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.

<table>
<thead>
<tr>
<th>Citation</th>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1038/ismej.2010.189">http://dx.doi.org/10.1038/ismej.2010.189</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Nature Publishing Group</td>
</tr>
<tr>
<td>Version</td>
<td>Author’s final manuscript</td>
</tr>
<tr>
<td>Accessed</td>
<td>Tue Jan 29 14:09:16 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/69640">http://hdl.handle.net/1721.1/69640</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-Noncommercial-Share Alike 3.0</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/3.0/">http://creativecommons.org/licenses/by-nc-sa/3.0/</a></td>
</tr>
</tbody>
</table>
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. Tyson¹,³, John M. Eppley¹, Edward F. DeLong¹,²,*

¹Departments of Civil and Environmental Engineering 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
Abstract

As part of an ongoing survey of microbial community gene expression in the ocean, we sequenced and compared a total of ~38 Mbp of community transcriptomes and ~157 Mbp of community genomes from four bacterioplankton samples, along a defined depth profile at Station ALOHA in North Pacific subtropical gyre (NPSG). Taxonomic analysis based on rRNA (as well as protein-coding sequences) suggested that the samples were dominated by three taxa: Prochlorales, Consistiales, and Cenarchaeales, that comprised 36-69% and 29-63% of the annotated sequences in the four DNA and four cDNA libraries, respectively. The remaining of the sequences represented a broad diversity of low abundance taxa (33 taxonomic groups detected with relative abundance of ≥ 1% in any of the eight libraries). The relative abundance of major taxonomic groups was sometimes inconsistently represented in the DNA and cDNA libraries. For example, the 125m sample genomic library was dominated by *Pelagibacter* (~36% of sequence reads), which contributed far fewer sequences to the community transcript pool (~11%). Those data suggest the utility of metatranscriptomics for assessing the relative transcriptional activities per cell for different taxonomic groups. Functional characterization in combination with taxonomic classification for highly expressed genes revealed taxon-specific contributions to active biogeochemical processes such as phototrophy and nitrogen metabolism. Examples included *Roseobacter*-relatives involved in aerobic anoxygenic phototrophy at 75m, and the unexpected contribution of ammonia oxidation by low abundance crenarchaeal population at 125m. Recruitment of DNA and cDNA reads to reference microbial genomes indicated depth-specific partition of coexisting microbial populations, as highlighted by the transcriptionally active HL-like
Prochlorococcus population in the bottom of the photic zone. Transcripts that mapped to Pelagibacter genomes suggested that nutrient uptake genes dominated Pelagibacter transcriptomes, with apparent enrichment for certain transporter types (e.g., the C4-dicarboxylate transport system) than others (e.g., phosphorus transporters). Collectively, the data support the utility of coupled DNA and cDNA analyses for describing the taxonomic and functional attributes of microbial communities in their natural habitats.

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

**Introduction**

Marine microbial communities, centrally involved in the fluxes of matter and energy in the global oceans, are major drivers of global biogeochemical cycling (Arrigo 2005, Karl and Lukas 1996). Our knowledge of abundance, diversity and gene content of planktonic microbes has been fundamentally advanced over the past three decades, by both model organism-based studies (Coleman and Chisholm 2007, Giovannoni et al 2005b), as well as metagenomic surveys of natural microbial communities (DeLong et al 2006, Dinsdale et al 2008, Rusch et al 2007). In particular, metagenomic comparisons of distinct microbiomes (DeLong et al 2006, Dinsdale et al 2008) have revealed habitat-dependent distribution of taxons and gene families, likely shaped by the biogeochemical conditions of each environment. Clearly, determining if and how such genomic variations are manifested at the level of gene expression and regulation represents another critical step towards understanding the interplay between microbes and their natural 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)_{16}VN 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 BmpI to remove poly(A) tails. Purified cDNA was then
directly sequenced by pyrosequencing (GS20). See Supplementary Methods for further
details.

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

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

Results and Discussion

Bacterioplankton samples and pyrosequencing data sets

The four sampling depths represent discrete zones in the water column at Station
ALOHA (22°45' N, 158°W), which includes the middle of the mixed layer (25 m), the
base of the mixed layer (75m), the deep chlorophyll maximum (DCM, 125m) at the top of the nutricline, and the upper mesopelagic zone (500m). On cruise HOT179, bacterioplankton samples were collected from each depth for RNA and DNA extraction and sequencing. Since the sampling times for these four sets of seawater samples were different (25m at 22:00 local time, 75m at 03:00, 125m at 06:00, and 500m at 06:00), we expected that the observed gene expression patterns would reflect spatial geochemical gradients (Supplementary Figure S1), as well as temporal differences (discussed below).

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

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

Roughly 0.3% of total DNA reads were designated as rRNA operon sequences (1188, 1117, 954, and 1029 reads for the 25m, 75m, 125m, and 500m samples, respectively), including bacterial, archaeal, and eukaryotic small and large subunit rRNAs, and intergenic spacer sequences. This sampling frequency was within the
expected range based on the rRNA operon size (~5,000 bp), assuming average genome size of ~2 Mbp for marine bacteria and archaea. To assess the taxonomic diversity within the four microbial communities, we classified these 16S rRNA gene sequences (Figure 1, upper panel), using the online Greengenes alignment and classification tools (http://greengenes.lbl.gov/cgi-bin/nph-classify.cgi) (DeSantis et al 2006), which was reported to yield the highest accuracy for assigning taxonomy to short pyrosequencing reads compared to other methods such as RDP classifier or BLAST (Liu et al 2008). These taxonomic assignments were further corroborated (Supplementary Figure S2; Pearson’s correlation > 0.95 for all four depths) using a full set of “shotgun” DNA library sequences (average read length 565 bp) from the same source DNA samples (Martinez et al 2010).

Each of the four microbial communities was dominated by two or three major groups (Figure 1, upper panel). Consistiales (predominantly Pelagibacter) recruited ~13-35% of the total classified 16S rRNA gene reads from all depths, supporting the high abundance of Pelagibacter populations throughout the water column (Eiler et al 2009) and their under-representation in large-insert metagenomic libraries, at least for the populations residing shallower depths (Pham et al 2008, Temperton et al 2009). The other major groups included Prochlorales in the photic zone (~17-51%), Cenarchaeales (~22%) and the uncultured delta-proteobacterial group SVA0853 (~9%) at 500m, and Acidimicrobidae (~2-8%) at all depths. This depth distribution was generally consistent with previous cultivation-independent surveys at this site, but variability (likely both biological and methodological) was apparent. For instance, a fosmid library-based survey (DeLong et al 2006) reported a significant decrease in the relative abundance of
Prochlorococcus populations at 75m depth, potentially caused by cyanophage infection, as suggested by the large number of cyanophage sequences recovered in the same cellular size fraction. In contrast, in this survey large numbers of phage sequences were not detected, and Prochlorococcus relative abundance peaked at 75 m depth, regardless of DNA library type and sequencing method (pyrosequencing, Figure 1; fosmid clone library, Table S1).

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

Another common approach to assess taxonomic composition from metagenomic data sets is to infer taxonomic origins from open reading frame (ORF) sequences (Huson et al 2007). Here, we observed both consistencies as well as some discrepancies when comparing the community composition derived from rRNA gene sequences (discussed above) to those derived from ORF sequences using MEGAN (Huson et al 2007). As seen in Figure 1 and Supplementary Figure S3, Pelagibacter relative abundance decreased from ~13-35% estimated from the 16S rRNA gene sequences, to ~9-23% from the ORF sequences, and the uncultured delta-proteobacterium SVA0853 was completely missed in the latter. In contrast, Prochlorococcus-like sequences represented ~39-71% of all annotated ORF sequences, much higher than that estimated from 16S rRNA gene sequences (~17-51%). Higher representation of Prochlorococcus-like mRNA transcripts relative to their cell abundance was noted by Poretsky et al in metatranscriptomic data sets from day and night samples from the same site, and was attributed to higher transcriptional activities of Prochlorococcus cells relative to coexisting heterotrophic 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
lower transcriptional activity of *Pelagibacter* may be influenced by the time of DCM sample collection: ~6AM local time, when photosynthetic microorganisms such as *Prochlorococcus* may be relatively more active.

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

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

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

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

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

*Nitrogen metabolism protein families*

A suite of nitrogen metabolism genes (ammonium transporter, \textit{amt}; dissimilatory nitrite reductase, \textit{nirK}; urea transporter, \textit{urt}; ammonia monooxygenase subunits, \textit{amoABC}) was among the most highly expressed of GOS protein families detected (Figure 3B). An essential macronutrient, nitrogen availability and turnover limits
biological production in many open ocean regions, including NPSG (Van Mooy and Devol 2008). Ammonia/ammonium is a key reduced nitrogen compound that can either be incorporated into carbon skeleton via the glutamine synthetase (GS; glnA)/glutamate synthase (GOGAT; glsF) cycle, or can serve as energy source fueling autotrophic metabolism (Konneke et al 2005). Thus, the transport of ammonia/ammonium is vital to planktonic microbes living in the nutrient deplete surface waters and energy constrained deep waters in an open ocean setting. Urea is another potentially important nitrogen source in the ocean, and is utilized by marine cyanobacteria (Moore et al 2002). The more oxidized forms of nitrogen, nitrite and nitrate require more metabolic energy to utilize but can serve as alternative nitrogen sources because of their much higher concentrations in deep euphotic zone and mesoplegic zone below the nitracline.

To assess the prevalent nitrogen utilizing pathways in the genomes of the most abundant planktonic microbial populations, we compared the observed frequency (normalized to gene length and data set size) of several essential nitrogen metabolism genes with that of the 16S rRNA gene of *Prochlorococcus* and marine group I crenarchaeota. The observed frequency of *Prochlorococcus*-related *amt*, *glnA*, *urt*, urease genes is equivalent to that of *Prochlorococcus* 16S rRNA gene (Supplementary Figure S4A, left panel), suggesting that ammonium and urea assimilation is preserved in naturally occurring *Prochlorococcus* populations. In contrast, the assimilatory nitrite reductase gene (*nirA*) was present in only a small fraction of *Prochlorococcus* cells (c.a., 7%, 8% and 15% at 25m, 75m, and 125m, respectively), consistent with expectation based on genomic and physiological studies of *Prochlorococcus* isolates (Moore et al 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
expression levels of crenarchaea-related *amt* and *amoABC* genes (especially *amoC* gene) was among the highest in our data sets (orders of magnitude higher than most other protein-coding genes) (Figure 3B). Interestingly, the anomalously high *amoC* gene expression appeared to be universal, as also observed in bacterial nitrifiers (Berube et al 2007), for as-yet unknown reasons. Consistent with a quantitative PCR-based study (Church et al 2010), the *amoABC* transcripts were detected in high abundance at 125m depth despite the small planktonic crenarchaeal population size (Supplementary Figure S4B, right panel). Together with previous report of remarkably high substrate affinity and kinetics of crenarcheal *amo* genes (Martens-Habbena et al 2009), these data further support a role for marine crenarchae in nitrification in the ocean via active ammonia oxidation.

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

**Photoheterotrophy**

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

Our depth profile data allow us to directly assess the in situ abundance and
taxonomic origins of PR gene and transcripts. Abundance of PR transcripts decreased
dramatically from euphotic zone to 500m (in which only 4 cDNA reads shared homology
to known PR genes) (Supplementary Figure S5A). While PR DNA and cDNA reads
appeared to be originated from a diverse range of taxa, the majority shared homology to
known PR genes from SAR11-like organisms (Supplementary Figure S5B). Notably, PR
genes were found most highly expressed in the 75m sample (collected at 22:00), followed
by the 25m and 125m samples (collected at 3:00 and 6:00, respectively) (Supplementary
Figure S5A; also see the Pelagibacter genome-wide gene expression analysis below),
suggesting PR genes may be constitutively expressed in the photic zone independent of
light conditions. Laboratory studies of PR-containing isolates as well as a recently
reported microcosm experiment have reported inconsistent observations, some suggesting
constitutive PR expression (Giovannoni et al 2005a, Riedel et al 2010), while others
suggesting light-regulation of PR expression (Gómez-Consarnau et al 2007, Lami et al
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 (Béjá 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
comparison (see Supplementary Methods). About 29%, 40%, 15% and 7% of total DNA reads, and 30%, 24%, 26%, and 18% of total cDNA reads were recruited to the reference genomic data for 25m, 75m, 125m, and 500m sample, respectively. Notably, the percentage of recruited cDNA reads for each sample was significantly higher than that of cDNA reads that could be assigned to NCBI-nr protein database (Table 1), a result of cDNA recruitment to expressed noncoding regions on the genomes. For instance, about 1539 reads in the 25m sample were recruited to an intergenic region of *Prochlorococcus* strain MIT 9215 genome, corresponding to the Group_2 small RNA previously reported by Shi *et al* (Shi *et al* 2009).

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

Such genome recruitment analysis provides direct measurement of vertical distribution of ecologically coherent populations (represented by reference genomes) in nature, such as high-light (HL) and low-light (LL) adapted *Prochlorococcus* “ecotypes” (Moore and Chisholm 1999). Notably, despite an expected significant increase of low-
light (LL) adapted *Prochlorococcus* populations (mostly eNATL2A) at 125m, where light intensity dramatically decreased compared to shallower depths, > 80% of the *Prochlorococcus*-like reads at 125m were most similar to sequences of high-light (HL) adapted isolates (mostly eMIT9312) (Supplementary Table S2). While possibly a result of physical homogenization of the water column due to deep mixing in the winter (Malmstrom et al 2010), these HL-like *Prochlorococcus* cells displayed elevated transcriptional activity at 125m (Supplementary Table S2), suggesting they were unlikely sinking dead cells. Zinser and colleagues (Zinser et al 2006) showed that in deeper waters (below 75 m) at the western North Atlantic site, a significant fraction of *Prochlorococcus* population cannot be detected by qPCR probes designed to capture currently known ecotypes, suggesting significant deep populations of *Prochlorococcus* yet to be identified and characterized. Results here suggest the presence of a HL-like *Prochlorococcus* population that may be well adapted to the lower euphotic zone, under low light conditions.

**Population transcriptomic analysis of Pelagibacter.** As the most abundant heterotrophic bacterial group throughout the ocean water column, *Pelagibacter* (member of the alphaproteobacteria SAR11 clade) provides a useful model example for how culture-based and metagenomic/metatranscriptomic data can be integrated to study the ecophysiology of wild populations. Subsets of DNA and cDNA reads from all 4 depths were mapped onto the reference genome of the open ocean *Pelagibacter* isolate HTCC7211 (see Supplementary Methods). The expression level of annotated protein coding genes provided clues on the prevailing metabolic activities of *Pelagibacter* populations at each depth (Figure 5; Supplementary Table S3). Overall, the expression
profile of protein coding genes confirmed the observation based on the rRNA profile (Figure 1), that Pelagibacter cells at 125m were less transcriptionally active at the time of sampling, compared to their counterparts at 25m and 75m. Indeed, ribosomal proteins were among the most highly expressed genes in 25m and 75m samples, and most ORFs showed lower expression levels in the 125m sample.

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

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

**HTCC7211-specific genes.** It has been well established that genomic plasticity of microbes, reflected by variations in gene content of closely related strains, may facilitate microbial adaptation to their natural habitats (Coleman *et al* 2006, Cuadros-Orellana *et al* 2007). We compared the genome sequences of two *Pelagibacter* coastal isolates (strains HTCC1062 and HTCC1002) and the open ocean isolate (HTCC7211, used as reference genome in the genome-centric analysis above), and asked which HTCC7211-specific genes might be highly expressed and thus functionally important in the open ocean 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 $\text{H}^+$ 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
correlate directly with cognate protein levels, and the kinetics that relates expression to phenotype vary among different transcript classes (Steunou et al 2008). Nonetheless, reasonably good correlation between transcriptomes and proteomes, especially for transcripts and peptides in higher abundance, has been observed in several model organisms (Corbin et al 2003, Eymann et al 2002, Scherl et al 2005). In the Pelagibacter genome-centric analyses reported here (Figure 5), we observed considerable overlap between highly abundant transcripts and the most represented peptides previously reported in a SAR11-centric metaproteomic study (Sowell et al 2008). This general consistency between the population transcriptomes and proteomes of the most abundant and ubiquitous heterotrophic bacteria clade in the open ocean, supports the use of metatranscriptomics to assess inventories of functionally relevant gene families based on their expression levels.

The work reported here, along with previous studies (Hewson et al 2009, Poretsky et al 2009), also illustrates several challenges for future metatranscriptomic studies. First, due to great diversity found in most natural systems and predominant transcriptional signal of the genes involved in central metabolism and protein synthesis machinery, sequencing depth needs to be greatly expanded (Stewart et al 2010). Our data showed that for one pyrosequencing run on an open ocean bacterioplankton sample, about 66-74% of sequences with putative taxonomic assignment belonged to the top two most abundant taxonomic groups (Supplementary Figure S3, lower panel). Thus, the majority of the diversity of the transcript pool was represented by low abundance reads with little statistical confidence, albeit these may well contain important information. As
a good demonstration, two technical replicate metatranscriptomic data sets were found to only share 37% of the NCBI-nr reference entries, suggesting the rarefaction curve of functional diversity is far from leveling off (Stewart et al 2010).

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

5 Conclusions and future directions

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

Acknowledgements:

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

Figure Legends

Figure 1. Taxonomic classification based on 16S rRNA-bearing reads in DNA and cDNA data sets. Taxonomic assignments were binned at the Order level, using the Hugenholtz taxonomy of Greengenes (see Supplementary Methods). 16S rRNA sequences that could not be classified were excluded from the analysis. Y axis scale represents the percentage of the total classified 16S rRNA reads. Only taxa that represented $\geq$ 1% of all classified reads are displayed.
Figure 2. Clustering of all cDNA and DNA data sets based on relative abundance of SEED subsystems. Only the most abundant subsystems that together recruited 95% of all reads are displayed. Hierarchical clustering of 4 DNA and 4 cDNA samples were performed with euclidean distance and single linkage method using MATLAB. Color scale represents the proportion of reads assigned to SEED categories relative to the total library size in each sample. Blue to red color indicates low to high representation of SEED categories.

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

Figure 4. Three-way comparison of representation of genomes and genome fragments (fully sequenced fosmids) in DNA and cDNA data sets. The 75m and 125m data sets were combined since they were the most similar. Each dot represents a genome (fragment), and its proximity to a vertex reflects the enrichment of the corresponding 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


Figure 1

DNA

Normalized percentage of DNA reads

Cellulose

cDNA

Normalized percentage of cDNA reads

Classified rRNA reads
Figure 2

High levels in 75m and 125m cDNA
High levels in 500m cDNA and 500m DNA
High levels in 500m cDNA

Medium levels in photic zone DNA
Medium levels in 75m and 125m cDNA
Medium levels in 75m and 125m cDNA

Medium levels in all DNA

Proportion:

0.00  0.05  0.10  0.15  0.20  0.25  0.30  0.35  0.40  0.45  0.50  0.55  0.60  0.65  0.70  0.75  0.80  0.85  0.90  0.95  1.00
Figure 3

A

Abundance rank

Cluster-based expression ratio

25m 75m 125m 500m

1. DNA repair
2. Translational initiation
3. ATP synthase
4. chlN
5. Iron transport
6. Nucleosome assembly
7. ABC-type sugar transporter
8. Proteorhodopsin
9. Arom. comp. catabolism
10. TatA/E
11. Amt

B

Cluster-based expression ratio

25m 75m 125m 500m

1. DNA repair
2. Antenna complex (pufBA)
3. Extracellular solute-binding
4. Oligopeptide transport
5. Proteorhodopsin
6. Porin
7. Photosynthetic RC H
8. ATP synthase
9. Iron transport
10. PuFM
11. Rubisco
12. Psal
13. Amt

2. DNA repair
3-8,12. Psa, Psb subunits
10. Rubisco
11. Proteorhodopsin
13. ATP synthase
14. Phosphate transport

1. Dissimilatory nitrite reductase
2. DNA repair
3-8,12. Psa, Psb subunits
10. Rubisco
11. Proteorhodopsin
13. ATP synthase
14. Phosphate transport

1. AmoC
2. 4Fe-4S ferredoxin
3. Amt
4. Like-Sm ribonucleoprotein
5. AmoA
6. Dissimilatory nitrite reductase
7. Phosphate transport
8. DNA repair
9. Urea transporter
10. AmoB
11. Response to phosphate starvation
12. Proteolysis
13. Thiamin biosynthesis
14. DNA transposition
Prochlorococcus phage (P-SSM2 and P-SSM4)
Prochlorococcus MIT9215
EBAC20E09 (gamma-proteobacteria)
Psychroflexus torquis ATCC 700755
High light Prochlorococcus (eMED4, eMIT9312)
Low light Prochlorococcus (eNATL2A, eMIT9211, eSS120)
Candidatus Pelagibacter ubique (HTCC1062, 1002)
Psychroflexus torquis ATCC 700755
High light Prochlorococcus (eMED4, eMIT9312)
Low light Prochlorococcus (eNATL2A, eMIT9211, eSS120)
Candidatus Pelagibacter ubique (HTCC1062, 1002)

H179 25m
H179 75m
H179 125m
H179 500m
H179 125m
H179 25m

Figure 4
Figure 5

Normalized expression level

ORF numbers along the genome of the *Pelagibacter* strain HTCC7211
## Table 1. Summary of 4 metagenomic data sets and 4 metatranscriptomic data sets.

<table>
<thead>
<tr>
<th>HOT 179 Depth</th>
<th># of total reads</th>
<th>Ave. read length (bp)</th>
<th># of rRNA reads</th>
<th>% of rRNA in total reads</th>
<th># of non rRNA reads</th>
<th>hits to protein db (% of non rRNA)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>COG</td>
</tr>
<tr>
<td>25 m</td>
<td>74638</td>
<td>99</td>
<td>33878</td>
<td>45.4</td>
<td>40760</td>
<td>7.5</td>
</tr>
<tr>
<td>75 m</td>
<td>106936</td>
<td>99</td>
<td>62096</td>
<td>58.1</td>
<td>44840</td>
<td>6.0</td>
</tr>
<tr>
<td>125 m</td>
<td>97915</td>
<td>97</td>
<td>45809</td>
<td>46.8</td>
<td>52106</td>
<td>6.2</td>
</tr>
<tr>
<td>500 m</td>
<td>109249</td>
<td>97</td>
<td>40537</td>
<td>37.1</td>
<td>68712</td>
<td>3.8</td>
</tr>
</tbody>
</table>

| cDNA           |                  |                       |                 |                          |                     | COG | SEED | NCBI-nr | GOS protein family |
|----------------|------------------|-----------------------|-----------------|--------------------------|---------------------|-------------------------------------|
| 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               |

| DNA            |                  |                       |                 |                          |                     | COG | SEED | NCBI-nr | GOS protein family |
|----------------|------------------|-----------------------|-----------------|--------------------------|---------------------|-------------------------------------|
| 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).

<table>
<thead>
<tr>
<th>Depth</th>
<th>Subsystem*</th>
<th>Representation in cDNA</th>
<th>Representation in DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ammonia_assimilation</td>
<td>1.52%</td>
<td>0.25%</td>
</tr>
<tr>
<td></td>
<td>Photosystem_I</td>
<td>1.72%</td>
<td>0.58%</td>
</tr>
<tr>
<td></td>
<td>Proteorhodopsin</td>
<td>1.00%</td>
<td>0.03%</td>
</tr>
<tr>
<td></td>
<td>Ribosome_LSU_bacterial</td>
<td>3.04%</td>
<td>1.23%</td>
</tr>
<tr>
<td></td>
<td>Ribosome_SSU_bacterial</td>
<td>2.58%</td>
<td>0.79%</td>
</tr>
<tr>
<td>25m</td>
<td>Universal_GTPases (mostly elongation factors)</td>
<td>2.36%</td>
<td>1.31%</td>
</tr>
<tr>
<td></td>
<td>RNA_polymerase_bacterial</td>
<td>2.46%</td>
<td>1.25%</td>
</tr>
<tr>
<td></td>
<td>Transcription_initiation__bacterial_sigma_factors</td>
<td>0.80%</td>
<td>0.21%</td>
</tr>
<tr>
<td></td>
<td>Terminal_cytochrome_C_oxidases</td>
<td>1.60%</td>
<td>0.38%</td>
</tr>
<tr>
<td></td>
<td>Ubiquinone_Menaquinone-cytochrome_c_reductase_complexes</td>
<td>0.58%</td>
<td>0.11%</td>
</tr>
<tr>
<td></td>
<td>Oxidative_stress</td>
<td>0.90%</td>
<td>0.28%</td>
</tr>
<tr>
<td></td>
<td>Ammonia_assimilation</td>
<td>1.09%</td>
<td>0.26%</td>
</tr>
<tr>
<td></td>
<td>Photosystem_I</td>
<td>2.38%</td>
<td>0.66%</td>
</tr>
<tr>
<td></td>
<td>Photosystem_II</td>
<td>2.31%</td>
<td>0.81%</td>
</tr>
<tr>
<td></td>
<td>Proteorhodopsin</td>
<td>0.80%</td>
<td>0.03%</td>
</tr>
<tr>
<td></td>
<td>Ribosome_LSU_bacterial</td>
<td>2.90%</td>
<td>1.20%</td>
</tr>
<tr>
<td></td>
<td>Ribosome_SSU_bacterial</td>
<td>1.97%</td>
<td>0.79%</td>
</tr>
<tr>
<td>75m</td>
<td>CO2_uptake__carboxysome</td>
<td>1.20%</td>
<td>0.49%</td>
</tr>
<tr>
<td></td>
<td>Peptidoglycan_Biosynthesis</td>
<td>2.28%</td>
<td>1.24%</td>
</tr>
<tr>
<td></td>
<td>Chlorophyll_Biosynthesis</td>
<td>2.34%</td>
<td>0.87%</td>
</tr>
<tr>
<td></td>
<td>Photosystem_I</td>
<td>5.24%</td>
<td>0.37%</td>
</tr>
<tr>
<td></td>
<td>Photosystem_II</td>
<td>5.21%</td>
<td>0.46%</td>
</tr>
<tr>
<td></td>
<td>Proteorhodopsin</td>
<td>1.34%</td>
<td>0.04%</td>
</tr>
<tr>
<td></td>
<td>Ribosome_LSU_bacterial</td>
<td>3.92%</td>
<td>1.32%</td>
</tr>
<tr>
<td></td>
<td>Ribosome_SSU_bacterial</td>
<td>2.69%</td>
<td>0.77%</td>
</tr>
<tr>
<td></td>
<td>Universal_GTPases (mostly elongation factors)</td>
<td>3.05%</td>
<td>1.37%</td>
</tr>
<tr>
<td></td>
<td>F0F1-type_ATP_synthase</td>
<td>2.14%</td>
<td>0.92%</td>
</tr>
<tr>
<td></td>
<td>Cytochrome_B6-F_complex</td>
<td>0.86%</td>
<td>0.16%</td>
</tr>
<tr>
<td></td>
<td>Transport_of_Iron</td>
<td>1.78%</td>
<td>0.40%</td>
</tr>
<tr>
<td>125m</td>
<td>Peptidoglycan_Biosynthesis</td>
<td>4.63%</td>
<td>1.12%</td>
</tr>
<tr>
<td></td>
<td>Ammonia_assimilation</td>
<td>3.43%</td>
<td>0.12%</td>
</tr>
<tr>
<td></td>
<td>Ribosome_SSU_bacterial</td>
<td>1.41%</td>
<td>0.67%</td>
</tr>
<tr>
<td></td>
<td>Terminal_cytochrome_C_oxidases</td>
<td>1.55%</td>
<td>0.51%</td>
</tr>
</tbody>
</table>

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