacterial and archaeal genomes), all fungal genomes available in the NCBI RefSeq database and JGI database, along with several representative protistan and plant genomes were added to the IMM database (using the customGenomicData.pl script available with the PhymmBL download) to facilitate identification of eukaryotic contigs. Cutoffs for annotation accuracy were chosen based on the default recommendations. Taxonomic identifications of contigs made using PhymmBL[30] were integrated with the functional annotations from CAMERA (BLASTx searches against the COG database and HMMer searches against Pfam database) and the read mapping information from assemblies. This was done using several custom PERL scripts that are available from the authors upon request.

Statistical analyses Analyses of overexpression of expressed genes relative to metagenome samples was performed using the R statistical package (http://www.r-project.org/), with the MG-RAST matR library (metagenomics.anl.gov). To maintain abundance information, assembled contig sequences from each sample were uploaded to MG RAST with the read mapping abundance added to the fasta headers as specified on the MG RAST website. Statistically significant differences in overexpressed functional genes relative to genes detected in metagenomes were determined by a Kruskal-Wallis test with a p value cutoff of 0.0005. All rRNA reads were removed from both metagenomic and metatranscriptomic datasets prior to comparison. Data were normalized in MG RAST with a log based transformation:

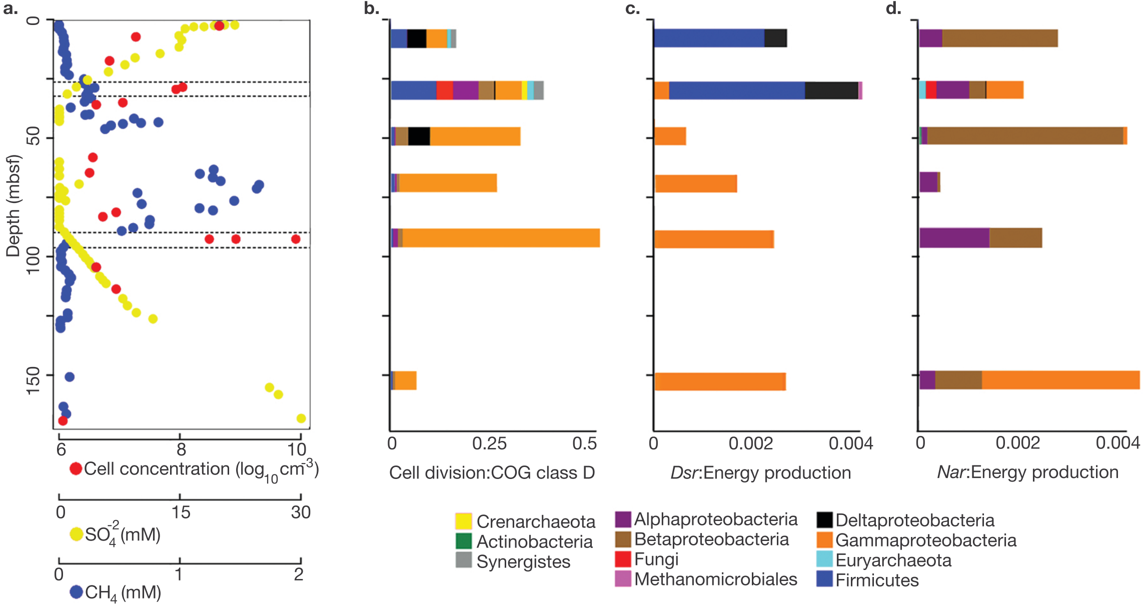
$$Y_{\mathrm{s,i}} = \log_2 (X_{\mathrm{s,i}} + 1)$$

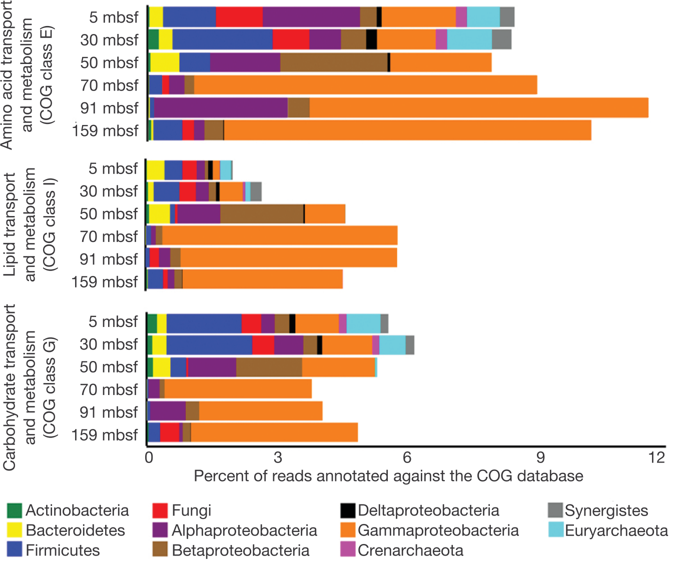
Where $X_{s,i}$ represents an abundance measure (i) in sample (s). Log transformed counts from each sample were then standardized (data centering) according to the following equation:

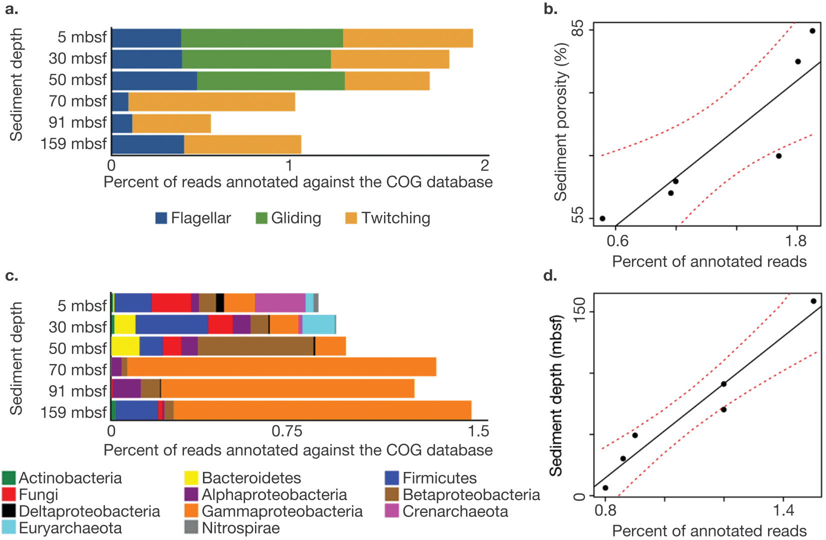
$$Z_{\mathrm{s,i}} = [(Y_{\mathrm{s,i}} - Y_{\mathrm{s}})/ \sigma_{\mathrm{s}})$$

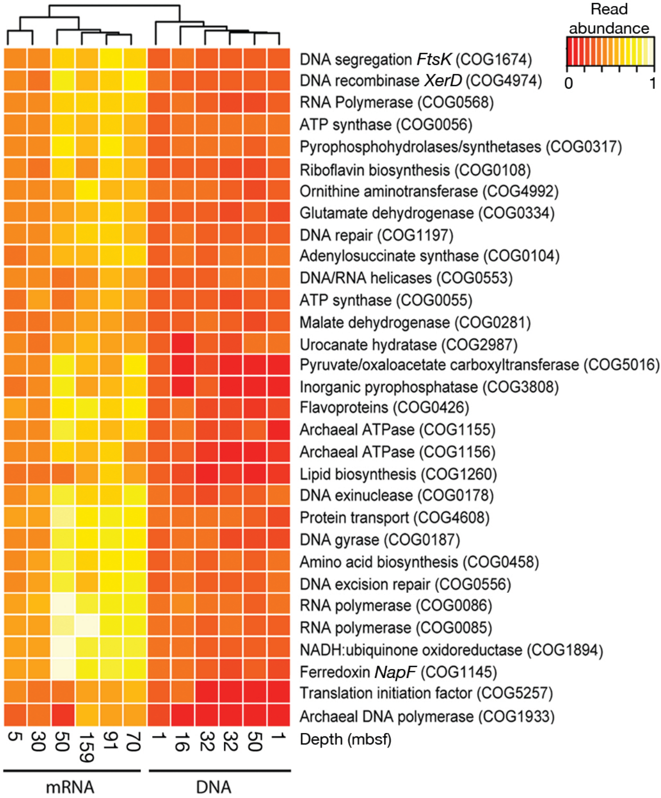
Where $Z_{\mathrm{s,i}}$ is the standardized abundance of an individual measure $Y_{\mathrm{s,i}}$ (log transformed from previous equation). From each log transformed measure of (i) in sample (s), the mean of all transformed values (Y_{s}) is subtracted and the difference is divided by the standard deviation (σ_{s}) of all log-transformed values for the given sample. After log transformation and standardization, the values for the functional categories within each sample were scaled from 0 (minimum value of all samples) to 1 (maximum value of all samples), which is a uniform scaling that does not affect the relative differences of values within a single sample or between 2 or more samples.

This procedure places the value of functional categories (*i.e*. COG categories) from each sample on a scale from 0 to 1 and was used to produce figures (*i.e*. heatmaps or principal component analysis) where the abundance range is on a scale from 0 to 1 (*i.e*. Figure 4). Normalized data that passed the Kruskal-Wallis test (p value cutoff criterion 0.0005) were used as input for heatmap presentation, UPGMA hierarchical clustering, and principal component analysis in R, using the matR package (metagenomics.anl.gov). Analysis of similarity (ANOSIM) analyses were performed on the normalized data in R, using the vegan package (http://vegan.r-forge.r-project.org/). ANOSIM was performed with 999 permutations using a Bray-Curtis distance metric. Correlations of gene expression data with geochemical and geophysical metadata were performed using the lm and predict commands in R, which are used to fit linear models to relationships between two different variables. The data for these analyses were normalized in the same fashion as Figures 1, 2, 3, S3, S4, S5, S6 and S8 (*i.e*. the relative abundance, per sample, of transcripts mapping to ORFs that were annotated to each functional COG category).


a.
b.
c.
d.
Depth (mbsf)
0
50
100
150
6
8
10
● Cell concentration ($\log_{10}$ cm^{-3})
0
15
30
● SO_4^{-2} (mM)
0
1
2
● CH_4 (mM)
0
0.25
0.5
Cell division:COG class D
0
0.002
0.004
Dsr:Energy production
0
0.002
0.004
Nar:Energy production
Crenarchaeota
Actinobacteria
Synergistes
Alphaproteobacteria
Betaproteobacteria
Fungi
Methanomicrobiales
Deltaproteobacteria
Gammaproteobacteria
Euryarchaeota
Firmicutes


Amino acid transport and metabolism (COG class E)
5 mbsf
30 mbsf
50 mbsf
70 mbsf
91 mbsf
159 mbsf
Lipid transport and metabolism (COG class I)
5 mbsf
30 mbsf
50 mbsf
70 mbsf
91 mbsf
159 mbsf
Carbohydrate transport and metabolism (COG class G)
5 mbsf
30 mbsf
50 mbsf
70 mbsf
91 mbsf
159 mbsf
0
3
6
9
12
Percent of reads annotated against the COG database
Actinobacteria
Bacteroidetes
Firmicutes
Fungi
Alphaproteobacteria
Betaproteobacteria
Deltaproteobacteria
Gammaproteobacteria
Crenarchaeota
Synergistes
Euryarchaeota


a.
Sediment depth
5 mbsf
30 mbsf
50 mbsf
70 mbsf
91 mbsf
159 mbsf
0
1
2
Percent of reads annotated against the COG database
Flagellar
Gliding
Twitching
b.
Sediment porosity (%)
85
55
0.6
1.8
Percent of annotated reads
c.
Sediment depth
5 mbsf
30 mbsf
50 mbsf
70 mbsf
91 mbsf
159 mbsf
0
0.75
1.5
Percent of reads annotated against the COG database
Actinobacteria
Bacteroidetes
Firmicutes
Fungi
Alphaproteobacteria
Betaproteobacteria
Deltaproteobacteria
Gammaproteobacteria
Crenarchaeota
Euryarchaeota
Nitrospirae
d.
Sediment depth (mbsf)
150
0
0.8
1.4
Percent of annotated reads


Read abundance
0
1
DNA segregation *FtsK* (COG1674)
DNA recombinase *XerD* (COG4974)
RNA Polymerase (COG0568)
ATP synthase (COG0056)
Pyrophosphohydrolases/synthetases (COG0317)
Riboflavin biosynthesis (COG0108)
Ornithine aminotransferase (COG4992)
Glutamate dehydrogenase (COG0334)
DNA repair (COG1197)
Adenylosuccinate synthase (COG0104)
DNA/RNA helicases (COG0553)
ATP synthase (COG0055)
Malate dehydrogenase (COG0281)
Urocanate hydratase (COG2987)
Pyruvate/oxaloacetate carboxyltransferase (COG5016)
Inorganic pyrophosphatase (COG3808)
Flavoproteins (COG0426)
Archaeal ATPase (COG1155)
Archaeal ATPase (COG1156)
Lipid biosynthesis (COG1260)
DNA exinuclease (COG0178)
Protein transport (COG4608)
DNA gyrase (COG0187)
Amino acid biosynthesis (COG0458)
DNA excision repair (COG0556)
RNA polymerase (COG0086)
RNA polymerase (COG0085)
NADH:ubiquinone oxidoreductase (COG1894)
Ferredoxin *NapF* (COG1145)
Translation initiation factor (COG5257)
Archaeal DNA polymerase (COG1933)
5
30
50
159
91
70
1
16
32
32
50
1
Depth (mbsf)
mRNA
DNA