

MIT Open Access Articles

Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the North Pacific Subtropical Gyre

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Shi, Yanmei et al. "Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean." The ISME Journal 5.6 (2010): 999-1013.

As Published: <http://dx.doi.org/10.1038/ismej.2010.189>

Publisher: Nature Publishing Group

Persistent URL: <http://hdl.handle.net/1721.1/69640>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike 3.0



**Integrated metatranscriptomic and metagenomic analyses of stratified microbial
assemblages in the North Pacific Subtropical Gyre**

(Running Title: Depth profile of microbial metatranscriptomics in the open ocean)

Yanmei Shi¹, Gene W. Tyson^{1,3}, John M. Eppley¹, Edward F. DeLong^{1,2,*}

¹Departments of Civil and Environmental Engineeringa and ²Biological Engineering,

Massachusetts Institute of Technology, Cambridge MA 02139

³Advanced Water Management Centre, University of Queensland, Brisbane, Queensland,

Australia

*Corresponding Author

Mailing address: Massachusetts Institute of Technology

15 Vassar Street

Cambridge, MA 02139

Tel: (617) 253-0252

1 **Abstract**

2 As part of an ongoing survey of microbial community gene expression in the
3 ocean, we sequenced and compared a total of ~38 Mbp of community transcriptomes
4 and ~157 Mbp of community genomes from four bacterioplankton samples, along a
5 defined depth profile at Station ALOHA in North Pacific subtropical gyre (NPSG).
6 Taxonomic analysis based on rRNA (as well as protein-coding sequences) suggested that
7 the samples were dominated by three taxa: Prochlorales, Consistiales, and Cenarchaeales,
8 that comprised 36-69% and 29-63% of the annotated sequences in the four DNA and four
9 cDNA libraries, respectively. The remaining of the sequences represented a broad
10 diversity of low abundance taxa (33 taxonomic groups detected with relative abundance
11 of $\geq 1\%$ in any of the eight libraries). The relative abundance of major taxonomic groups
12 was sometimes inconsistently represented in the DNA and cDNA libraries. For example,
13 the 125m sample genomic library was dominated by *Pelagibacter* (~36% of sequence
14 reads), which contributed far fewer sequences to the community transcript pool (~11%).
15 Those data suggest the utility of metatranscriptomics for assessing the relative
16 transcriptional activities per cell for different taxonomic groups. Functional
17 characterization in combination with taxonomic classification for highly expressed genes
18 revealed taxon-specific contributions to active biogeochemical processes such as
19 phototrophy and nitrogen metabolism. Examples included *Roseobacter*-relatives involved
20 in aerobic anoxygenic phototrophy at 75m, and the unexpected contribution of ammonia
21 oxidation by low abundance crenarchaeal population at 125m. Recruitment of DNA and
22 cDNA reads to reference microbial genomes indicated depth-specific partition of
23 coexisting microbial populations, as highlighted by the transcriptionally active HL-like

1 *Prochlorococcus* population in the bottom of the photic zone. Transcripts that mapped to
2 *Pelagibacter* genomes suggested that nutrient uptake genes dominated *Pelagibacter*
3 transcriptomes, with apparent enrichment for certain transporter types (e.g., the C4-
4 dicarboxylate transport system) than others (e.g., phosphorus transporters). Collectively,
5 the data support the utility of coupled DNA and cDNA analyses for describing the
6 taxonomic and functional attributes of microbial communities in their natural habitats.

7 **Keywords:** metatranscriptomics; metagenomics; bacterioplankton samples;
8 biogeochemical processes

9

10 **Introduction**

11 Marine microbial communities, centrally involved in the fluxes of matter and
12 energy in the global oceans, are major drivers of global biogeochemical cycling (Arrigo
13 2005, Karl and Lukas 1996). Our knowledge of abundance, diversity and gene content of
14 planktonic microbes has been fundamentally advanced over the past three decades, by
15 both model organism-based studies (Coleman and Chisholm 2007, Giovannoni et al
16 2005b), as well as metagenomic surveys of natural microbial communities (DeLong et al
17 2006, Dinsdale et al 2008, Rusch et al 2007). In particular, metagenomic comparisons of
18 distinct microbiomes (DeLong et al 2006, Dinsdale et al 2008) have revealed habitat-
19 dependent distribution of taxons and gene families, likely shaped by the biogeochemical
20 conditions of each environment. Clearly, determining if and how such genomic variations
21 are manifested at the level of gene expression and regulation represents another critical
22 step towards understanding the interplay between microbes and their natural
23 environment, as well as their metabolic strategies to exploit distinct ecological niches.

Metatranscriptomics involves the direct sampling and sequencing of gene transcripts from natural microbial assemblages, and provides quantitative assessment of microbial gene expression, without requiring *a priori* knowledge of community taxonomic and genomic compositions. We first carried out a pilot metatranscriptomic study at the Hawaii Ocean Time-series (HOT) Station ALOHA (Frias-Lopez et al 2008), where community transcripts were analyzed in parallel with genomic sequences for a bacterioplankton assemblage at 75m depth (within the mixed layer). One unexpected finding from that study was that many highly abundant transcripts (most of which were designated as hypothetical genes) were absent or in low abundance in the coupled DNA library, suggesting they originated from low abundance microorganisms (or less frequently represented genes in hypervariable genomic regions). Subsequently, comparative analyses of surface water samples have shed light on the day/night and geographical differences in community gene expression (Hewson et al 2010, Poretsky et al 2009). More recently, to effectively enhance sequencing coverage across the functional transcript pool, Stewart *et al* developed a universal rRNA-subtraction protocol that was shown to physically remove large amount of rRNA molecules from RNA samples, reducing rRNA transcript abundance by 40-58% (Stewart et al 2010). The implications of these metatranscriptomic studies are clear: although the sequencing of microbial community transcripts has just begun and is far from comprehensive, it complements the metagenomic approach and has already yielded valuable information on the active components of microbial genomes.

Here we analyze coupled metatranscriptomic and metagenomic data from four bacterioplankton samples taken at Station ALOHA, along the stratified water column

characteristic of warm, nutrient-depleted surface waters underlain by a steep pycnocline and nutricline (Dore and Karl 1996, Karl and Lukas 1996). The goal was to assess in parallel microbial metabolic potential (in DNA) and functional gene expression (in cDNA) along the vertical gradient. In addition to the recent use of these data sets to search and compare putatively novel RNA regulatory elements (small RNAs) highly abundant in these habitats (Shi et al 2009), the results here demonstrate that coupled metagenomic and metatranscriptomic analyses provide useful perspectives on microbial activity, biogeochemical potential, and regulation in indigenous microbial populations.

Methods

Sample Collection. Bacterioplankton samples (size fraction 0.22 μm – 1.6 mm) from the photic zone (25m, 75m, 125m) and the mesopelagic zone (500m) were collected from the Hawaii Ocean Time-series (HOT) Station ALOHA site in March 2006, as described previously (Shi et al 2009). See Supplementary Methods for further details on the seawater collection and RNA/DNA extraction.

Complementary DNA (cDNA) synthesis and sequencing. The synthesis of microbial community cDNA from small amounts of mixed-population microbial RNA was performed as previously described (Frias-Lopez et al 2008). Briefly, ~100 ng of total RNA was amplified using MessageAmp II (Ambion, Foster City CA) following the manufacturer's instructions and substituting the T7-BpmI-(dT)¹⁶VN oligo in place of the oligo(dT) supplied with the kit. The SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen) was used to convert amplified RNA to microgram quantities of cDNA, which was then digested with BpmI to remove poly(A) tails. Purified cDNA was then

1 directly sequenced by pyrosequencing (GS20). See Supplementary Methods for further
2 details.

3 **Bioinformatic analyses.** Ribosomal RNA sequences were first identified by
4 comparing the data sets to a combined 5S, 16S, 18S, 23S, and 28S rRNA database
5 derived from available microbial genomes and sequences from the ARB SILVA LSU and
6 SSU databases (www.arb-silva.de). 16S rRNA reads were further selected and subjected
7 to taxonomic classification. Non-rRNA sequences were compared to NCBI-nr, SEED,
8 and GOS protein clusters databases using BLASTX for functional gene analyses as
9 previously described (Frias-Lopez et al 2008, Shi et al 2009). Two custom databases (one
10 nucleotide and one amino acid) were constructed from then publicly available 2067
11 microbial genome sequences, and were used to recruit cDNA and DNA reads. See
12 Supplementary Methods for further details.

13 **Data deposit.** The nucleotide sequences are available from the NCBI Sequence
14 Read Archive under with accession numbers SRA007802.3, SRA000263, SRA007804.3
15 and SRA007806.3 corresponding to cDNA sequences, and SRA007801.5, SRA000262,
16 SRA007803.3 and SRA007805.4 corresponding to DNA sequences, for 25 m, 75 m,
17 125 m and 500 m samples, respectively.

18

19 **Results and Discussion**

20 **Bacterioplankton samples and pyrosequencing data sets**

21 The four sampling depths represent discrete zones in the water column at Station
22 ALOHA (22°45' N, 158°W), which includes the middle of the mixed layer (25m), the

1 base of the mixed layer (75m), the deep chlorophyll maximum (DCM, 125m) at the top
2 of the nutricline, and the upper mesopelagic zone (500m). On cruise HOT179,
3 bacterioplankton samples were collected from each depth for RNA and DNA extraction
4 and sequencing. Since the sampling times for these four sets of seawater samples were
5 different (25m at 22:00 local time, 75m at 03:00, 125m at 06:00, and 500m at 06:00), we
6 expected that the observed gene expression patterns would reflect spatial geochemical
7 gradients (Supplementary Figure S1), as well as temporal differences (discussed below).

8 A total of ~38 Mbp and ~157 Mbp of sequences were obtained for the four
9 metatranscriptomic and four metagenomic data sets, respectively (Table 1). The number
10 of cDNA reads per GS20 run is roughly a quarter of that of the DNA reads, likely due to
11 incomplete removal of poly(A) tags added during RNA amplification step (Frias-Lopez
12 et al 2008). (Subsequent to the work reported here, significant improvements have been
13 made in the cDNA preparing and sequencing protocols, using the GS-FLX platform
14 (Stewart et al 2010)). Nevertheless, these earlier datasets reported here represent the first
15 set of coupled metagenomic and metatranscriptomic datasets, and provide new
16 information of gene expression in parallel with community structure, gene abundance,
17 and genetic variation.

18 19 **Taxonomic composition: ribosomal RNA (rRNA) sequence-based analyses**

20 Roughly 0.3% of total DNA reads were designated as rRNA operon sequences
21 (1188, 1117, 954, and 1029 reads for the 25m, 75m, 125m, and 500m samples,
22 respectively), including bacterial, archaeal, and eukaryotic small and large subunit
23 rRNAs, and intergenic spacer sequences. This sampling frequency was within the

1 expected range based on the rRNA operon size (~5,000 bp), assuming average genome
2 size of ~2 Mbp for marine bacteria and archaea. To assess the taxonomic diversity within
3 the four microbial communities, we classified these 16S rRNA gene sequences (Figure 1,
4 upper panel), using the online Greengenes alignment and classification tools
5 (<http://greengenes.lbl.gov/cgi-bin/nph-classify.cgi>) (DeSantis et al 2006), which was
6 reported to yield the highest accuracy for assigning taxonomy to short pyrosequencing
7 reads compared to other methods such as RDP classifier or BLAST (Liu et al 2008).
8 These taxonomic assignments were further corroborated (Supplementary Figure S2;
9 Pearson's correlation > 0.95 for all four depths) using a full set of "shotgun" DNA library
10 sequences (average read length 565 bp) from the same source DNA samples (Martinez et
11 al 2010).

12 Each of the four microbial communities was dominated by two or three major
13 groups (Figure 1, upper panel). Consistiales (predominantly *Pelagibacter*) recruited ~13-
14 35% of the total classified 16S rRNA gene reads from all depths, supporting the high
15 abundance of *Pelagibacter* populations throughout the water column (Eiler et al 2009)
16 and their under-representation in large-insert metagenomic libraries, at least for the
17 populations residing shallower depths (Pham et al 2008, Temperton et al 2009). The other
18 major groups included Prochlorales in the photic zone (~17-51%), Cenarchaeales (~22%)
19 and the uncultured delta-proteobacterial group SVA0853 (~9%) at 500m, and
20 Acidimicrobidae (~2-8%) at all depths. This depth distribution was generally consistent
21 with previous cultivation-independent surveys at this site, but variability (likely both
22 biological and methodological) was apparent. For instance, a fosmid library-based survey
23 (DeLong et al 2006) reported a significant decrease in the relative abundance of

1 *Prochlorococcus* populations at 75m depth, potentially caused by cyanophage infection,
2 as suggested by the large number of cyanophage sequences recovered in the same cellular
3 size fraction. In contrast, in this survey large numbers of phage sequences were not
4 detected, and *Prochlorococcus* relative abundance peaked at 75 m depth, regardless of
5 DNA library type and sequencing method (pyrosequencing, Figure 1; fosmid clone
6 library, Table S1).

7

8 **Taxonomic composition: Protein-coding sequence-based analyses**

9 Another common approach to assess taxonomic composition from metagenomic
10 data sets is to infer taxonomic origins from open reading frame (ORF) sequences (Huson
11 et al 2007). Here, we observed both consistencies as well as some discrepancies when
12 comparing the community composition derived from rRNA gene sequences (discussed
13 above) to those derived from ORF sequences using MEGAN (Huson et al 2007). As seen
14 in Figure 1 and Supplementary Figure S3, *Pelagibacter* relative abundance decreased
15 from ~13-35% estimated from the 16S rRNA gene sequences, to ~9-23% from the ORF
16 sequences, and the uncultured delta-proteobacterium SVA0853 was completely missed in
17 the latter. In contrast, *Prochlorococcus*-like sequences represented ~39-71% of all
18 annotated ORF sequences, much higher than that estimated from 16S rRNA gene
19 sequences (~17-51%). Higher representation of *Prochlorococcus*-like mRNA transcripts
20 relative to their cell abundance was noted by Poretsky *et al* in metatranscriptomic data
21 sets from day and night samples from the same site, and was attributed to higher
22 transcriptional activities of *Prochlorococcus* cells relative to coexisting heterotrophic
23 microbes (Poretsky et al 2009). However, it appears that differences in transcriptional

activities may not be the explanation, since our DNA data sets showed the same trend of overrepresentation of *Prochlorococcus*-related ORF sequences. Assuming similar genome sizes, a more likely explanation is that the higher representation of *Prochlorococcus*-derived sequences reflects the uneven representation of taxa in current databases. That is, sequence annotation is biased in favor of taxa with more sequenced isolates, such as *Prochlorococcus*, than those with fewer or no sequenced isolates such as *Pelagibacter* and SVA0853-related delta-proteobacteria.

Taxonomic origin of transcripts in the cDNA samples

The simultaneous recovery of rRNA and mRNA transcripts from RNA samples provided a unique opportunity to assess the contribution of each taxon to the community metabolic processes (as judged by transcript abundance). We performed taxonomic analyses with the 16S rRNA as well as protein-coding mRNA transcript sequences exactly as described above for DNA samples (Figure 1, lower panel; Supplementary Figure S3, lower panel). *Prochlorococcus* populations inhabiting DCM layer (125m) displayed highest transcriptional activity, relative to their abundance at that depth. In contrast, *Pelagibacter*, the most numerically abundant heterotrophic bacteria in the open ocean, appeared to be relatively more abundant in cell numbers but less active transcriptionally within DCM layer (also evident in the *Pelagibacter* genome-wide gene expression analysis below). The DCM layer is characterized by two opposing resource gradients: light supplied from above and nutrients supplied from below, and thus co-existing photoautotrophic and heterotrophic microbes might alternate dominance at different times of a day or in different seasons of a year. Specifically, this apparently

1 lower transcriptional activity of *Pelagibacter* may be influenced by the time of DCM
2 sample collection: ~6AM local time, when photosynthetic microorganisms such as
3 *Prochlorococcus* may be relatively more active.

4 Finally, for the relatively under-studied mesopelagic zone (500m), two
5 observations are clear. Marine group I crenarchaeota and *Pelagibacter* constitute a major
6 fraction of microbial community both by abundance and metabolic activity. Meanwhile,
7 groups in lower abundance such as *Alteromonadales* and *Sphingomonadales* showed a
8 dramatically higher transcript per gene ratio, suggesting that these groups exhibit higher
9 transcriptional activity than expected based on their DNA abundance.

11 **Global analysis of metabolic potential and functional activities**

12 The majority of the non-rRNA cDNA reads (> 50%), especially those derived
13 from the 500m sample (> 70%), did not share any significant match against NCBI non-
14 redundant (NCBI nr) and the SEED (Meyer et al 2008) databases (Table 1). Not
15 surprisingly, a significantly higher fraction of cDNA reads shared homology to sequences
16 in the Global Ocean Sampling (GOS) peptide database, the largest marine-specific
17 sequence database available (Yooseph et al 2007). Furthermore, a large fraction of these
18 cDNA sequences were not present in the coupled DNA libraries at the current sequencing
19 depth (data not shown). These novel sequences likely represented actively expressed
20 ORFs from low abundance microbial groups (alternatively, hyperdynamic genomic
21 regions of well known taxa), or noncoding regions that by definition are not translated
22 into proteins but instead function as RNA molecules (Shi et al 2009).

23 For sequences that were annotated as protein coding, we compared gene and

transcript abundance in parallel, in order to investigate gene expression in a normalized fashion (see Supplementary Methods). Such normalization accounts for differences in community structure and gene content among samples, allowing detection of metabolic pathways and gene families in lower abundance but with relatively high transcriptional activity (see the example of crenarchaeal-mediated ammonia oxidation at 125m below).

Known metabolic pathways. Several metabolic pathways exhibited high expression levels, as evidenced by a number of SEED subsystems that were found significantly enriched (at the 98% confidence level) in each transcript library, relative to the corresponding DNA library (Figure 2; Table 2). In the surface sample (25m) collected at 22:00 local time, the active expression of oxidative stress-related genes was likely a result of high UV doses during daytime. Aerobic respiration, expected to be enriched relative to photosynthesis at night, was reflected in the expression of cytochrome c oxidases and menaquinone-cytochrome c reductase complexes. The sample collected from DCM layer (125m) at 6:00 AM local time, exhibited high abundance of transcripts associated with carbon fixation and photosynthesis, compared with the other two photic zone samples (despite the relatively lower abundance of photosynthetic genes in the DNA, see Table 2). This is consistent with laboratory observations where *Prochlorococcus* carbon fixation genes were maximally expressed at dawn, and photosynthetic gene expression was elevated upon the appearance of light (Zinser et al 2009). Highly expressed subsystems in the mesopelagic sample (500m) included peptidoglycan biosynthesis that may be involved in maintenance of cell wall integrity at greater depths, and ammonia assimilation that plays a significant role in energy metabolism for mesopelagic crenarchaeota (Konneke et al 2005).

1 Not surprisingly, light-harvesting cellular subsystems were among the most
2 highly expressed in the photic zone. The differentiated clustering of photic zone DNA
3 and cDNA samples observed (Figure 2; Supplementary Figure 5) may be partly
4 attributable to sampling times, given the commonality of diel rhythms among
5 photosynthetic microbes (Zinser et al 2009). As expected, the metabolic signatures of
6 mesopelagic communities suggested completely different modalities, including energy
7 sources, cellular structures, catabolic and anabolic biochemical pathways.

8 **GOS protein families.** The recent global ocean sampling (GOS) expedition
9 (Rusch et al 2007, Yooseph et al 2007) has greatly expanded our knowledge of open
10 ocean-derived protein families. Among all protein families identified based on sequence
11 similarity clustering, 3,995 protein clusters consisted of only GOS sequences, 1,700 of
12 which have no detectable homology to previously known protein families (Yooseph et al
13 2007). Many of these GOS-only protein clusters of unknown functions were detected in
14 our transcript libraries, some in high abundance (Figure. 3A), underscoring ecologically
15 relevant functions associated with these novel/hypothetical protein families. Meanwhile,
16 analysis of protein families with known or predicted functions highlighted genes that are
17 highly expressed and therefore likely play active roles in maintaining ecosystem
18 functions at each habitat (Figure 3B).

19 Nitrogen metabolism protein families

20 A suite of nitrogen metabolism genes (ammonium transporter, *amt*; dissimilatory
21 nitrite reductase, *nirK*; urea transporter, *urt*; ammonia monooxygenase subunits,
22 *amoABC*) was among the most highly expressed of GOS protein families detected
23 (Figure 3B). An essential macronutrient, nitrogen availability and turnover limits

1 biological production in many open ocean regions, including NPSG (Van Mooy and
2 Devol 2008). Ammonia/ammonium is a key reduced nitrogen compound that can either
3 be incorporated into carbon skeleton via the glutamine synthetase (GS; *glnA*)/glutamate
4 synthase (GOGAT; *glsF*) cycle, or can serve as energy source fueling autotrophic
5 metabolism (Konneke et al 2005). Thus, the transport of ammonia/ammonium is vital to
6 planktonic microbes living in the nutrient deplete surface waters and energy constrained
7 deep waters in an open ocean setting. Urea is another potentially important nitrogen
8 source in the ocean, and is utilized by marine cyanobacteria (Moore et al 2002). The
9 more oxidized forms of nitrogen, nitrite and nitrate require more metabolic energy to
10 utilize but can serve as alternative nitrogen sources because of their much higher
11 concentrations in deep euphotic zone and mesoplegic zone below the nitracline.

12 To assess the prevalent nitrogen utilizing pathways in the genomes of the most
13 abundant planktonic microbial populations, we compared the observed frequency
14 (normalized to gene length and data set size) of several essential nitrogen metabolism
15 genes with that of the 16S rRNA gene of *Prochlorococcus* and marine group I
16 crenarchaeota. The observed frequency of *Prochlorococcus*-related *amt*, *glnA*, *urt*, urease
17 genes is equivalent to that of *Prochlorococcus* 16S rRNA gene (Supplementary Figure
18 S4A, left panel), suggesting that ammonium and urea assimilation is preserved in
19 naturally occurring *Prochlorococcus* populations. In contrast, the assimilatory nitrite
20 reductase gene (*nirA*) was present in only a small fraction of *Prochlorococcus* cells (c.a.,
21 7%, 8% and 15% at 25m, 75m, and 125m, respectively), consistent with expectation
22 based on genomic and physiological studies of *Prochlorococcus* isolates (Moore et al
23 2002, Rocap et al 2003). Furthermore, the transcripts of these nitrogen metabolism genes

(except *nirA*) were also detected in our metatranscriptomic data sets (Supplementary Figure S4A, right panel), suggesting active deployment of these nitrogen metabolism pathways by *Prochlorococcus* cells *in situ*. The *amt* gene was the most actively transcribed, likely an adaptive mechanism to efficiently scavenge low-concentration ammonium as the most preferred nitrogen source. The dramatic decrease in *amt* gene expression at 125m however, was not expected. It is possible that the apparently higher primary production at 125m (DCM) has caused accumulation of ammonium via active nutrient regeneration processes. In fact, ammonium maxima near the DCM layer are common in stratified oligotrophic waters (Brzezinski 1988). As a result, the presumably elevated ammonium concentration may result in down-regulation of the *amt* gene expression, as observed in many cyanobacteria isolates.

Marine group I crenarchaeota exist in high abundance in mesopelagic zone, where distinct forms and concentrations of nitrogen species (e.g., nitrate, nitrite, urea) are present. *Nitrosopumilus maritimus*, an isolate of related crenarchaea from marine aquarium, has been shown definitively to grow chemolithoautotrophically on ammonia (Konneke et al 2005). Further genomic analyses of marine group I crenarchaeota have provided insights into the metabolism of other forms of nitrogen compounds (Hallam et al 2006, Walker et al 2010). Here, our data showed that *amt*, *amoABC*, and *glnA* genes were prevalent and expressed in planktonic crenarchaeal populations, whereas urea utilization genes, while present and expressed, appeared in lower abundance (Supplementary Figure S4B, left panel). Clearly, despite the apparent lack of such genes in the *N. maritimus* genome (Walker et al 2010), a fraction of planktonic crenarchaeal populations encode genes for utilizing urea as nutrient or energy source. The normalized

1 expression levels of crenarchaea-related *amt* and *amoABC* genes (especially *amoC* gene)
2 was among the highest in our data sets (orders of magnitude higher than most other
3 protein-coding genes) (Figure 3B). Interestingly, the anomalously high *amoC* gene
4 expression appeared to be universal, as also observed in bacterial nitrifiers (Berube et al
5 2007), for as-yet unknown reasons. Consistent with a quantitative PCR-based study
6 (Church et al 2010), the *amoABC* transcripts were detected in high abundance at 125m
7 depth despite the small planktonic crenarchaeal population size (Supplementary Figure
8 S4B, right panel). Together with previous report of remarkably high substrate affinity and
9 kinetics of crenarchaeal *amo* genes (Martens-Habben et al 2009), these data further
10 support a role for marine crenarchaea in nitrification in the ocean via active ammonia
11 oxidation.

12 Nitrite, an end product of archaeal ammonia oxidation, could exert toxic effects to
13 cells if accumulated, and an upper primary nitrite maximum (UPNM) is often observed
14 near DCM layer (125m in this study) in the open ocean (Dore and Karl 1996). Consistent
15 with the hypothesis that dissimilatory nitrite reductase (*nirK*) in ammonia-oxidizing
16 microbes is involved in nitrite detoxification (Casciotti and Ward 2001, Hallam et al
17 2006), *nirK* was found highly expressed at 125m (Supplementary Figure S4B, right
18 panel). Finally, nitrate reductase genes (*narH* and *narG*) and transcripts were frequently
19 detected in the 500m data sets, and appeared to be most similar to homologs found in
20 Candidatus *Kuenenia stuttgartiensis* (data not shown), suggesting that planktonic
21 crenarahaeta may not participate in the first step of nitrate respiration.

22 Photoheterotrophy

23 We detected in the photic-zone active expression of genes involved in

1 photoheterotrophy, including those encoding proteorhodopsins. Proteorhodopsin (PR) is a
2 photoprotein that functions as light-driven proton pump, generating biochemical energy
3 via proton motive force (Béjà et al 2000). PR photosystems have been detected in a large
4 percentage (up to 80%) of ocean surface-dwelling bacteria and archaea (DeLong and
5 Béjà 2010), and were suggested to be horizontally transferred among phylogenetically
6 divergent microbial taxa (Frigaard et al 2006, McCarren and DeLong 2007). Laboratory-
7 based experiments have suggested that PR photosystem increases cellular fitness to
8 bacterial cells under adverse growth conditions (Gómez-Consarnau et al 2007, Gómez-
9 Consarnau et al 2010, González et al 2008).

10 Our depth profile data allow us to directly assess the *in situ* abundance and
11 taxonomic origins of PR gene and transcripts. Abundance of PR transcripts decreased
12 dramatically from euphotic zone to 500m (in which only 4 cDNA reads shared homology
13 to known PR genes) (Supplementary Figure S5A). While PR DNA and cDNA reads
14 appeared to be originated from a diverse range of taxa, the majority shared homology to
15 known PR genes from SAR11-like organisms (Supplementary Figure S5B). Notably, PR
16 genes were found most highly expressed in the 75m sample (collected at 22:00), followed
17 by the 25m and 125m samples (collected at 3:00 and 6:00, respectively) (Supplementary
18 Figure S5A; also see the *Pelagibacter* genome-wide gene expression analysis below),
19 suggesting PR genes may be constitutively expressed in the photic zone independent of
20 light conditions. Laboratory studies of PR-containing isolates as well as a recently
21 reported microcosm experiment have reported inconsistent observations, some suggesting
22 constitutive PR expression (Giovannoni et al 2005a, Riedel et al 2010), while others
23 suggesting light-regulation of PR expression (Gómez-Consarnau et al 2007, Lami et al

2009). Higher-resolution metatranscriptomic studies are necessary to provide further insight into light effects on PR gene expression in different taxa, and in different oceanographic provinces.

Evidence for another form of phototrophy mediated by aerobic anoxygenic phototrophic (AAP) bacteria was also observed. Recent studies suggest that AAPs constitute a considerable fraction of marine planktonic community, and may contribute significantly to the carbon cycle in the ocean via facultative photoheterotrophy (Béjà et al 2002, Kolber et al 2001). Living in an oligotrophic environment, oceanic AAPs likely are capable of efficiently controlling the expression of their photosynthetic apparatus, supplementing heterotrophic metabolism with light-dependent energy harvest. In this depth profile, AAPs were most abundant in 25m and 75m samples based on observed gene frequencies of bacteriochlorophyll biosynthesis genes (*bchXYZ*), light-harvesting complex I genes (*pufAB*) and the reaction center genes (*pufLM*). The majority of these photosynthetic genes were closely related to *Roseobacter*-like AAP sequences, particularly a BAC clone insert retrieved from the Red Sea (eBACred25D05; accession number: AY671989) (Oz et al 2005). GOS protein clusters associated with these AAP genes were found highly expressed in the 75m sample (Figure 3B), and most of this AAP gene expression originated from the *puf* operon (Supplementary Figure S6). Collectively, the data indicate photosynthetically active population of AAPs, at 75m in particular.

Reference genome-centric analyses

We used a total of 2067 genomic references (including finished and draft genomes), to recruit DNA and cDNA reads at high stringency, based on BLASTN

1 comparison (see Supplementary Methods). About 29%, 40%, 15% and 7% of total DNA
2 reads, and 30%, 24%, 26%, and 18% of total cDNA reads were recruited to the reference
3 genomic data for 25m, 75m, 125m, and 500m sample, respectively. Notably, the
4 percentage of recruited cDNA reads for each sample was significantly higher than that of
5 cDNA reads that could be assigned to NCBI-nr protein database (Table 1), a result of
6 cDNA recruitment to expressed noncoding regions on the genomes. For instance, about
7 1539 reads in the 25m sample were recruited to an intergenic region of *Prochlorococcus*
8 strain MIT 9215 genome, corresponding to the Group_2 small RNA previously reported
9 by Shi *et al* (Shi et al 2009).

10 The relative representation of genomes/genome fragments is shown in a three-
11 way comparison plot, to illustrate the similarities and differences of communities
12 dwelling in specific habitats (Figure 4). For this analysis, the 75m and 125m samples
13 were pooled together, since they share similar profile at both DNA and cDNA levels
14 (Figure 2). All genomes recruiting > 50 DNA reads are also listed in Supplementary
15 Table S2. Here, general separation of photic zone populations with mesopelagic
16 populations was observed, with a few exceptions that were found more evenly distributed
17 along the depth, including the ubiquitous *Pelagibacter*, and the alphaproteobacterium
18 *Erythrobacter* sp. SD-21, a Mn(II) oxidizing bacterium that has been isolated from many
19 diverse marine environments including surface and deep oceans (Francis et al 2001).

20 Such genome recruitment analysis provides direct measurement of vertical
21 distribution of ecologically coherent populations (represented by reference genomes) in
22 nature, such as high-light (HL) and low-light (LL) adapted *Prochlorococcus* “ecotypes”
23 (Moore and Chisholm 1999). Notably, despite an expected significant increase of low-

1 light (LL) adapted *Prochlorococcus* populations (mostly eNATL2A) at 125m, where
2 light intensity dramatically decreased compared to shallower depths, > 80% of the
3 *Prochlorococcus*-like reads at 125m were most similar to sequences of high-light (HL)
4 adapted isolates (mostly eMIT9312) (Supplementary Table S2). While possibly a result
5 of physical homogenization of the water column due to deep mixing in the winter
6 (Malmstrom et al 2010), these HL-like *Prochlorococcus* cells displayed elevated
7 transcriptional activity at 125m (Supplementary Table S2), suggesting they were unlikely
8 sinking dead cells. Zinser and colleagues (Zinser et al 2006) showed that in deeper waters
9 (below 75 m) at the western North Atlantic site, a significant fraction of *Prochlorococcus*
10 population cannot be detected by qPCR probes designed to capture currently known
11 ecotypes, suggesting significant deep populations of *Prochlorococcus* yet to be identified
12 and characterized. Results here suggest the presence of a HL-like *Prochlorococcus*
13 population that may be well adapted to the lower euphotic zone, under low light
14 conditions.

15 **Population transcriptomic analysis of *Pelagibacter*.** As the most abundant
16 heterotrophic bacterial group throughout the ocean water column, *Pelagibacter* (member
17 of the alphaproteobacteria SAR11 clade) provides a useful model example for how
18 culture-based and metagenomic/metatranscriptomic data can be integrated to study the
19 ecophysiology of wild populations. Subsets of DNA and cDNA reads from all 4 depths
20 were mapped onto the reference genome of the open ocean *Pelagibacter* isolate
21 HTCC7211 (see Supplementary Methods). The expression level of annotated protein
22 coding genes provided clues on the prevailing metabolic activities of *Pelagibacter*
23 populations at each depth (Figure 5; Supplementary Table S3). Overall, the expression

1 profile of protein coding genes confirmed the observation based on the rRNA profile
2 (Figure 1), that *Pelagibacter* cells at 125m were less transcriptionally active at the time of
3 sampling, compared to their counterparts at 25m and 75m. Indeed, ribosomal proteins
4 were among the most highly expressed genes in 25m and 75m samples, and most ORFs
5 showed lower expression levels in the 125m sample.

6 Nutrient-uptake genes of *Pelagibacter*, particularly those encoding periplasmic
7 solute binding proteins of ATP-binding cassette (ABC) families, represented the most
8 abundant class of transcripts (Figure 5). The disproportionally high abundance of
9 transporter genes in *Pelagibacter* genomes is believed to contribute to their capability of
10 efficiently utilizing a broad variety of substrates (Giovannoni et al 2005b). Here we
11 observed high transcriptional levels of solute-binding proteins families 1, 3, and 7 (Figure
12 5), which involve in the uptake of sugars, polar amino acids, and organic polyanions,
13 respectively (Tam and Saier 1993). Polyamines (e.g., spermidine/putrescine), trace
14 elements (e.g., selenium), and possible osmolytes (e.g., glycine betaine) also appeared to
15 be actively transported. In addition, a few transporter families other than the ABC
16 superfamily were also expressed, including Na⁺/solute symporter (Ssf family) and
17 tripartite ATP-independent periplasmic (TRAP) dicarboxylate transporter genes for the
18 uptake of mannitol and/or C4-dicarboxylates, which relies on proton motive force rather
19 than ATP hydrolysis. Notably, different expression levels among the four depths were
20 discernible for these transporter genes, potentially a result of substrate availability and
21 preference for *Pelagibacter* populations residing different depths.

22 Sowell and colleagues have observed in *Pelagibacter* metaproteomes collected
23 from the Sargasso Sea surface water a dominant signal of periplasmic transport proteins

1 for substrates such as phosphate, amino acids, phosphonate and spermidine/putrescine
2 (Sowell et al 2008). The overall consistent observation that nutrient-uptake transporters
3 were most highly expressed both at transcriptional level (this study) and translational
4 level (Sowell et al 2008), corroborates the oligotrophic nature of both oceanic sites.
5 However, significant differences in peptide versus transcript expression levels were also
6 apparent among certain categories of transporters. For example, we did not detect gene
7 expression for phosphate and phosphonate transporter genes (*pstS* and *phnD*) related to
8 *Pelagibacter* in our data sets. In fact, no *phnD*-related sequences were detected in the
9 DNA reads recruited to the *Pelagibacter* HTCC7211 genome, suggesting *phnD* gene is
10 absent in most *Pelagibacter* cells at Station ALOHA. This observation contrasts sharply
11 with the that of Sowell *et al*, reflecting the significant biogeochemical difference between
12 the two oceanic sites (e.g., phosphate concentrations at BATS are much lower than that at
13 Station ALOHA (Wu et al 2000)). The effect of geography-dependent phosphorus
14 limitation appears to be reflected in the gene content of native *Prochlorococcus* cells
15 (Martiny et al 2009), as well as other picoplankton populations (Martinez et al 2010).

16 **HTCC7211-specific genes.** It has been well established that genomic plasticity of
17 microbes, reflected by variations in gene content of closely related strains, may facilitate
18 microbial adaptation to their natural habitats (Coleman et al 2006, Cuadros-Orellana et al
19 2007). We compared the genome sequences of two *Pelagibacter* coastal isolates (strains
20 HTCC1062 and HTCC1002) and the open ocean isolate (HTCC7211, used as reference
21 genome in the genome-centric analysis above), and asked which HTCC7211-specific
22 genes might be highly expressed and thus functionally important in the open ocean
23 environment.

There are 296 HTCC7211-specific genes (see Supplementary Methods), 154 detected in at least one of our metatranscriptomic data sets (Supplementary Figure S7). Two ORFs encoding ABC-type periplasmic solute binding proteins appeared to be specific to open ocean-dwelling *Pelagibacter*, and were highly expressed. One ORF encodes a selenium-binding protein, which may contribute to the synthesis of selenoproteins (Zhang and Gladyshev 2008). The other ORF encodes an extracellular solute-binding protein family 1, which is associated with the uptake of malto-oligosaccharides, multiple sugars, alpha-glycerol phosphate, and iron (Tam and Saier 1993). In addition, the C4-dicarboxylate transport (Dct) system, which relies on highly specific and affine extracytoplasmic solute binding receptors, appeared to be important in oceanic *Pelagibacter* populations. Not only were four *dct* operons present in the strain HTCC7211 (as opposed to apparently only one copy in coastal strains HTCC1062 and HTCC1002), but the three HTCC7211-specific *dctP* paralogues (encoding a periplasmic C4-dicarboxylate-binding protein) were also expressed (Supplementary Figure S7). Dct transporters are secondary carriers that use an electrochemical H⁺ gradient as the driving force for transport rather than ATP hydrolysis, and allow the uptake of mannitol and/or C4-dicarboxylates like succinate, fumarate, and malate, pointing to such organic compounds as important carbon and energy source for oceanic *Pelagibacter*.

Caveats and challenges

Given the complex, nonlinear relationship between gene expression, protein expression and biochemical function, the transcript profiles need to be carefully interpreted in the context of other supporting data. Transcript abundance will not always

1 correlate directly with cognate protein levels, and the kinetics that relates expression to
2 phenotype vary among different transcript classes (Steunou et al 2008). Nonetheless,
3 reasonably good correlation between transcriptomes and proteomes, especially for
4 transcripts and peptides in higher abundance, has been observed in several model
5 organisms (Corbin et al 2003, Eymann et al 2002, Scherl et al 2005). In the *Pelagibacter*
6 genome-centric analyses reported here (Figure 5), we observed considerable overlap
7 between highly abundant transcripts and the most represented peptides previously
8 reported in a SAR11-centric metaproteomic study (Sowell et al 2008). This general
9 consistency between the population transcriptomes and proteomes of the most abundant
10 and ubiquitous heterotrophic bacteria clade in the open ocean, supports the use of
11 metatranscriptomics to assess inventories of functionally relevant gene families based on
12 their expression levels.

13 The work reported here, along with previous studies (Hewson et al 2009,
14 Poretsky et al 2009), also illustrates several challenges for future metatranscriptomic
15 studies. First, due to great diversity found in most natural systems and predominant
16 transcriptional signal of the genes involved in central metabolism and protein synthesis
17 machinery, sequencing depth needs to be greatly expanded (Stewart et al 2010). Our data
18 showed that for one pyrosequencing run on an open ocean bacterioplankton sample,
19 about 66-74% of sequences with putative taxonomic assignment belonged to the top two
20 most abundant taxonomic groups (Supplementary Figure S3, lower panel). Thus, the
21 majority of the diversity of the transcript pool was represented by low abundance reads
22 with little statistical confidence, albeit these may well contain important information. As

1 a good demonstration, two technical replicate metatranscriptomic data sets were found to
2 only share 37% of the NCBI-nr reference entries, suggesting the rarefaction curve of
3 functional diversity is far from leveling off (Stewart et al 2010).

4 Another challenge is associated with the frequent observation that hypothetical
5 genes are among the most highly expressed genes in the genomes examined
6 (Supplementary Figure S8). Such hypothetical genes are potentially of great relevance to
7 the ecology of host populations in their native environment, but understanding and
8 characterizing unknown functions in these hypothetical ORFs represents a continuing
9 challenge. It has recently been reported that about two thirds of the gene families with
10 unknown functions likely represent very divergent branches of known and well-
11 characterized families (Jaroszewski et al 2009). Expression patterns in the environment,
12 combined with structure and sequence homology search, provides a starting point for
13 formulating and testing hypotheses about the biological functions of these uncharacterized
14 ORFs. As an example, four highly expressed hypothetical genes, putatively originating
15 from marine crenarchaeal genomes, were annotated to contain putative polycystic kidney
16 disease (PKD) domain (PF00801). PKD domains are mostly present on the cell surface,
17 and are involved in protein-protein interactions. Particularly, PKD domains were found
18 predominant in archaeal surface layer proteins that were thought to protect the cell from
19 extreme environments (Jing et al 2002), or in exported proteins of marine heterotrophic
20 bacteria that may be involved in the binding and degradation of extracellular polymers
21 (carbohydrate and protein) (Zhao et al 2008). The taxonomic origins and prevalence in

1 community transcriptomes of these PKD domain-containing hypothetical genes now
2 render them reasonable targets for future functional characterizations in planktonic marine
3 crenarchaea.

4

5 **Conclusions and future directions**

6 Through analysis of four coupled metagenomic and metatranscriptomic data sets,
7 we have demonstrated that microbial community transcriptomes *in situ* can be profiled
8 (for abundant microbial populations, at a reasonable coverage), and compared in the
9 context of genomic compositions and ambient environmental conditions. Our results
10 provide insight into: 1) sequence characteristics, such as the uniqueness and vast
11 diversity, of microbial community transcriptomes in the open ocean ecosystem; 2)
12 specific metabolic processes that characterize each of the four habitats investigated; 3)
13 highly expressed gene families, and their putative taxonomic breakdown; and 4)
14 population variability and physiological signals from abundant taxa of the microbial
15 assemblages, inferred via reference genome-centric analyses. Given the great complexity
16 found in the transcriptome of even small genomes (Guell et al 2009), it must be assumed
17 that we are yet scratching the surface of the dynamic, complex transcriptional network
18 orchestrated by microbe-microbe and microbe-environment interactions. Future
19 metagenomic and metatranscriptomic surveys at more highly resolved spatial and
20 temporal dimensions will help provide a more comprehensive picture of microbial
21 functional diversity in natural settings. Additionally, the application of coupled
22 metagenomics and metatranscriptomics in experimental settings will allow more

1 controlled observation of microbial community responses to environmental changes, and
2 allow simultaneous study of microbial community physiology, population and
3 community dynamics.

5 **Acknowledgements:**

6 We thank the captain and crew of the R/V Kilo Moana for facilitating sample
7 collection. Thanks also to Stephan Schuster for collaboration on pyrosequencing. We are
8 grateful to the J. Craig Venter Institute, and the Gordon and Betty Moore Foundation for
9 the microbial genome sequences. This work was supported by the Gordon and Betty
10 Moore Foundation, National Science Foundation Microbial Observatory Award MCB-
11 0348001, the Department of Energy Genomics GTL Program, and the Department of
12 Energy Microbial Genomics Program, and an NSF Science and Technology award, C-
13 MORE.

15 **Figure Legends**

17 **Figure 1. Taxonomic classification based on 16S rRNA-bearing reads in DNA**
18 **and cDNA data sets.** Taxonomic assignments were binned at the Order level, using the
19 Hugenholtz taxonomy of Greengenes (see Supplementary Methods). 16S rRNA
20 sequences that could not be classified were excluded from the analysis. Y axis scale
21 represents the percentage of the total classified 16S rRNA reads. Only taxa that
22 represented $\geq 1\%$ of all classified reads are displayed.

1

2 **Figure 2. Clustering of all cDNA and DNA data sets based on relative**
3 **abundance of SEED subsystems.** Only the most abundant subsystems that together
4 recruited 95% of all reads are displayed. Hierarchical clustering of 4 DNA and 4 cDNA
5 samples were performed with euclidean distance and single linkage method using
6 MATLAB. Color scale represents the proportion of reads assigned to SEED categories
7 relative to the total library size in each sample. Blue to red color indicates low to high
8 representation of SEED categories.

9

10 **Figure 3. Community-level gene expression profiles based on the GOS**
11 **protein family database.** Cluster-based expression ratio was defined as representation of
12 each GOS cluster in the cDNA library normalized by its representation in the DNA
13 library. GOS clusters that recruited only cDNA reads were arbitrarily set a value of 1
14 copy of DNA read, to avoid a denominator of 0. (A) GOS clusters were ranked by their
15 cluster-based expression ratios for four depths; (B) The most highly expressed GOS
16 clusters with known or predicted functions were highlighted for each depth.

17

18 **Figure 4. Three-way comparison of representation of genomes and genome**
19 **fragments (fully sequenced fosmids) in DNA and cDNA data sets.** The 75m and 125m
20 data sets were combined since they were the most similar. Each dot represents a genome
21 (fragment), and its proximity to a vertex reflects the enrichment of the corresponding
22 genome (fragment) in the respective sample. Only genomes recruited > 0.1% of total

reads are displayed. Station ALOHA fosmids represent fosmid sequences that were reported by DeLong et al (DeLong et al 2006). See Supplementary Methods for detail.

Figure 5. Genome-wide expression profiles of *Pelagibacter*-related populations, in all four depths. X-axis shows the arbitrary numbering of ORFs along the genome of *Pelagibacter* strain HTCC7211. Y-axis scale represents normalized cDNA to DNA ratio (normalized expression level; see Supplementary Methods) for each ORF. Each colored circle in the stem plot represents a given ORF at a given depth.

REFERENCES

- Arrigo KR (2005). Marine microorganisms and global nutrient cycles. *Nature* **437**: 349-355.
- Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al* (2000). Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1906.
- Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T *et al* (2002). Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**: 630-633.
- Berube PM, Samudrala R, Stahl DA (2007). Transcription of all amoC copies is associated with recovery of *Nitrosomonas europaea* from ammonia starvation. *J Bacteriol* **189**: 3935-3944.
- Brzezinski MA (1988). Vertical-Distribution of Ammonium in Stratified Oligotrophic Waters. *Limnol Oceanogr* **33**: 1176-1182.
- Casciotti KL, Ward BB (2001). Dissimilatory nitrite reductase genes from autotrophic ammonia-oxidizing bacteria. *Appl Environ Microbiol* **67**: 2213-2221.
- Church MJ, Wai B, Karl DM, DeLong EF (2010). Abundances of crenarchaeal amoA genes and transcripts in the Pacific Ocean. *Environ Microbiol* **12**: 679-688.

1 Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF *et al* (2006).
 2 Genomic islands and the ecology and evolution of Prochlorococcus. *Science* **311**: 1768-
 3 1770.
 4
 5 Coleman ML, Chisholm SW (2007). Code and context: Prochlorococcus as a model for
 6 cross-scale biology. *Trends Microbiol* **15**: 398-407.
 7
 8 Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE *et al* (2003). Toward a
 9 protein profile of Escherichia coli: Comparison to its transcription profile. *Proc Natl*
 10 *Acad Sci USA* **100**: 9232-9237.
 11
 12 Cuadros-Orellana S, Martin-Cuadrado AB, Legault B, D'Auria G, Zhaxybayeva O, Papke
 13 RT *et al* (2007). Genomic plasticity in prokaryotes: the case of the square haloarchaeon.
 14 *Isme Journal* **1**: 235-245.
 15
 16 DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al* (2006).
 17 Community genomics among stratified microbial assemblages in the ocean's interior.
 18 *Science* **311**: 496-503.
 19
 20 DeLong EF, Béjà O (2010). The Light-Driven Proton Pump Proteorhodopsin Enhances
 21 Bacterial Survival during Tough Times. *PLoS Biol* **8**: e1000359.
 22
 23 DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al* (2006).
 24 Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible
 25 with ARB. *Appl Environ Microbiol* **72**: 5069-5072.
 26
 27 Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM *et al* (2008).
 28 Functional metagenomic profiling of nine biomes. *Nature* **452**: 629-632.
 29
 30 Dore JE, Karl DM (1996). Nitrite distributions and dynamics at Station ALOHA. *Deep*
 31 *Sea Research Part II: Topical Studies in Oceanography* **43**: 385-402.
 32
 33 Eiler A, Hayakawa DH, Church MJ, Karl DM, Rappe MS (2009). Dynamics of the
 34 SAR11 bacterioplankton lineage in relation to environmental conditions in the
 35 oligotrophic North Pacific subtropical gyre. *Environ Microbiol* **11**: 2291-2300.
 36
 37 Eymann C, Homuth G, Scharf C, Hecker M (2002). Bacillus subtilis functional
 38 genomics: Global characterization of the stringent response by proteome and
 39 transcriptome analysis. *J Bacteriol* **184**: 2500-2520.
 40
 41 Francis CA, Co EM, Tebo BM (2001). Enzymatic manganese(II) oxidation by a marine
 42 alpha-proteobacterium. *Appl Environ Microbiol* **67**: 4024-4029.
 43
 44 Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al* (2008).
 45 Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA*
 46 **105**: 3805-3810.

- 1
2 Frigaard NU, Martinez A, Mincer TJ, DeLong EF (2006). Proteorhodopsin lateral gene
3 transfer between marine planktonic Bacteria and Archaea. *Nature* **439**: 847-850.
4
5 Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL *et al* (2005a).
6 Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**: 82-85.
7
8 Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al* (2005b).
9 Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242-1245.
10
11 Gómez-Consarnau L, Gonzalez JM, Coll-Llado M, Gourdon P, Pascher T, Neutze R *et al*
12 (2007). Light stimulates growth of proteorhodopsin-containing marine Flavobacteria.
13 *Nature* **445**: 210-213.
14
15 Gómez-Consarnau L, Akram N, Lindell K, Pedersen A, Neutze R, Milton DL *et al*
16 (2010). Proteorhodopsin Phototrophy Promotes Survival of Marine Bacteria during
17 Starvation. *PLoS Biol* **8**: e1000358.
18
19 González JM, Fernandez-Gomez B, Fernandez-Guerra A, Gomez-Consarnau L, Sanchez
20 O, Coll-Llado M *et al* (2008). Genome analysis of the proteorhodopsin-containing marine
21 bacterium Polaribacter sp MED152 (Flavobacteria). *Proc Natl Acad Sci USA* **105**: 8724-
22 8729.
23
24 Guell M, van Noort V, Yus E, Chen W-H, Leigh-Bell J, Michalodimitrakis K *et al*
25 (2009). Transcriptome Complexity in a Genome-Reduced Bacterium. *Science* **326**: 1268-
26 1271.
27
28 Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM *et al* (2006).
29 Pathways of carbon assimilation and ammonia oxidation suggested by environmental
30 genomic analyses of marine Crenarchaeota. *PLoS Biol* **4**: 520-536.
31
32 Hewson I, Poretsky RS, Dyhrman ST, Zielinski B, White AE, Tripp HJ *et al* (2009).
33 Microbial community gene expression within colonies of the diazotroph, Trichodesmium,
34 from the Southwest Pacific Ocean. *Isme Journal* **3**: 1286-1300.
35
36 Hewson I, Rachel SP, Tripp HJ, Joseph PM, Jonathan PZ (2010). Spatial patterns and
37 light-driven variation of microbial population gene expression in surface waters of the
38 oligotrophic open ocean. *Environ Microbiol* **12**: 1940-1956.
39
40 Huson DH, Auch AF, Qi J, Schuster SC (2007). MEGAN analysis of metagenomic data.
41 *Genome Res* **17**: 377-386.
42
43 Jaroszewski L, Li ZW, Krishna SS, Bakolitsa C, Wooley J, Deacon AM *et al* (2009).
44 Exploration of Uncharted Regions of the Protein Universe. *PLoS Biol* **7**: e1000205.
45

- Jing H, Takagi J, Liu JH, Lindgren S, Zhang RG, Joachimiak A *et al* (2002). Archaeal surface layer proteins contain beta propeller, PKD, and beta helix domains and are related to metazoan cell surface proteins. *Structure* **10**: 1453-1464.
- Karl DM, Lukas R (1996). The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep Sea Research Part II: Topical Studies in Oceanography* **43**: 129-156.
- Kolber ZS, Plumley FG, Lang AS, Beatty JT, Blankenship RE, VanDover CL *et al* (2001). Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean. *Science* **292**: 2492-2495.
- Konneke 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.
- Lami R, Cottrell M, T., Campbell B, J., Kirchman D, L. (2009). Light-dependent growth and proteorhodopsin expression by *Flavobacteria* and SAR11 in experiments with Delaware coastal waters. *Environ Microbiol* **11**: 3201–3209.
- Liu ZZ, DeSantis TZ, Andersen GL, Knight R (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* **36**: e120.
- Malmstrom RR, Coe A, Kettler GC, Martiny AC, Frias-Lopez J, Zinser ER *et al* (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J*: doi: 10.1038/ismej.2010.1060.
- Martens-Habbena W, Berube PM, Urakawa H, de la Torre JR, Stahl DA (2009). Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* **461**: 976-979.
- Martinez A, Tyson GW, DeLong EF (2010). Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ Microbiol* **12**: 222-238.
- Martiny AC, Huang Y, Li WZ (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* **11**: 1340-1347.
- McCarren J, DeLong EF (2007). Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. *Environ Microbiol* **9**: 846-858.
- Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al* (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: doi:10.1186/1471-2105-1189-1386.

- 1
- 2 Moore LR, Chisholm SW (1999). Photophysiology of the marine cyanobacterium
- 3 Prochlorococcus: Ecotypic differences among cultured isolates. *Limnol Oceanogr* **44**:
- 4 628-638.
- 5
- 6 Moore LR, Post AF, Rocap G, Chisholm SW (2002). Utilization of different nitrogen
- 7 sources by the marine cyanobacteria Prochlorococcus and Synechococcus. *Limnol*
- 8 *Oceanogr* **47**: 989-996.
- 9
- 10 Oz A, Sabehi G, Koblizek M, Massana R, Beja O (2005). Roseobacter-like bacteria in
- 11 Red and Mediterranean Sea aerobic anoxygenic photosynthetic populations. *Appl*
- 12 *Environ Microbiol* **71**: 344-353.
- 13
- 14 Pham VD, Konstantinidis KT, Palden T, DeLong EF (2008). Phylogenetic analyses of
- 15 ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical
- 16 profile in the North Pacific Subtropical Gyre. *Environ Microbiol* **10**: 2313-2330.
- 17
- 18 Poretsky RS, Hewson I, Sun SL, Allen AE, Zehr JP, Moran MA (2009). Comparative
- 19 day/night metatranscriptomic analysis of microbial communities in the North Pacific
- 20 subtropical gyre. *Environ Microbiol* **11**: 1358-1375.
- 21
- 22 Riedel T, Tomasch J, Buchholz I, Jacobs J, Kollenberg M, Gerdt G *et al* (2010).
- 23 Constitutive Expression of the Proteorhodopsin Gene by a Flavobacterium Strain
- 24 Representative of the Proteorhodopsin-Producing Microbial Community in the North
- 25 Sea. *Appl Environ Microbiol* **76**: 3187-3197.
- 26
- 27 Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al* (2003).
- 28 Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche
- 29 differentiation. *Nature* **424**: 1042-1047.
- 30
- 31 Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al*
- 32 (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through
- 33 Eastern Tropical Pacific. *PLoS Biol* **5**: 398-431.
- 34
- 35 Scherl A, Fran ois P, Bento M, Deshusses JM, Charbonnier Y, Converset V *et al* (2005).
- 36 Correlation of proteomic and transcriptomic profiles of Staphylococcus aureus during the
- 37 post-exponential phase of growth. *J Microbiol Methods* **60**: 247-257.
- 38
- 39 Shi Y, Tyson GW, DeLong EF (2009). Metatranscriptomics reveals unique microbial
- 40 small RNAs in the ocean's water column. *Nature* **459**: 266-269.
- 41
- 42 Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF *et al* (2008).
- 43 Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the
- 44 Sargasso Sea. *ISME J* **3**: 93-105.
- 45

- 1 Steunou AS, Jensen SI, Brecht E, Becraft ED, Bateson MM, Kilian O *et al* (2008).
2 Regulation of nif gene expression and the energetics of N₂ fixation over the diel cycle in
3 a hot spring microbial mat. *ISME J* **2**: 364-378.
- 4
5 Stewart FJ, Ottesen EA, DeLong EF (2010). Development and quantitative analyses of a
6 universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J* **4**: 896-
7 907.
- 8
9 Tam R, Saier MH (1993). Structural, Functional, and Evolutionary Relationships among
10 Extracellular Solute-Binding Receptors of Bacteria. *Microbiological Reviews* **57**: 320-
11 346.
- 12
13 Temperton B, Field D, Oliver A, Tiwari B, Muhling M, Joint I *et al* (2009). Bias in
14 assessments of marine microbial biodiversity in fosmid libraries as evaluated by
15 pyrosequencing. *Isme Journal* **3**: 792-796.
- 16
17 Van Mooy BAS, Devol AH (2008). Assessing nutrient limitation of Prochlorococcus in
18 the North Pacific subtropical gyre by using an RNA capture method. *Limnol Oceanogr*
19 **53**: 78-88.
- 20
21 Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ *et al* (2010).
22 Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and
23 autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci USA* **107**:
24 8818-8823.
- 25
26 Wu JF, Sunda W, Boyle EA, Karl DM (2000). Phosphate depletion in the western North
27 Atlantic Ocean. *Science* **289**: 759-762.
- 28
29 Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K *et al* (2007).
30 The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein
31 families. *PLoS Biol* **5**: 432-466.
- 32
33 Zhang Y, Gladyshev VN (2008). Trends in Selenium Utilization in Marine Microbial
34 World Revealed through the Analysis of the Global Ocean Sampling (GOS) Project.
35 *PLoS Genet* **4**: e1000095.
- 36
37 Zhao G-Y, Chen X-L, Zhao H-L, Xie B-B, Zhou B-C, Zhang Y-Z (2008). Hydrolysis of
38 Insoluble Collagen by Deseasin MCP-01 from Deep-sea Pseudoalteromonas sp. SM9913.
39 *J Biol Chem* **283**: 36100-36107.
- 40
41 Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, Scanlan DJ *et al* (2006).
42 Prochlorococcus ecotype abundances in the North Atlantic Ocean as revealed by an
43 improved quantitative PCR method. *Appl Environ Microbiol* **72**: 723-732.
- 44

- 1 Zinser ER, Lindell D, Johnson ZI, Futschik ME, Steglich C, Coleman ML *et al* (2009).
- 2 Choreography of the Transcriptome, Photophysiology, and Cell Cycle of a Minimal
- 3 Photoautotroph, *Prochlorococcus*. *PLoS ONE* **4**: e5135.

Figure 1

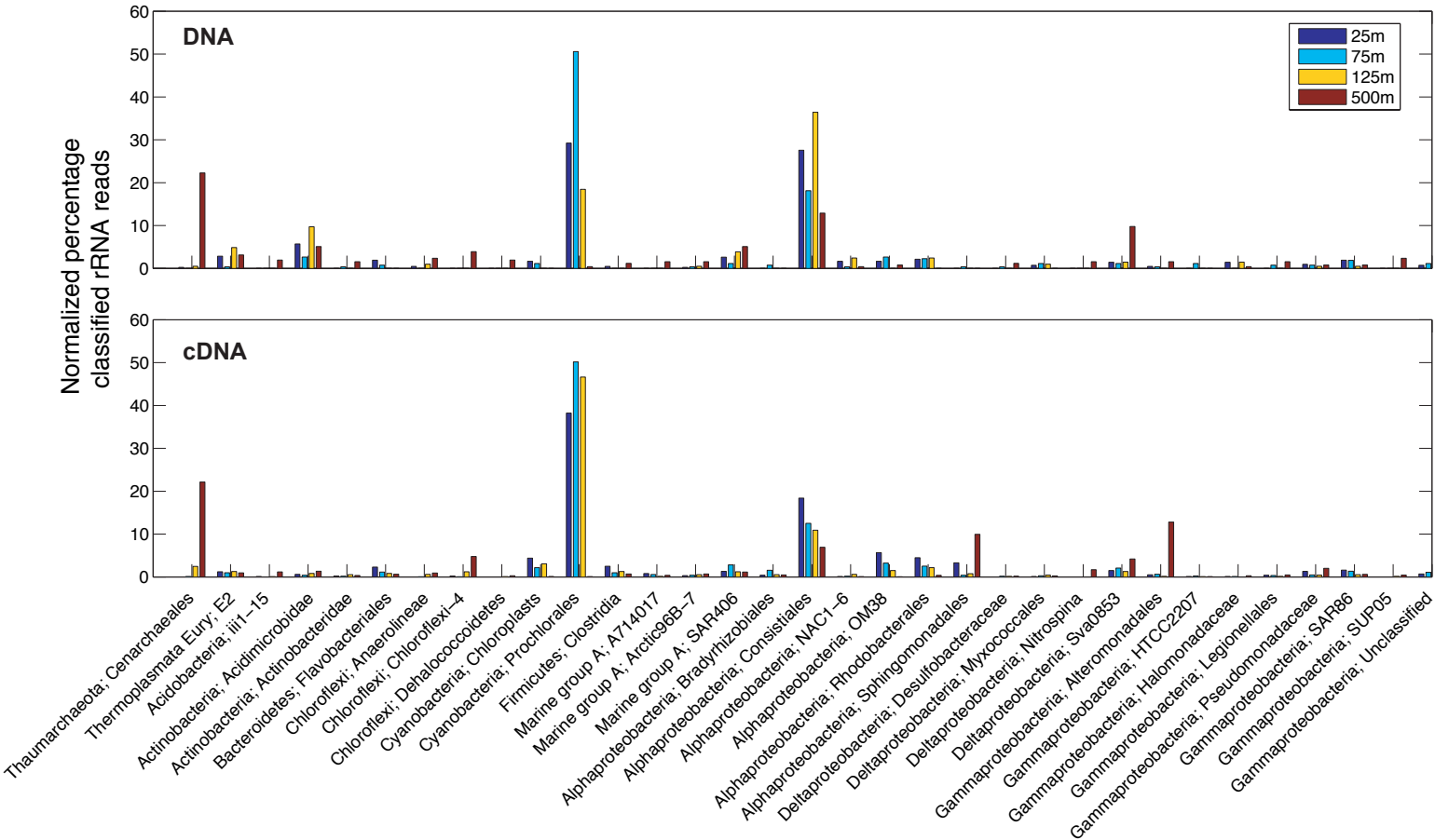


Figure 2

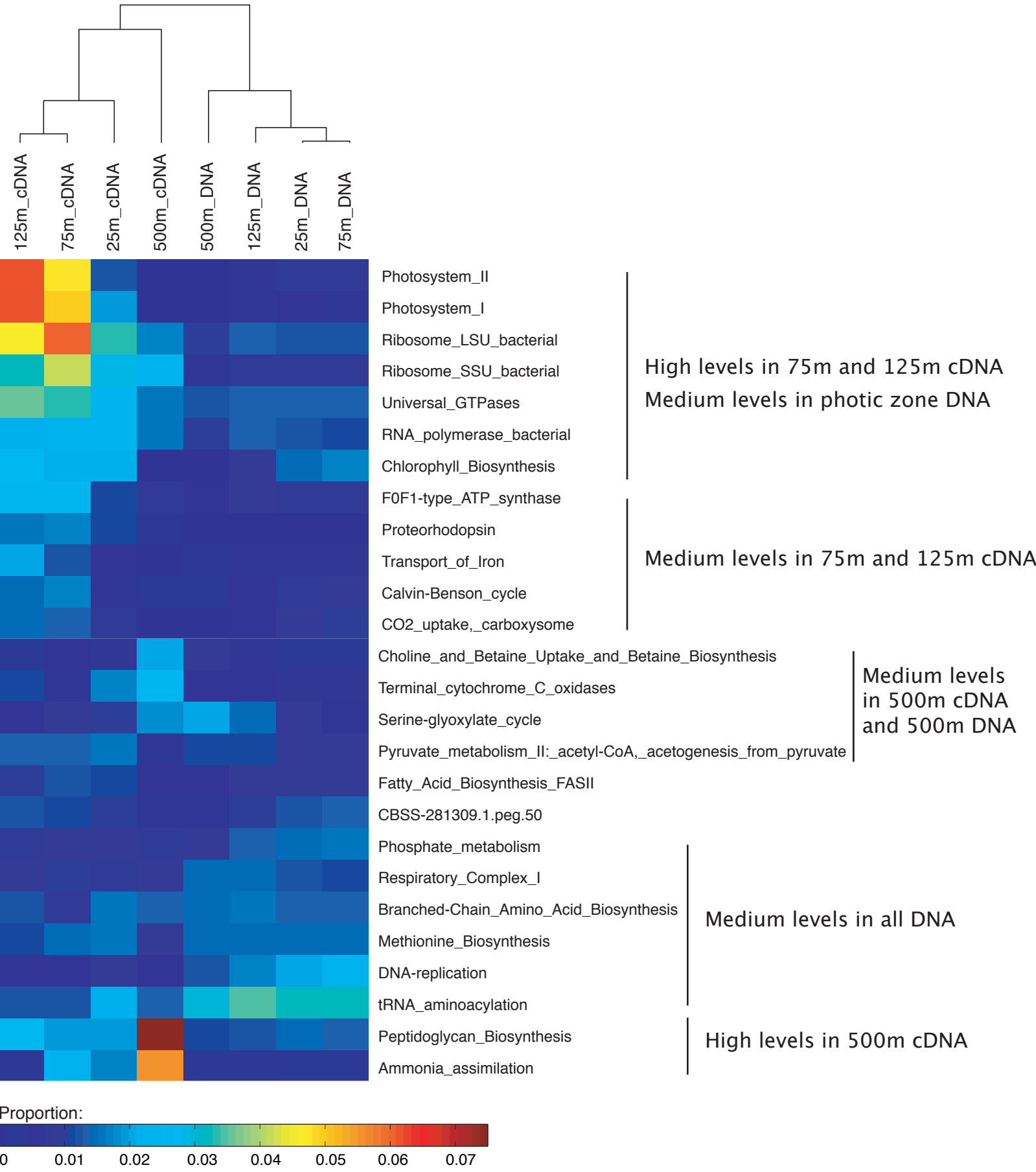


Figure 3

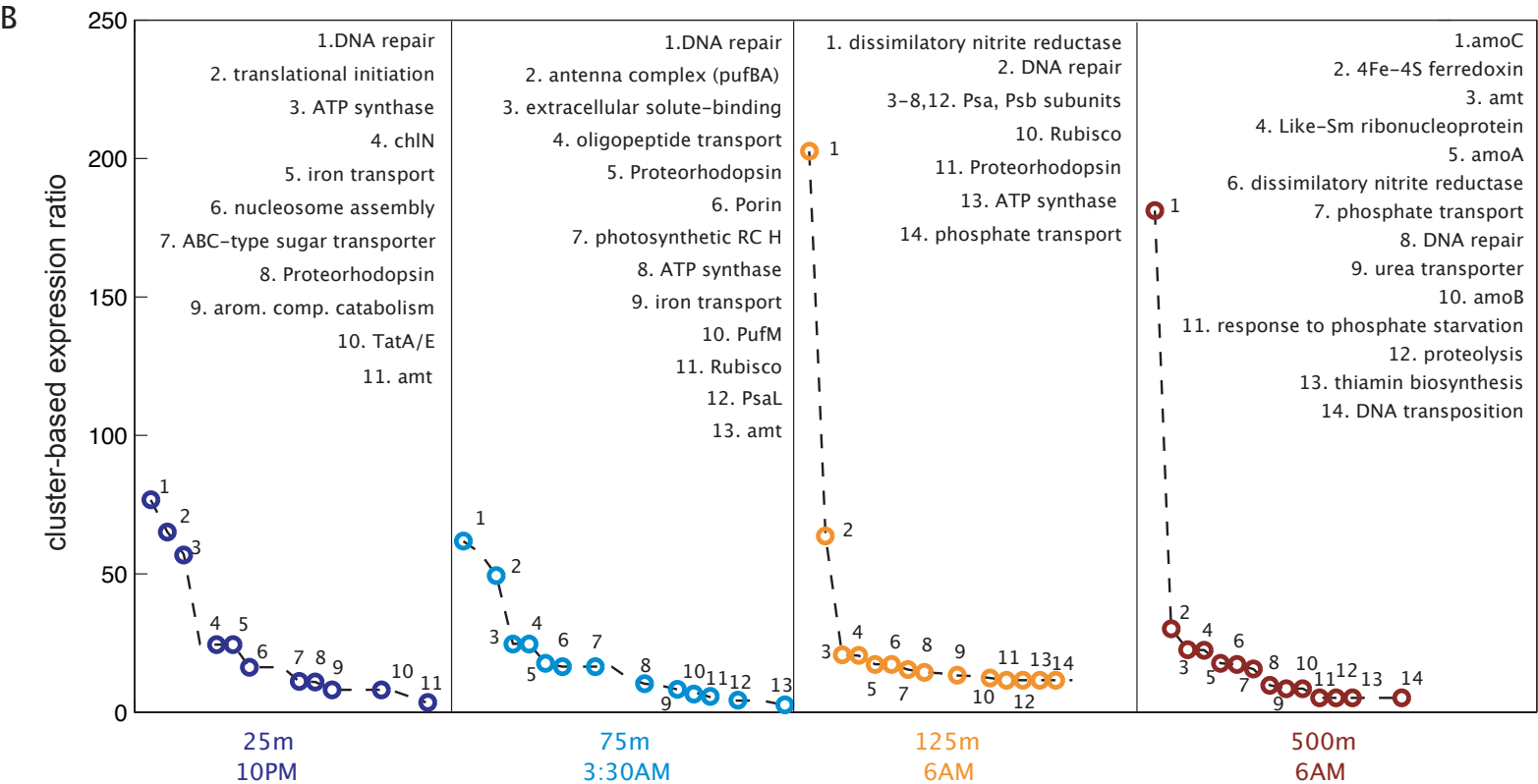
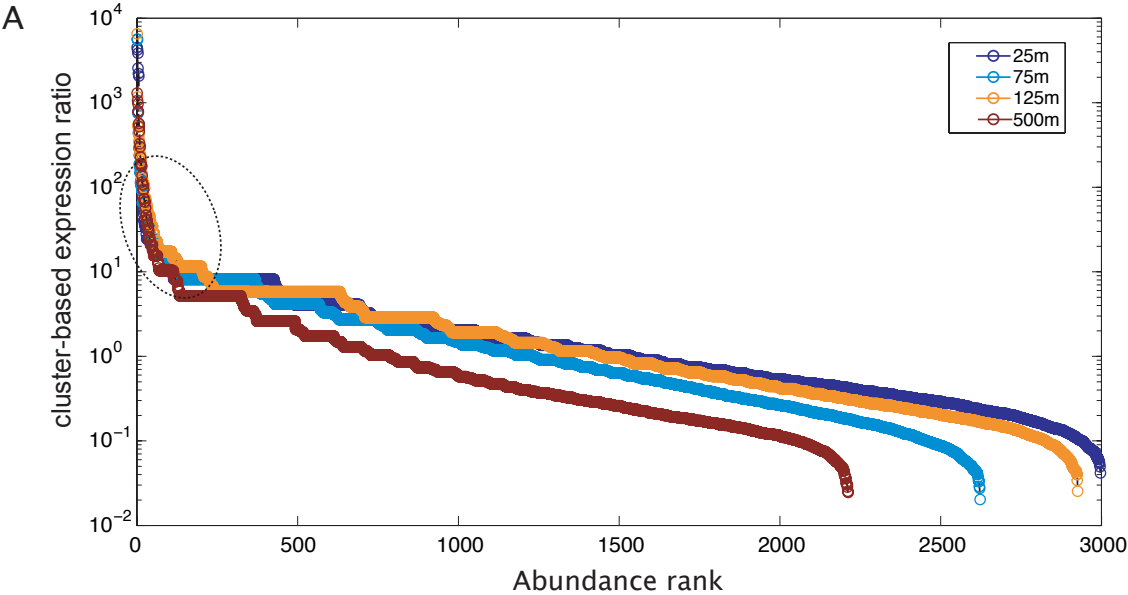


Figure 4

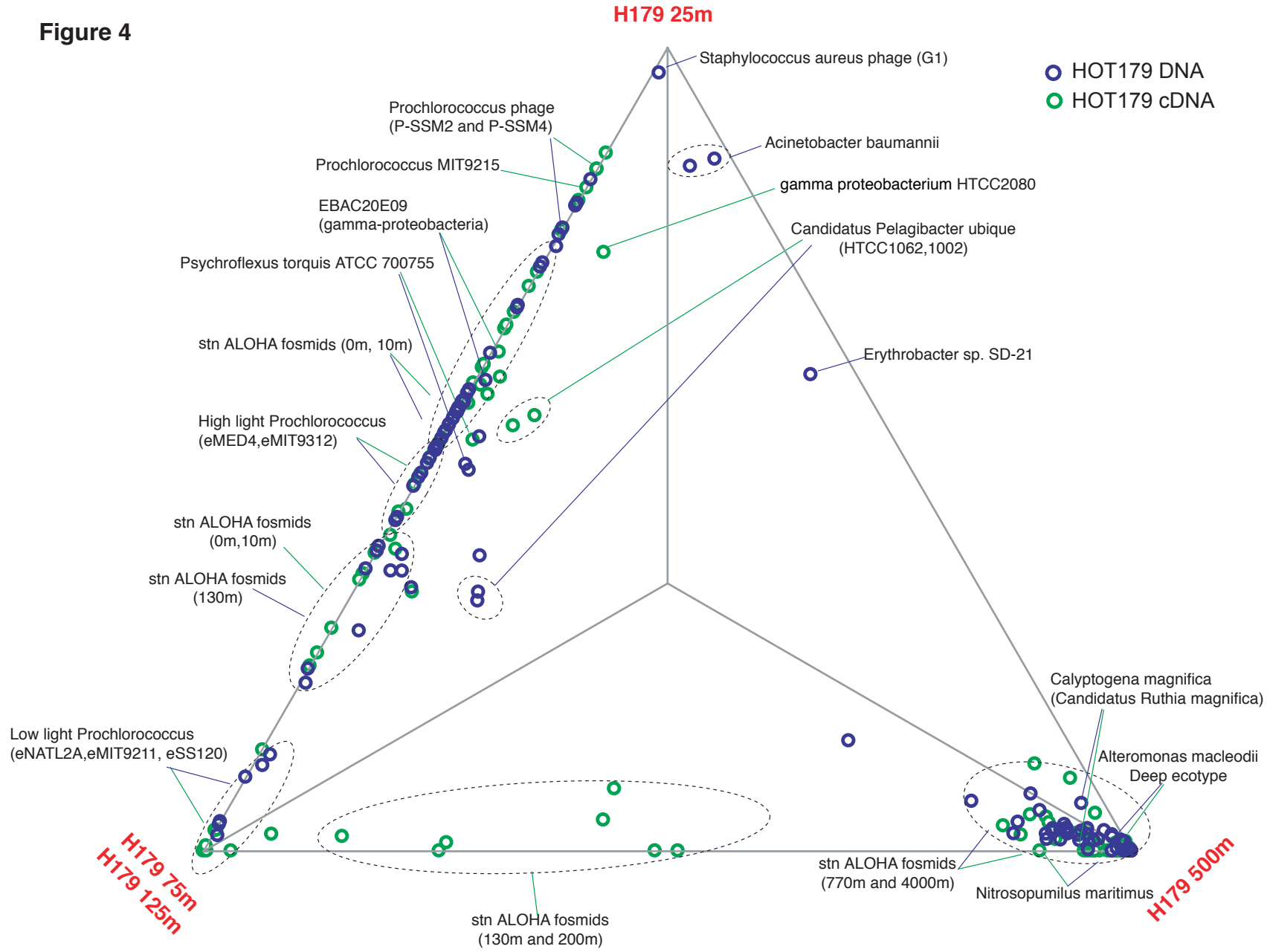


Figure 5

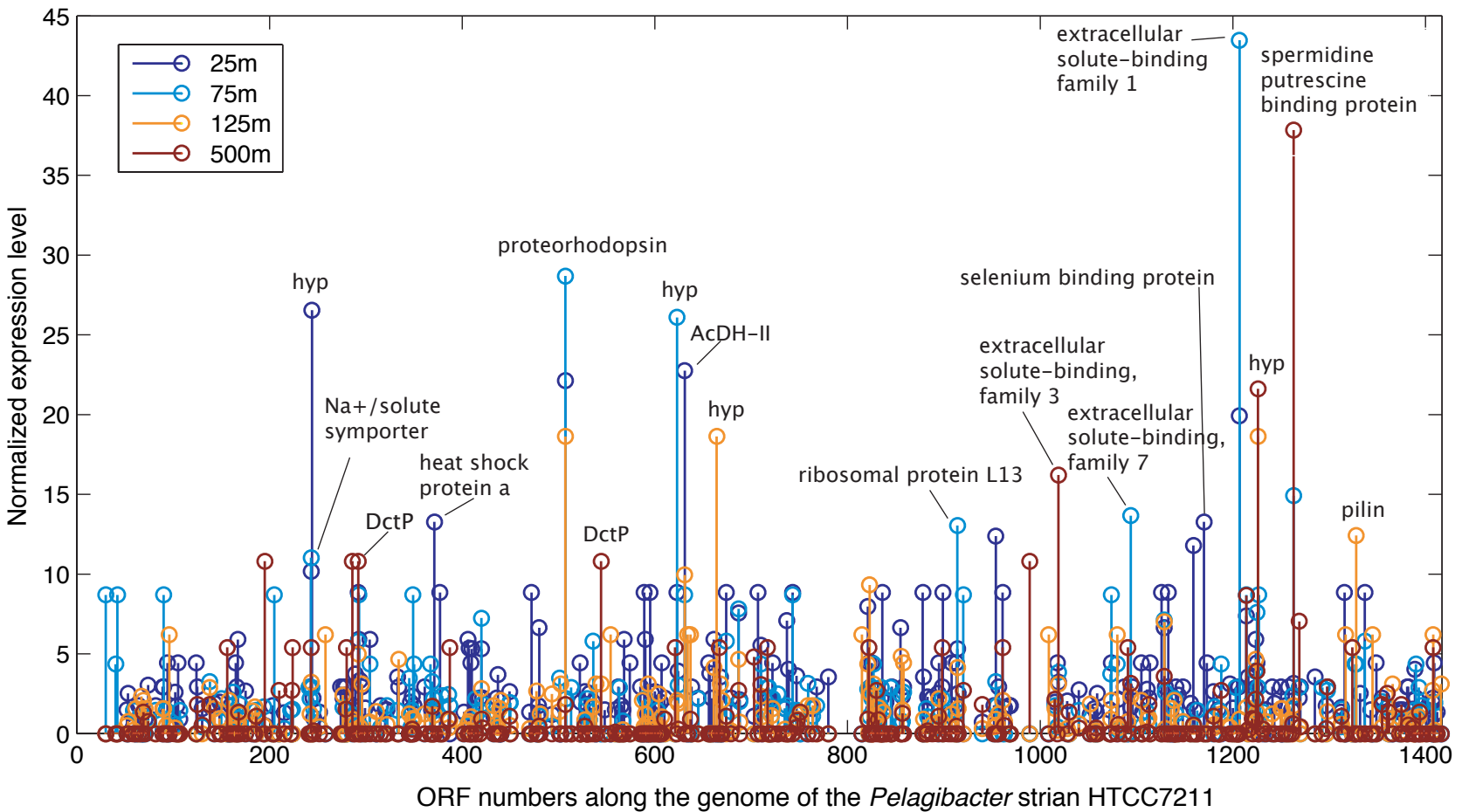


Table 1. Summary of 4 metagenomic data sets and 4 metatranscriptomic data sets.

HOT 179	Depth	# of total reads	Ave. read length (bp)	# of rRNA reads	% of rRNA in total reads	# of non rRNA reads	hits to protein db (% of non rRNA)			
							COG	SEED	NCBI-nr	GOS protein family
cDNA	25 m	74638	99	33878	45.4	40760	7.5	11.2	17.1	45.3
	75 m	106936	99	62096	58.1	44840	6.0	9.9	15.3	49.4
	125 m	97915	97	45809	46.8	52106	6.2	10.4	16.1	46.2
	500 m	109249	97	40537	37.1	68712	3.8	4.4	10.1	26.3
DNA	25 m	359665	109	1188	0.3	358477	19.1	26.7	42.0	63.5
	75 m	388652	110	1117	0.3	387535	22.4	33.2	51.3	71.9
	125 m	322751	109	954	0.3	321797	18.1	23.4	36.3	60.9
	500 m	371071	107	1029	0.3	370042	17.3	18.3	30.5	49.0

Table 2. SEED subsystems that are significantly enriched in cDNA data sets relative to DNA data sets (0.98 confidence level, based on the method described in Rodriguez-Brito et al, 2006).

Depth	Subsystem*	Representation in cDNA	Representation in DNA
25m	Ammonia_assimilation	1.52%	0.25%
	Photosystem_I	1.72%	0.58%
	Proteorhodopsin	1.00%	0.03%
	Ribosome_LSU_bacterial	3.04%	1.23%
	Ribosome_SSU_bacterial	2.58%	0.79%
	Universal_GTPases (mostly elongation factors)	2.36%	1.31%
	RNA_polymerase_bacterial	2.46%	1.25%
	Transcription_initiation,_bacterial_sigma_factors	0.80%	0.21%
	Terminal_cytochrome_C_oxidases	1.60%	0.38%
	Ubiquinone_Menaquinone-cytochrome_c_reductase_complexes	0.58%	0.11%
	Oxidative_stress	0.90%	0.28%
75m	Ammonia_assimilation	1.09%	0.26%
	Photosystem_I	2.38%	0.66%
	Photosystem_II	2.31%	0.81%
	Proteorhodopsin	0.80%	0.03%
	Ribosome_LSU_bacterial	2.90%	1.20%
	Ribosome_SSU_bacterial	1.97%	0.79%
125m	CO2_uptake,_carboxysome	1.20%	0.49%
	Peptidoglycan_Biosynthesis	2.28%	1.24%
	Chlorophyll_Biosynthesis	2.34%	0.87%
	Photosystem_I	5.24%	0.37%
	Photosystem_II	5.21%	0.46%
	Proteorhodopsin	1.34%	0.04%
	Ribosome_LSU_bacterial	3.92%	1.32%
	Ribosome_SSU_bacterial	2.69%	0.77%
	Universal_GTPases (mostly elongation factors)	3.05%	1.37%
	F0F1-type_ATP_synthase	2.14%	0.92%
	Cytochrome_B6-F_complex	0.86%	0.16%
	Transport_of_Iron	1.78%	0.40%
500m	Peptidoglycan_Biosynthesis	4.63%	1.12%
	Ammonia_assimilation	3.43%	0.12%
	Ribosome_SSU_bacterial	1.41%	0.67%
	Terminal_cytochrome_C_oxidases	1.55%	0.51%

* Subsystems listed are significantly enriched in cDNA samples at the 0.98 confidence level