MIT Open Access Articles

Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton

Massachusetts Institute of Technology

1

Metatranscriptomic analysis of autonomously collected and preserved marine

bacterioplankton

4

Elizabeth A. Ottesen[1], Roman Marin III[2], Christina M. Preston[2], Curtis R. Young[1], John

P. Ryan[2], Christopher A. Scholin[2], Edward F. DeLong*[1]

7

8

9

10

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of

Technology, Cambridge, MA 02139

[2]Monterey Bay Aquarium Research Institute, Moss Landing, CA 95039

14
15
16
17
*Corresponding author. Contact Information:
19
20
        Division of Biological Engineering & Department of Civil and
        Environmental Engineering
        Massachusetts Institute of Technology
        Parsons Laboratory 48-427
        15 Vassar Street
        Cambridge, MA 02139
        Phone: (617) 253-5271
        Email: delong@MIT.edu
29

30
Running Title: Autonomous sampling for marine metatranscriptomics

32 **Abstract**

33   Planktonic microbial activity and community structure is dynamic, and can change on

34 time scales of hours to days. Yet for logistical reasons, this temporal scale is typically under-

35 sampled in the marine environment.   In order to facilitate higher-resolution, long-term

36 observation of microbial diversity and activity, we developed a protocol for automated collection

37 and fixation of marine microbes using the Environmental Sample Processor (ESP) platform. The

38 protocol applies a preservative (RNALater) to cells collected on filters, for long-term storage and

39 preservation of total cellular RNA. Microbial samples preserved using this protocol yielded

40 high-quality RNA after thirty days of storage at room temperature, or onboard the ESP at *in situ*

41 temperatures.   Pyrosequencing of cDNA libraries generated from ESP-collected and preserved

42 samples yielded transcript abundance profiles nearly indistinguishable from those derived from

43 conventionally-treated replicate samples. To demonstrate the utility of the method, we used a

44 moored ESP to remotely and autonomously collect Monterey Bay seawater for

45 metatranscriptomic analysis. Community RNA was extracted and pyrosequenced from samples

46 collected at four time points over the course of a single day. In all four samples, the oxygenic

47 photoautotrophs were predominantly eukaryotic, while the bacterial community was dominated

48 by *Polaribacter*-like *Flavobacteria* and a *Rhodobacterales* bacterium sharing high similarity

49 with *Rhodobacterales* sp. HTCC2255. However, each time point was associated with distinct

50 species abundance and gene transcript profiles. These laboratory and field tests confirmed that

51 autonomous collection and preservation is a feasible and useful approach for characterizing the

52 expressed genes and environmental responses of marine microbial communities.

53

Introduction

Community sequencing techniques have become a prominent tool in microbial ecology. Marine environments have been the focus of major metagenomic surveys, which have provided insight into microbial gene content and community structure (DeLong et al 2006, Feingersch et al, 2009, Martin-Cuadrado et al 2007, Rusch et al 2007, Venter et al 2004). Metatranscriptomic analyses, with their focus on the transcriptional activity of the community, are also yielding insight into community function and gene regulation (Frias-Lopez et al 2008, Hewson et al, Poretsky et al 2009, Shi et al 2009). However, current protocols for such studies require manual sample processing and shipboard collections, and as a result sampling schemes are often limited by ship availability, shipboard sampling logistics, and expense. Given the importance of episodic nutrient delivery events in modulating biogeochemical cycles (Fasham et al 2001, Karl et al 2001, Karl 2002), new tools for observation and sampling of microbes are needed to facilitate observation of microbial processes at ecologically meaningful temporal and spatial scales.

The purpose of this study was to develop protocols for automated collection and preservation of samples for transcriptomic analysis using the Environmental Sample Processor (ESP), an automated platform for water sampling and molecular analysis. The ESP is an automated fluid handling system that collects and processes biological samples from seawater (Scholin et al 2009). Current real-time capabilities include array-based detection of target organisms including harmful algal species (Greenfield et al 2006, Greenfield et al 2008, Haywood et al 2007), invertebrate larvae (Goffredi et al 2006, Jones et al 2008), and major bacterial and archaeal clades (Preston et al 2009). Quantitative PCR capabilities are also currently in development (Scholin et al 2009). The ESP can return samples preserved in a saline-ethanol solution, but sample analysis has been primarily limited to *in situ* hybridization techniques (Goffredi et al 2006, Greenfield et al 2006, Greenfield et al 2008, Jones et al 2008). To extend the scope of laboratory analysis of archived samples, new protocols for sample preservation were developed for gene expression analysis and transcriptomics.

83   In this study, we validated and field-tested protocols for automated collection and

84 preservation of community mRNA from samples of marine bacterioplankton. Together with the

85 ESP platform (Scholin et al 2009), these protocols enable the autonomous collection of samples

86 and their return to the laboratory for transcriptional analysis. As the ESP also incorporates

87 standard oceanographic instrumentation and the ability to transmit collected data and receive

88 remote commands, environmental context monitoring and event response are also possible.

89 Following successful laboratory tests, our protocols were applied in field deployments. We used

90 the ESP and previously developed metatranscriptomic protocols to analyze microbial community

91 gene expression in Monterey Bay, a coastal system that has been the focus of previous molecular

92 microbial community analyses (Mincer et al 2007, O'Mullan and Ward 2005, Rich et al 2010,

93 Suzuki et al 2001, Suzuki et al 2004). Transcriptomic analysis of ESP-collected surface water

94 samples provided high resolution sequence data useful in determining the identity, relative

95 abundance, and expressed gene profiles of predominant marine bacterioplankton.

96

97 **Materials and Methods**

98 *ESP Operation for Sample Archival*

99   Only methods for ESP sample archival are presented here; for a full description of ESP

100 operation see Scholin et al (2009) and Roman et al (2007). Samples archived by the ESP for

101 metatranscriptomic analysis were collected in titanium sample "pucks" (filter holders) containing

102 the following 25 mm diameter filters stacked from top to bottom: a 5 μm Durapore prefilter

103 (Millipore), a 0.22 μm Durapore sample filter, 0.45 μm Metricell backing filter (Pall Gelman

104 Corporation), and a 10 μm sintered frit (Chand Eisenmann Metallurgical); see Greenfield et al.

105 (2006) for details. Pucks are stored in a rotating carousel in the ESP. A robotic mechanism

106 transfers fresh pucks from the carousel to a collection position where they are immobilized and

107 connected to the ESP's sample acquisition and reagent fluid handling system. Collection of

108 samples was achieved by drawing seawater through the stacked filters with repeated pulls of a 25

109 cc syringe. Between pulls, filtrate was exhausted back to the environment. The instrument

110   maintained a +10 psi pressure differential across the filter puck throughout sample collection.

111   Filtration continued until the desired sample volume (1 L) was reached, or until the flow rate fell

112   below 25 ml in 2.5 minutes (after which filtration was terminated and the filtered volume

113   recorded). The material retained on the filter was then preserved with two, 20 min incubations

114   with 1 ml of RNAlater (Ambion). Following sample collection and preservation, the puck was

115   removed from the collection station and returned to the storage carousel. The sample intake line

116   was then flushed with a 0.2% (v/v) sodium hypochlorite solution, which remained in the line

117   until the next sampling event. Immediately prior to collection of the next sample, the sample

118   intake line was flushed with a dilute Tween-20 solution (0.05 %, v/v), a fresh puck was loaded

119   from the carousel and the sample archival procedure repeated. Pucks remained onboard the

120   instrument (at *in situ* temperatures and under an N_2 gas atmosphere) until the end of the full

121   deployment. After deployment, the 5 and 0.22 μm filters were recovered and stored at -80 °C

122   until use. Metagenomic and metaranscriptomic studies were performed only on the 0.22 μm

123   filters.

124   To mimic *in situ* sampling conditions, seawater collected for processing in the laboratory

125   was loaded into a dispensing pressure vessel (Millipore), attached to the intake and exhaust

126   valves of the ESP and pressurized to 20 psi, to simulate conditions at ~18 m from the sea surface.

127   Collection and processing then proceeded as described above.

128

129   *Validation of Fixation Protocols*

130   To examine the long-term stability of RNA following fixation, a near-shore sample was

131   collected from the Monterey Bay Municipal Pier and pre-filtered through a 10 μm Nitex mesh.

132   The ESP then collected and preserved five replicate samples using the protocol described above.

133   Following collection of all samples, filter pucks were removed from the instrument and stored at

134   room temperature under conditions that mimic deployment (in a sealed container with an N_2

135   atmosphere and damp paper towels to generate humidity). Pucks were retrieved at one week

136   intervals, and the sample filters removed and stored at -80 °C. Following completion of the time

series, total RNA was extracted from all 0.22 μm filters simultaneously as described below. The integrity of the recovered RNA was evaluated using a Bioanalyzer high-sensitivity electrophoresis system and the RNA 6000 Pico mRNA protocol (Agilent).

To evaluate the effect of long-term preservation on metatranscriptomic profiles, an ESP-collected and preserved sample was compared against a sample collected by vacuum filtration. Seawater was collected from the Santa Cruz (California) municipal wharf with one aliquot dedicated for ESP processing (as above) and a second for collection using traditional laboratory vacuum filtration. The material collected using the latter method was immediately flash frozen in liquid nitrogen, and stored at -80 °C until extraction. The ESP-processed aliquot of that same sample remained on the instrument during a field test of the ESP at MBARI's station M0 (36.83 °N, 121.90 °W) from April 6, 2009 to April 29, 2009. Following recovery of the instrument, both ESP- and vacuum-collected samples were extracted and used for metatranscriptomic analysis as described below. To further validate collection and preservation protocols, samples collected *in situ* over the course of the deployment were also used for community RNA extraction and the RNA quality examined by size fractionation.

To further evaluate the effect of automated collection and preservation on metatranscriptomic profiles, ESP-collected and preserved samples were compared to a replicate sample processed by peristaltic pump filtration. Water was collected in Monterey Bay (36.7173 °N, 122.1147 °W, ~9 km from MBARI station M1) at 00:13 local time on June 8, 2009 by CTD rosette at 30 m depth. 1 L was processed by ship-board ESP under normal conditions (~ 1 hr sample time). A replicate sample of 0.5 L was processed for RNA extraction using a standard peristaltic pump filtration protocol, as described previously (Frias-Lopez et al 2008). Following peristaltic pump filtration, the 0.22 μm filter was immediately submersed in 300 uL of RNA Later and stored at -80 °C, with a total time of ~30 min from seawater collection to preservation. To provide DNA template for synthesis of sample-specific rRNA subtractive hybridization probes, an additional 9.8 L of water from the same sample was collected for DNA extraction with the same prefilter but using a 0.22 μm pore size Sterivex filter (Millipore). The Sterivex

164   filter was subsequently filled with 2 mL lysis buffer (50 mM Tris-Hcl, 40 mM EDTA, and 0.75

165   M sucrose) and stored at -80 °C.

166

167   *Field Tests in Monterey Bay*

168   Monterey Bay time series samples were collected during a deployment of the ESP at

169   MBARI's station M0 from May 14 to June 11, 2009. Sample pucks from that deployment were

170   recovered and processed (filters separated and placed in sterile tubes at -80 °C) on June 12. The

171   AUV *Dorado* was repeatedly deployed from June 2-4, 2009, to survey water masses and

172   phytoplankton variability in an area surrounding the ESP mooring. Details of AUV sensors,

173   operation, and data processing can be found in Ryan et al. (2010b).

174   DNA for metagenomic sequencing and rRNA probe synthesis was extracted from

175   seawater collected by CTD rosette from a ship in close proximity to the ESP on June 2, 2009 at

176   0830 local time. A seawater sample collected by the AUV *Dorodo* using the *Gulper* water

177   sampling system (Ryan et al 2010a) on June 4th was used for DNA extraction and rRNA probe

178   synthesis, but was not sequenced. Both DNA samples were filtered using Sterivex filters and

179   stored in lysis buffer as described in the previous section.

180

181   *Nucleic Acid Extraction and Subtractive Hybridization*

182   Total RNA was extracted from filters as described previously (Frias-Lopez et al 2008).

183   Briefly, community RNA was extracted using the *mir*Vana kit (Ambion). Turbo DNase

184   (Ambion) was used to remove genomic DNA, and the resulting samples purified and

185   concentrated using the RNeasy MinElute cleanup kit (Qiagen). RNA extraction yields for all

186   samples are summarized in Table 1.

187   DNA was extracted and purified using the QuickGene 610L system (Fujifilm) and DNA

188   Tissue Kit L with a modified lysis protocol. 2 mg of lysozyme in lysis buffer (described above)

189   was added to thawed Sterivex filters, which were incubated with rotation to mix at 37 °C for 45

190   min. 100 µL each buffers DET and MDT were added, and the sample incubated at 55 °C for 2

191 hr. with rotation. The lysate was decanted from the filter, 2 ml LDT solution, and incubated at

192 55 °C for a further 15 min. 2.7 ml EtOH was added, and the sample loaded onto the QuickGene

193 instrument for purification according to the DNA Tissue protocol.

194   Antisense rRNA probes for subtractive hybridization were prepared as described

195 previously (Stewart et al 2010). In brief, universal bacterial, archaeal, and eukaryotic SSU and

196 LSU primers with attached T7 promoters were used in PCR reactions with Herculase II Fusion

197 DNA polymerase (Stratagene) to generate templates for antisense-rRNA probe synthesis.

198 Biotin-labeled antisense rRNA probes were generated from the PCR products using the

199 MegaScript T7 kit (Ambion). The Santa Cruz municipal pier samples used for the vacuum/ESP

200 comparison lacked a paired DNA sample, so PCR was instead performed on first-strand cDNA

201 prepared with the SuperScript III kit (Invitrogen) with random primers and 40 ng of the ESP-

202 collected total RNA sample.

203   Subtractive hybridization was carried out using published protocols (Stewart et al 2010).

204 Hybridization reactions were carried out on 200 ng of total RNA and sample-specific antisense

205 rRNA probe mixtures. For the Santa Cruz wharf vacuum/ESP comparison, the original, two-step

206 hybridization protocol was followed, using 200 ng total RNA and 250 ng each of the SSU and

207 LSU rRNA bacterial probes. For the remaining samples, the amended protocol presented in the

208 supplementary materials of that paper was utilized. Station M1 samples utilized probes

209 synthesized from the paired DNA sample, with 400 ng each bacterial SSU and LSU, 200 ng each

210 archaeal SSU and LSU, and 300ng each eukaryotic SSU and LSU. Station M0 samples utilized

211 probes generated from both the June 2 (CTD-collected) and June 4 (AUV-collected) DNA

212 samples at 0.75X concentration (eg. 300ng June 2 bacterial SSU + 300ng June 4 bacterial SSU).

213 For the station M0 samples, the archaeal SSU rRNA primers exhibited non-specific

214 amplification, and as a result no archaeal SSU rRNA probes were included. rRNA-probe

215 duplexes were subsequently bound to Strepavidin-coated magnetic beads, and removed from the

216 total RNA preparation. Following this procedure, samples were purified and concentrated using

217 the RNeasy MinElute cleanup kit.

218   rRNA-subtracted (and unsubtracted total RNA) samples were amplified as described

219   previously (Frias-Lopez et al 2008). In brief, RNA was amplified using the MessageAmp II

220   Bacteria kit (Ambion) and a poly-T primer with an additional 5´ BpmI restriction site. First-

221   strand cDNA was synthesized from the aRNA using random primers and SuperScript III

222   (Invitrogen), second-strand cDNA synthesized using DNA pol I, *E. coli* ligase, and T4 DNA

223   polymerase (Invitrogen), and remaining poly-A tails removed by digestion with BpmI (NEB).

224   All samples were sequenced using the 454 Genome Sequencer (Roche).

225   Metatranscriptomic samples were prepared and sequenced using the GS FLX protocol, and the

226   metagenomic DNA sample using the GS FLX Titanium protocol. Library preparation and

227   sequencing was carried out according to the manufacturer's protocols.

228

229   *Sequence Processing and Annotation*

230   Sequencing and annotation statistics for each sample are summarized in Table 2.

231   Sequences derived from rRNA were identified using BLASTN with a bit score cutoff of 50

232   against a database composed of 5S, 16S, 18S, 23S, and 28S rRNA sequences from microbial

233   genomes and the SILVA LSU and SSU databases (www.arb-silva.de). Non-rRNA sequences

234   with identical start sites (first 3 bp), 99% identity, and < 1 bp length difference were identified as

235   probable artificially duplicated sequences (Stewart et al 2010) and removed using the cd-hit

236   program (Li and Godzik 2006) and scripts developed by Gomez-Alvarez et al. (2009). Non-

237   rRNA sequences were compared to the November 3, 2009 version of NCBI's non-redundant (nr)

238   protein database reference databases using BLASTX. Unless otherwise specified, a bit score

239   cutoff of 50 was used to identify significant matches to the database.

240   For pairwise comparisons of metatranscriptomic profiles, each sequence was assigned to

241   a single reference gene in the NCBI-nr database based on BLASTX alignment bit score. When a

242   single sequence aligned equally well to multiple potential reference genes, it was assigned to the

243   reference gene that was most frequently identified in the dataset. Reference gene abundances

244   were compared using the cumulative form of the AC Test (Audic and Claverie 1997) and an

245   FDR correction for multiple comparisons (Benjamini and Hochberg 1995); details of how these

246   tests were conducted are in the supplementary online materials.

247   The MEGAN program (Huson et al 2007) was used to assign sequences to a higher-order

248   taxonomy. All analyses used a bit score cutoff of 50 and database matches with bit scores within

249   10% of the top-scoring hit. Unique non-rRNA sequences from both subtracted and unsubtracted

250   sequence datasets were pooled and assigned to the NCBI taxonomy based on the results of a

251   BLASTX search of the NCBI nr database. rRNA genes were assigned to the NCBI taxonomy

252   using manually curated rRNA databases constructed based on the approach used by Urich et al.

253   (2008) as described in the supplementary online materials. Only rRNA-unsubtracted samples

254   were utilized in taxonomic analysis of putative rRNA sequences. Due to the higher copy number

255   and lower genetic diversity of rRNA genes compared to mRNA genes, rRNA taxonomies were

256   constructed without removal of duplicates.

257   Analyses of gene expression in the *Rhodobacterales* sp. HTCC2255 and *Polaribacter*

258   taxonomic groups used all sequences for which the taxon in question was among the top-scoring

259   database matches (all matches with bit scores equal to the highest-scoring alignment were

260   considered). For composite analyses of *Polaribacter* expression, read counts and annotations for

261   genes shared by both *Polaribacter irgensii* 23-P and *Polaribacter* sp. MED152 were combined.

262   Shared genes were defined as reciprocal best BLASTP hits with e-value $< 1 \times 10^{-5}$, and at least

263   80% alignment coverage for both genes. The draft version of the *Rhodobacterales* sp.

264   HTCC2255 genome has a large number of contigs that were annotated as contamination and

265   removed from the genome scaffolds. These contigs (and the 2267 CDS identified within them)

266   were not identified in surveys of Monterey Bay and were excluded from our genome analyses.

267

268   *Functional Comparison using KEGG Gene Categories*

269   For comparisons of differences in community function, sequences were assigned to

270   functional categories based on KEGG orthology groups (Kanehisa and Goto 2000). For bulk

271   community-level analyses, sequence reads were assigned a single reference gene in the Nov. 7,

272   2009 version of the KEGG database as described above for the NCBI-nr database, with the

273   additional weighting factor that proteins that were assigned to a KEGG ortholog category were

274   preferred when choosing between multiple matches with identical alignment scores and

275   frequencies in the dataset. For analyses focused on *Polaribacter* and HTCC2255, the KAAS

276   automated annotation pipeline (Moriya et al 2007) was used to annotate each reference genome.

277   KEGG ortholog counts for each taxon were then compiled using all sequences for which the

278   taxon in question is among the top-scoring hits by BLASTX against NCBI-nr database (all hits

279   with bit scores equal to the highest-scoring alignment).

280   KEGG Pathway counts were generated based on the total number of sequences assigned

281   to KO annotations within that pathway (due to functional overlap, some orthologs were

282   represented in multiple pathways). All comparisons used KEGG Pathway rather than BRITE

283   hierarchies, and pathways within the "Human Diseases" or "Organismal Systems" hierarchies

284   were not analyzed. Both ortholog and pathway counts for each sample were normalized to the

285   total number of non-rRNA sequences with significant hits to the KEGG database (for bulk

286   community analyses) or to the total number of sequences assigned to the taxon in question

287   (analyses focused on HTCC2255 or *Polaribacter*). Statistical evaluation of KEGG pathway

288   abundances used in-house R scripts utilizing a methodology explained in detail in the

289   supplementary online materials.

290

291   **Results and Discussion**

292   *Validation of ESP Preservation Protocols*

293   Following collection and preservation on the ESP, marine bacterioplankton samples were

294   found to be stable for at least four weeks (Figure 1). In one experiment, replicate samples of a

295   near-shore surface water sample were filtered and preserved using an ESP in the laboratory and

296   stored under ESP-like conditions at room temperature (high humidity, N_2 atmosphere). Each of

297 the five samples that were processed at weekly intervals had similar yields (Table 1), and gel

298 electrophoresis indicated comparable RNA quality (discrete 16S and 23S rRNA peaks with no

299 obvious degradation) (Figure 1a). In a separate field test, RNA was extracted from cells that had

300 been filtered and fixed by the ESP over the course of a 29-day deployment in Monterey Bay.

301 These *in situ* filtered and fixed cells also yielded high yields of RNA (Table 1) with comparable

302 high quality over all time points sampled (Figure 1b). We conclude that the ESP sampling and

303 preservation protocol can provides material that is sufficient for downstream extraction of high

304 quality, high yield total cellular RNA.

305   To evaluate preservation of mRNA, a near-shore surface water sample taken from the

306 municipal wharf in Santa Cruz, California was processed via the ESP for filtration and fixation,

307 and then subjected to 29 days of storage on board the instrument during a deployment in

308 Monterey Bay. This sample, and a control sample that was filtered in parallel by vacuum

309 filtration and immediately flash frozen, were used to prepare a community transcriptome

310 pyrosequencing library. Transcript abundance profiles for the conventionally-processed flash-

311 frozen sample and the 29-day ESP-preserved sample were highly similar, with only 6 out of

312 17,284 transcripts showing significantly different abundances in the two samples (Figure 2,

313 Table 3). An ESP-processed and preserved sample collected near MBARI station M1 in

314 Monterey Bay was also compared to a replicate sample processed by rapid peristaltic pump

315 filtration. This pair of samples again yielded similar transcript abundance profiles, with

316 significant differences for 28 out of 35036 reference genes (Figure S1, Table S1).

317   Both of our comparisons of ESP- and manually-processed samples yielded numbers of

318 significantly different references, and percentages of sequence reads mapping to significantly

319 different transcripts, that were within the range typically observed for technically replicated

320 metatranscriptomic profiles (Stewart et al, 2010; see Table S2 for comparisons recalculated using

321 our updated database and cutoffs). In fact, the sample subjected to long-term storage on board

322 the instrument (Santa Cruz wharf ESP) was more similar to its manually processed control (in

323 terms of taxonomic composition and significantly different transcripts), than were the paired

324 ship-board samples (Station M1 PP and ESP), in which the ESP-collected sample was

325 immediately removed from the instrument and stored at -80 °C (Table 4). This suggests that the

326 water filtration method used may have more impact on the observed metatranscriptomic profile

327 than the length of time between preservation and removal to low-temperature, permanent

328 storage.

329

330 *Metatranscriptomic Analysis of Monterey Bay Surface Water Samples*

331 Four time points during a deployment of the ESP in Monterey Bay surface waters were

332 chosen for transcriptomic sequencing. These time points represented transcriptomic profiles of

333 microbial communities at dawn (5-6 am), late morning (10-11 am), dusk (6-7 pm), and night (10-

334 11 pm) on June 2, 2009. As the ESP was deployed in a moored configuration, these samples

335 represented distinct microbial populations from water masses with different physical and

336 chemical conditions (Figures S2 and S3, Table S3). To provide genomic context and additional

337 information on population structure, community DNA was extracted and sequenced from a 10L

338 water sample collected near the ESP on June 2, 2009.

339

340 *Taxonomic Composition of Station M0 Metagenomic and Metatranscriptomic Samples*

341 Metagenomic and metatranscriptomic sequences were assigned to taxonomic groups

342 using MEGAN (Huson et al 2007) to analyze both protein-coding regions and rRNA small

343 (SSU) and large (LSU) subunit sequences (Figure 3). The metagenomic dataset showed a

344 similar taxonomic composition among all transcript types, with a primarily bacterial community

345 dominated by flavobacteria and alpha- and gamma-proteobacteria. Cyanobacterial transcripts

346 represented only a small proportion of the library, suggesting that primary production in these

347 samples was dominated by eukaryotic phytoplankton. A diversity of transcripts from eukaryotic

348 nano- and pico-plankton capable of passing through the 5um prefilter were detected in the

349 samples. SSU rRNA-based analysis showed a higher representation of eukaryotic organisms in

350 metatranscriptomic samples than in the metagenome, consistent with previous observations that

351 eukaryotic picoplankton exhibit higher transcriptional activity relative to genomic abundance

352 than bacteria (Man-Aharonovich et al 2010). In taxonomic analyses of LSU rRNA and coding

353 sequences, eukaryotes represented a much smaller proportion of assigned sequences than

354 expected based on the SSU rRNA results, but this may be due to limited database coverage of

355 marine picoeukaryotes.

356 The annotated protein-coding transcript pool contained two particularly abundant

357 organisms/groups. The first was an alpha proteobacterium closely related to *Rhodobacterales* sp.

358 HTCC2255. HTCC2255 is a proteorhodopsin-containing *Roseobacter* of the NAC11-7 clade

359 (Newton et al 2010, Yooseph et al 2010) isolated by dilution-to-extinction near the coast of

360 Oregon as part of the high throughput culture collection (Connon and Giovannoni 2002).

361 HTCC2255 was also the most frequently identified taxon in both of the water samples used for

362 validation studies, constituting >25% of assignable non-rRNA sequences in the two Santa Cruz

363 wharf transcriptomes and ~10% of assignable sequences from the Station M1 transcriptomes

364 (Figure S4). The second was a *Flavobacterium* most closely related to the two sequenced

365 *Polaribacter* species, *Polaribacter* sp. MED152 (González et al 2008) and *P. irgensii* 23-P

366 (Gosink et al 1998). These two reference genomes together represented 8-9% of sequences from

367 the Santa Cruz wharf, but < 2% of sequences from the Station M1 samples.

368 Other bacteria representing significant proportions of the community included gamma

369 proteobacterium HTCC2207 (Stingl et al 2007), a diversity of *Flavobacteria* including

370 *Flavobacterium* MS024-2A (Woyke et al 2009), and eubacterium SCB49 (Yooseph et al 2010).

371 *Ostreococcus* (Derelle et al 2006, Palenik et al 2007) and *Micromonas* (Worden et al 2009)

372 picoplankton were among the most abundant eukaryotic sequences identified in annotated

373 transcripts, but rRNA-based analysis suggests a large diversity of eukaryotes without sequenced

374 genomes were present in the samples.

375 Across the four time points, the community composition remained broadly similar, with

376 Whittaker's index of association (Whittaker 1952) values between 0.68-0.85 for NCBI taxon

377 counts (Table S2). The five most abundant bacterial taxa dominated in all four samples (Figure

378   S5), although the picoeukaryote *Ostreococcus lucimarinus* showed more dramatic shifts in

379   abundance between the different samples. All taxa present at >0.02% of uniquely assignable

380   transcripts in at least one sample were detected in all four samples. However, the relative

381   abundances of these organisms varied dramatically. One of the largest changes was in *O.*

382   *lucimarinus*-like sequences, which ranged from 0.2% to 6.2% of transcripts, while MED152-like

383   sequences ranged from 5.5-16.8% and HTCC2255-like sequences from 9.12-16.7%. This

384   variability is greater than the differences seen in comparisons of abundant taxa in transcriptomic

385   profiles at different depths in the euphotic zone at station ALOHA (Shi et al 2010) or of day and

386   night samples collected two days apart at a single station in the North Pacific Subtropical gyre

387   (Poretsky et al 2009). However, such comparisons cannot independently discriminate between

388   changes in organism abundance and changes in activity level (cellular RNA content). For

389   example, in incubation experiments where dimethylsulfoniopropionate (DMSP) was added to

390   oligotrophic waters, noticeable shifts in transcript taxonomic composition were observed within

391   30 min, presumably too short an interval for extensive growth (Vila-Costa et al 2010).

392

393   *Rhodobacterales sp. HTCC2255*

394   HTCC2255-like transcripts were abundant at all four time points. HTCC2255 was

395   among the top-scoring alignments in the NCBI-nr database for a total of 12%, 14%, 15%, and

396   8% of assignable sequences in the 5am, 10am, 6pm, and 10pm metatranscriptomic datasets, and

397   a corresponding 9%, 10%, 11%, and 6% of sequences were unambiguously assigned to

398   HTCC2255 by the Megan LCA algorithm. HTCC2255-like protein coding genes averaged 95%

399   identity to the sequenced strain at the amino acid level (Figure 4). 2092 out of 2240 of

400   HTCC2255 coding sequences were identified in at least one of the four Station M0

401   transcriptomes, including genes involved in DMSP degradation, the full TCA cycle,

402   proteorhodopsin, and sulfur oxidation. The draft genome sequence of HTCC2255 has two

403   scaffolds, NZ_DS022282 and NZ_DS022288, of which the smaller, NZ_DS022288 has much

404   lower coverage in both the metagenome and the transcriptomes (Figure S6). This may represent

405 a less conserved genomic island, or a plasmid that is missing from the Monterey Bay genotype.

406 Alternatively, the metadata associated with the draft genome notes that contaminating gamma

407 proteobacteria sequences were present in the raw sequence data and removed from the scaffolds

408 during draft assembly; NZ_DS022288 may represent a scaffold that was inappropriately

409 assigned to HTCC2255.

410 HTCC2255 appears to be a very common component of the Monterey Bay microbial

411 community. Several BAC sequences previously isolated from Monterey Bay (Rich et al 2010,

412 Rich et al 2008) share a high percent identity and synteny with HTCC2255, and probes targeting

413 these BAC clones and the HTCC2255 reference genome identified these organisms in 97 -100%

414 of surface water samples spanning a 4 year time series at Monterey Bay station M1 (Rich et al

415 2010). Additionally, the HTCC2255-like BAC EB000-55B11 was also detected in near-shore

416 samples from Woods Hole, MA during experiments with the prototype genome-proxy array

417 (Rich et al 2008). In general, HTCC2255-like organisms appear to be widely present in marine

418 communities (Yooseph et al 2010), but may be particularly abundant in near-shore waters. The

419 reference strain was isolated off the Oregon coast, relatives have been detected off both the

420 California and Massachusetts coasts, and HTCC2255-like sequences were reported as the most

421 abundant sequence type in a proteorhodopsin library from the North Sea (Riedel et al 2010).

422

423 *Polaribacter-like sequences*

424 *Polaribacter*-like transcripts were identified at all four time points, but were most

425 abundant in the 5am sample. A *Polaribacter*-derived sequence was among the top-scoring

426 alignments in the NCBI-nr database for 27%, 12%, 12%, and 9% of assignable sequences in the

427 5am, 10am, 6pm, and 10pm metatranscriptomic datasets, and a corresponding 15%, 6%, 5%, and

428 5% were unambiguously assigned to this genus by the Megan LCA algorithm. These sequences

429 appear to represent one or more unsequenced organisms related to *Polaribacter*, as sequence

430 reads with top database hits to a *Polaribacter* sp. averaged 82% amino acid identity to *P. irgensii*

431 23P and 83% identity to *Polaribacter* sp. MED152 (Figure 4). The two sequenced organisms

432 average 72% amino acid identity among their shared genes (defined as reciprocal best hits with

433 e-value < 1 x 10^{-5} and 80% of the gene aligned). 1519 out of 1636 shared *Polaribacter* genes

434 were identified in one or more of the four transcriptomic samples, while 361 of 974 MED152-

435 specific and 259 of 920 23P-specific genes were identified (Figures S7 and S8).

436   There is less prior evidence for *Polaribacter* as a common component of the Monterey

437 Bay bacterioplankton community than there is for HTCC2255. *Polaribacter*-like sequences

438 were not detected in experiments using genome proxy arrays to examine community structure at

439 Monterey Bay station M1, despite inclusion of both sequenced *Polaribacter* genotypes in the

440 array (Rich et al 2010). However, the sequences recovered in this study averaged ~82% amino

441 acid identity to the reference strains, which is too genetically dissimilar to show strong

442 hybridization signal on the array (Rich et al 2010, Rich et al 2008). Additionally, our study used

443 a 5 μm prefilter during sample collection, while the array experiments used a 1.6 μm prefilter,

444 which may change the representation of larger and/or particle-attached bacterial cells.

445 *Polaribacter*-like sequences were identified in 16S rRNA libraries prepared from samples

446 collected at station M0 in September-October 2004, during development of their sandwich

447 hybridization assay (Preston et al 2009). Other studies of Monterey Bay have not specifically

448 examined abundance of *Flavobacteria* or *Polaribacter*, although Suzuki et al. (2004) identified

449 the CFB group as representing 8.5% of bacteria in a surface water sample. However,

450 *Flavobacteria* and the CFB group as a whole is thought to play a major role in degradation of

451 particulate and high molecular weight dissolved organic matter in the ocean (Kirchman 2002),

452 and *Polaribacter* was found to be the most abundant *Flavobacterial* group across a transect in

453 the North Atlantic (Gómez-Pereira et al 2010).

454

455 *Nutrient acquisition strategies of Monterey Bay Rhodobacterales and Polaribacter*

456   Although both HTCC2255 and the two *Polaribacter* reference strains are

457 proteorhodopsin-bearing heterotrophs, their genome characteristics and transcriptional activity

458 are consistent with distinctly different nutrient acquisition strategies. Among the most highly

459   expressed HTCC2255-like genes were substrate binding proteins associated with TRAP and

460   ABC transporters of amino acids, sugars and sugar alcohols (Table S4 and Table S5).   In

461   contrast, González et al. (2008) found that *Polaribacter* sp. MED152 carries relatively few

462   transporters for free amino acids or sugars, and no sugar-specific ABC transporters.   Both

463   *Polaribacter* genomes appear to lack TRAP and TRAP-T transporters, and few transcripts

464   mapped to those ABC transporters they do carry (<1% of *Polaribacter*-like sequences in contrast

465   to 10-14% HTCC2255-like sequences; details in Table S6).   Similarly, in an examination of

466   transporters in a coastal transcriptome, Poretsky et al. (2010) found an abundance of

467   *Rhodobacterales* and SAR11 associated ABC- and TRAP-related transcripts in coastal

468   environmental transcriptomes, and relatively few *Flavobacterales* associated sequences.   A

469   metaproteomic investigation of SAR11 in the Sargasso Sea (Sowell et al 2009) found that

470   transport functions similarly dominated the proteome of that alpha proteobacterium, with the

471   most abundant proteins being ABC and TRAP transporters.   This is consistent with previous

472   studies showing that *Alphaproteobacteria* dominate uptake of amino acids and monomers, while

473   *Bacteroidetes* specialize in utilization of polymers (Cottrell and Kirchman 2000, Kirchman

474   2002).

475   The most abundant group of transport-related transcripts within the *Polaribacter*-

476   associated sequences are TonB-dependent/ligand gated channels (Table S7).   TonB-dependent

477   channels were also among the most highly expressed proteins from *gamma proteobacterium*

478   HTCC2207 and *Flavobacterium* MS024-2A. *Rhodobacterales* sp. HTCC2255 appears to lack a

479   TonB system; the draft genome contained no significant hits to the pfam profiles of either TonB

480   or the two TonB-dependent receptor domains.   TonB-dependent transporters from a variety of

481   taxonomic groups were the most abundant family of membrane proteins identified in a

482   metaproteomic analysis of samples from the South Atlantic (Morris et al 2010).   TonB-related

483   proteins were also identified as among the most abundant transcripts assigned to DOM-

484   responsive *Idiomarinaceae* and *Alteromonadacea* in a DOM-enriched marine microcosm

485   (McCarren et al 2010).   As mentioned above, ABC transporters from *Flavobacterales* were not

486   found to be abundant in a coastal metatranscriptome (Poretsky et al 2010). However, the authors

487   did mention an abundance of *Flavobacterales* transporters for inorganic compounds, and

488   COG1629, which includes some TonB transporters, is included in this group under the COG

489   classification scheme. TonB-dependent channels were originally identified in the context of iron

490   transport, but have since been associated with the transport of a large variety of compounds

491   (Schauer et al 2008). Of particular interest is their newly recognized association with

492   degradation of polymers and complex carbohydrates (Blanvillain et al 2007, Martens et al 2009).

493   Several of the TonB-dependent channels in *Polaribacter* sp, MED152 were associated with

494   predicted peptidases and glycosyl hydrolases (González et al 2008), suggesting they may be

495   involved in utilization of high molecular weight substrates in this organism.

496

497   *Differences in Functional and Metabolic Profiles of Metatranscriptomic Samples*

498   In order to examine transcript abundance dynamics, we used KEGG pathways to

499   functionally profile the four transcriptomic samples. KEGG profiles were generated for bulk

500   community data, and individually for *Rhodobacterales* sp. HTCC2255-like and *Polaribacter*-like

501   transcripts. Metabolic genes represented the most abundant class of annotated transcripts,

502   although they were out-numbered by unassigned transcripts for both the bulk community and

503   *Polaribacter*-like sequences (Table 5). Transcripts involved in environmental information

504   processing were more abundant in the HTCC2255 fraction compared to the bulk community (14-

505   19% vs 4.4-4.9% of transcripts); this signal was dominated by ABC transporters.

506   KEGG-annotated genes could be assigned to 188 (bulk community), 127 (HTCC2255),

507   and 129 (*Polaribacter*) pathways, excluding BRITE hierarchies and pathways associated with

508   human diseases and organismal systems. Of these, 133, 76, and 69 had significantly different

509   levels of expression between samples (Table S8 and S9). Many central metabolic pathways,

510   including oxidative phosphorylation, photosynthesis, the TCA cycle, pyruvate metabolism, and

511   glycolysis were overrepresented in the 10am sample (Figure 5), both at the bulk community and

512   taxon-specific levels. In contrast, ribosomal proteins peaked in morning and evening samples,

513 and had different maxima and minima for the bulk community, HTCC2255 and *Polaribacter*

514 groups. RNA polymerase also displayed different trends among the three groups. Given the

515 changes in taxonomic composition discussed above, we cannot rule out population-level effects

516 in comparisons of relative transcript abundance. However, the synchronous changes in central

517 carbohydrate metabolism and energy metabolism suggests that these pathways may be tuned to

518 broader environmental factors, while the dynamics of translation and transcription suggest more

519 population-specific controls.

520 Although complicating population-specific factors cannot be ruled out, these

521 transcriptional profiles suggest that light may play a role in the metabolism of *Rhodobacterales*

522 sp. HTCC2255 and *Polaribacter*. In recent years, a number of potential mechanisms by which

523 light might influence the metabolism of heterotrophs in the ocean have been discovered (Béjà et

524 al 2000, Béjà et al 2002, Kolber et al 2001, Venter et al 2004). Proteorhodopsin made up 0.18-

525 0.82% of HTCC2255-like and 0.42-1.18% of *Polaribacter*-like transcripts. Surprisingly, in both

526 groups the representation of proteorhodopsin transcripts was highest in the nighttime (10pm)

527 sample. *Dokdonia* sp. MED134, another proteorhodopin-carrying flavobacterium, has been

528 demonstrated to have higher levels of proteorhodopsin in light-incubated vs dark-incubated

529 cultures, but these changes were examined at time scales of 3-13 days, not hours as in this study

530 (Gómez-Consarnau et al 2007). Proteorhodopsin was one of the most abundant transcripts

531 associated with *gamma proteobacterium* HTCC2207 (0.55-0.91% of HTCC2207 transcripts) and

532 *Flavobacterium* MS024-2A (0.86-3.9% of MS024-2A transcripts). HTCC2207-like

533 proteorhodopsin transcripts appeared most abundant at night, but MS024-2A-like

534 proteorhodopsin transcripts had the highest relative abundance at 10am. However, while the

535 daytime increase in proteorhodopsin expression for MS024-2A was significant in the context of

536 the total number of transcripts assigned to this organism, the relatively low coverage (1610-5210

537 assigned sequences) precludes rigorous transcriptional analysis.

538 Even in the absence of photo-regulation of proteorhodopsin expression, the HTCC2255-

539 like transcripts did exhibit potential light-dependent changes in energy metabolism. In

540 particular, HTCC2255-like F-type ATP synthase transcripts appeared to be downregulated at

541 night; 5 out of 9 ATP synthase-associated proteins had significant differences in abundance, and

542 all were least abundant in the 10pm sample. This may indicate light-dependent changes in the

543 cross-membrane proton gradient. In contrast, although ATP synthase as a whole (photosynthesis

544 pathway in Fig 5) showed a slight but significant decrease in transcript abundance at night for

545 *Polaribacter*-like sequences, only one subunit exhibited a significant change in expression at the

546 transcript level. One explanation for this difference in light-dependence of ATP synthase

547 expression is that HTCC2255-like organisms may be more dependent on ATP to power transport

548 than *Polaribacter*, as a result of the expanded use of ABC transporters in this organism (TonB-

549 dependent transporters utilize the proton gradient directly). More broadly, the transcriptomic

550 profile of HTCC2255-like microbes at 10pm showed a larger decrease in many metabolic

551 activities than seen in co-ocurring *Polaribacter*-like bacteria, suggesting potentially greater light

552 regulation in this organism. Alternatively, the particular population of HTCC2255-like

553 organisms sampled during the 10pm timepoint could have been in a lower-energy metabolic

554 state, for reasons independent of the time of day.

555

556 **Implications**

557 While a number of metatranscriptomic studies of marine microbial communities have

558 been conducted, most have represented single or relatively few time points that were spatially

559 segregated. Although these studies have proven useful for general surveys of expressed genes

560 and non-coding RNAs, the utility of such comparisons is limited in the absence of data on the

561 spatial and temporal scales of natural environmental variability. In this study, we demonstrate

562 the feasibility of *in situ* autonomous collection of metatranscriptomic samples using the ESP

563 platform, along with synoptic data on environmental conditions. A distinct advantage of this

564 approach is that it allowed longer term deployments and continuous monitoring of environmental

565 fluctuations over the full time course of multiple sample collections. Consistent with the known

566 dynamic variability in coastal systems, our observations reflected continuously changing

conditions, consistent with high current velocity at the sample site. Each of the four metatranscriptomic datasets thus represents a different water mass, and a distinct microbial community. Although similar taxa were present in each time point, these taxa showed different bulk activity levels (as reflected in rRNA and mRNA abundances) and expression profiles in each of the four samples. As a result, although differences in gene expression levels could be observed, it was difficult to differentiate changes that reflect the specific water masses and microbial populations sampled, versus discrete organismal responses to broader environmental parameters (such as time of day). Our study demonstrates that such effects are large enough to require serious consideration, even when a fixed location is sampled across relatively short (24 hr) time scales.

Automated sample collection has the potential to greatly reduce the costs associated with long-term environmental monitoring, allowing longer duration and/or higher frequency sampling schemes. Different deployment schemes for the sampler, for example on "drifters" or autonomous vehicles, may also facilitate short-term temporal sampling within coherent water masses. This may prove to be important for developing a picture of the temporal and spatial scales of natural variability in microbial populations. In this study, we found that the identities of the most abundant microbial populations did not shift dramatically in samples collected over the course of a day, but their relative abundances did. With ESP technology, it will be possible to examine such differences over longer time scales, and using different sampling modes (e.g. Lagrangian versus Eularian) to better correlate changes in environmental conditions with shifts in microbial community composition and activity.

600

601   References

602 Audic S, Claverie JM (1997). The significance of digital gene expression profiles. *Genome Res*
603 **7:** 986-995.

605 Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al* (2000). Bacterial
606 rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289:** 1902-1906.

608 Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T *et al* (2002).
609 Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415:** 630-633.

611 Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate - a practical and powerful
612 approach to multiple testing. *J R Stat Soc Series B Stat Methodo* **57:** 289-300.

614 Blanvillain S, Meyer D, Boulanger A, Lautier M, Guynet C, Denancé N *et al* (2007). Plant
615 carbohydrate scavenging through tonB-dependent receptors: a feature shared by phytopathogenic
616 and aquatic bacteria. *PLoS One* **2:** e224.

618 Connon SA, Giovannoni SJ (2002). High-throughput methods for culturing microorganisms in
619 very-low-nutrient media yield diverse new marine isolates. *Appl Environ Microbiol* **68:** 3878-
620 3885.

622 Cottrell MT, Kirchman DL (2000). Natural assemblages of marine proteobacteria and members
623 of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved
624 organic matter. *Appl Environ Microbiol* **66:** 1692-1697.

626 DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al* (2006). Community
627 genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496-503.

629 Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S *et al* (2006). Genome
630 analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features.
631 *Proc Natl Acad Sci U S A* **103:** 11647-11652.

633 Fasham MJR, Baliño BM, Bowles MC, Anderson R, Archer D, Bathmann U *et al* (2001). A new
634 vision of ocean biogeochemistry after a decade of the Joint Global Ocean Flux Study (JGOFS).
635 *Ambio***:** 4-31.

637 Feingersch R, Suzuki MT, Shmoish M, Sharon I, Sabehi G, Partensky F *et al* (2010). Microbial
638 community genomics in eastern Mediterranean Sea surface waters. *ISME J* **4:** 78-87.
639

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

Goffredi SK, Jones WJ, Scholin CA, Marin R, 3rd, Vrijenhoek RC (2006). Molecular detection of marine invertebrate larvae. *Mar Biotechnol (NY)* **8:** 149-160.

Gomez-Alvarez V, Teal TK, Schmidt TM (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3:** 1314-1317.

Gómez-Consarnau L, González JM, Coll-Lladó M, Gourdon P, Pascher T, Neutze R *et al* (2007). Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* **445:** 210-213.

Gómez-Pereira PR, Fuchs BM, Alonso C, Oliver MJ, van Beusekom JE, Amann R (2010). Distinct flavobacterial communities in contrasting water masses of the north Atlantic Ocean. *ISME J* **4:** 472-487.

González JM, Fernández-Gómez B, Fernàndez-Guerra A, Gómez-Consarnau L, Sánchez O, Coll-Lladó M *et al* (2008). Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (Flavobacteria). *Proc Natl Acad Sci U S A* **105:** 8724-8729.

Gosink JJ, Woese CR, Staley JT (1998). *Polaribacter* gen. nov., with three new species, *P. irgensii* sp. nov., *P. franzmannii* sp. nov. and *P. filamentus* sp. nov., gas vacuolate polar marine bacteria of the *Cytophaga-Flavobacterium-Bacteroides* group and reclassification of '*Flectobacillus glomeratus*' as *Polaribacter glomeratus* comb. nov. *Int J Syst Bacteriol* **48 Pt 1:** 223-235.

Greenfield DI, Marin R, Jensen S, Massion E, Roman B, Feldman J *et al* (2006). Application of environmental sample processor (ESP) methodology for quantifying *Pseudo-nitzschia australis* using ribosomal RNA-targeted probes in sandwich and fluorescent in situ hybridization formats. *Limnol Oceanog Methods* **4:** 426-435.

Greenfield DI, Marin III R, Doucette G, Mikulski C, Jones K, Jensen S *et al* (2008). Field applications of the second-generation Environmental Sample Processor (ESP) for remote detection of harmful algae: 2006-2007. *Limnol Oceanog Methods* **6:** 667-679.

Haywood AJ, Scholin CA, Marin R, Steidinger KA, Heil C, Ray J (2007). Molecular detection of the brevetoxin-producing dinoflagellate *Karenia brevis* and closely related species using rRNA-targeted probes and a semiautomated sandwich hybridization assay. *J Phycol* **43:** 1271-1286.

Hewson I, Poretsky RS, Tripp HJ, Montoya JP, Zehr JP (2010). Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. *Environ Microbiol* **12:** 1940-1956.

Huson DH, Auch AF, Qi J, Schuster SC (2007). MEGAN analysis of metagenomic data. *Genome Res* **17:** 377-386.

Jones WJ, Preston CM, Marin R, Scholin CA, Vrijenhoek RC (2008). A robotic molecular method for in situ detection of marine invertebrate larvae. *Mol Ecol Resour* **8:** 540-550.

Kanehisa M, Goto S (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27-30.

Karl DM, Dore JE, Lukas R, Michaels AF, Bates NR, Knap AH (2001). Building the long-term picture: The U.S. JGOFS time-series programs. *Oceanography* **14:** 6-17.

Karl DM (2002). Nutrient dynamics in the deep blue sea. *Trends Microbiol* **10:** 410-418.

Kirchman DL (2002). The ecology of *Cytophaga-Flavobacteria* in aquatic environments. *FEMS Microbiol Ecol* **39:** 91-100.

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.

Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009). Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol* **75:** 5345-5355.

Li W, Godzik A (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658-1659.

Man-Aharonovich D, Philosof A, Kirkup BC, Le Gall F, Yogev T, Berman-Frank I *et al* (2010). Diversity of active marine picoeukaryotes in the Eastern Mediterranean Sea unveiled using photosystem-II *psbA* transcripts. *ISME J* **4:** 1044-1052.

Martens EC, Koropatkin NM, Smith TJ, Gordon JI (2009). Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J Biol Chem* **284:** 24673-24677.

Martin-Cuadrado AB, Lopez-Garcia P, Alba JC, Moreira D, Monticelli L, Strittmatter A *et al* (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One* **2:** e914.

McCarren J, Becker JW, Repeta DJ, Shi Y, Young CR, Malmstrom RR *et al* (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proc Natl Acad Sci U S A* **107:** 16420-16427.

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

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35:** W182-185.

Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J* **4:** 673-685.

Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE *et al* (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME J* **4:** 784-798.

O'Mullan GD, Ward BB (2005). Relationship of temporal and spatial variabilities of ammonia-oxidizing bacteria to nitrification rates in Monterey Bay, California. *Appl Environ Microbiol* **71:** 697-705.

747 Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N *et al* (2007). The tiny
748 eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc*
749 *Natl Acad Sci U S A* **104:** 7705-7710.

751 Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA (2009). Comparative day/night
752 metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre.
753 *Environ Microbiol* **11:** 1358-1375.

755 Poretsky RS, Sun S, Mou X, Moran MA (2010). Transporter genes expressed by coastal
756 bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* **12:** 616-627.

758 Preston CM, Marin R, 3rd, Jensen SD, Feldman J, Birch JM, Massion EI *et al* (2009). Near real-
759 time, autonomous detection of marine bacterioplankton on a coastal mooring in Monterey Bay,
760 California, using rRNA-targeted DNA probes. *Environ Microbiol* **11:** 1168-1180.

762 Rich V, Pham V, Eppley J, Shi Y, Delong EF (2010). Time-series analyses of Monterey Bay
763 coastal microbial picoplankton using a 'genome proxy' microarray. *Environ Microbiol*.

765 Rich VI, Konstantinidis K, DeLong EF (2008). Design and testing of 'genome-proxy'
766 microarrays to profile marine microbial communities. *Environ Microbiol* **10:** 506-521.

768 Riedel T, Tomasch J, Buchholz I, Jacobs J, Kollenberg M, Gerdts G *et al* (2010). Constitutive
769 expression of the proteorhodopsin gene by a flavobacterium strain representative of the
770 proteorhodopsin-producing microbial community in the North Sea. *Appl Environ Microbiol* **76:**
771 3187-3197.

773 Roman B, Scholin C, Jensen S, Massion E, Marin III R, Preston C *et al* (2007). Controlling a
774 robotic marine environmental sampler with the Ruby scripting language. *JALA* **12:** 56-61.

776 Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al* (2007). The
777 Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical
778 Pacific. *PLoS Biol* **5:** e77.

780 Ryan JP, Johnson SB, Sherman A, Rajan K, Py F, Thomas H *et al* (2010a). Mobile autonomous
781 process sampling within coastal ocean observing systems. *Limnol Oceanog Methods* **8:** 394-402.

783 Ryan JP, McManus MA, Sullivan JM (2010b). Interacting physical, chemical and biological
784 forcing of phytoplankton thin-layer variability in Monterey Bay, California. *Continental Shelf*
785 *Res* **30:** 7-16.

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

790 Scholin C, Doucette G, Jensen S, Roman B, Pargett D, Marin R *et al* (2009). Remote detection
791 of marine microbes, small invertebrates, harmful algae, and biotoxins using the Environmental
792 Sample Processor (ESP). *Oceanography* **22:** 158-167.

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

797 Shi Y, Tyson GW, Eppley JM, Delong EF (2010). Integrated metatranscriptomic and
798 metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J* **Epub Dec**
799 **9**.

Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF *et al* (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J* **3:** 93-105.

Stewart FJ, Ottesen EA, DeLong EF (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J* **4:** 896-907.

Stingl U, Desiderio RA, Cho JC, Vergin KL, Giovannoni SJ (2007). The SAR92 clade: an abundant coastal clade of culturable marine bacteria possessing proteorhodopsin. *Appl Environ Microbiol* **73:** 2290-2296.

Suzuki MT, Béjà O, Taylor LT, DeLong EF (2001). Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ Microbiol* **3:** 323-331.

Suzuki MT, Preston CM, Béjà O, de la Torre JR, Steward GF, DeLong EF (2004). Phylogenetic screening of ribosomal RNA gene-containing clones in Bacterial Artificial Chromosome (BAC) libraries from different depths in Monterey Bay. *Microb Ecol* **48:** 473-488.

Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3:** e2527.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Vila-Costa M, Rinta-Kanto JM, Sun S, Sharma S, Poretsky R, Moran MA (2010). Transcriptomic analysis of a marine bacterial community enriched with dimethylsulfoniopropionate. *ISME J* **4:** 1410-1420.

Whittaker RH (1952). A study of summer foliage insect communities in the great smoky mountains. *Ecological Monographs* **22:** 1-44.

Worden AZ, Lee JH, Mock T, Rouzé P, Simmons MP, Aerts AL *et al* (2009). Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324:** 268-272.

Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H *et al* (2009). Assembling the marine metagenome, one cell at a time. *Plos One* **4**.

Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M *et al* (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468:** 60-66.

845 **Table 1. Samples and RNA extraction efficiencies**

Sample	Vol (ml)[1]	Yield (µg)
Monterey wharf – Time 0	500	3.2
Monterey wharf – 1 week	500	5.0
Monterey wharf – 2 weeks	500	4.7
Monterey wharf – 3 weeks	500	5.1
Monterey wharf – 4 weeks	500	4.8
Station M0 – 4/7/09	1000	1.4
Station M0 – 4/9/09	1000	2.5
Station M0 – 4/12/09	1000	1.9
Station M0 – 4/16/09	1000	2.1
Station M0 – 4/20/09	1000	2.2
Station M0 – 4/25/09	1000	2.4
Santa Cruz wharf -- Vacuum	300	4.0
Santa Cruz wharf -- ESP	1000	11.8
Station M1 – Peristaltic Pump	500	0.7
Station M1 – ESP	1000	0.9
Station M0 – 6/2/09 5am	1000	0.7
Station M0 – 6/2/09 10am	1000	2.5
Station M0 – 6/2/09 6pm	1000	0.6
Station M0 – 6/2/09 10pm	1000	1.2

846 [1]Total volume of seawater filtered.

847

847 **Table 2. Read numbers and statistics**

Sample	Reads[1]	% rRNA[2]	non-rRNA reads		
			Replicates[3]	nr Hits[4]	KEGG Hits[4]
Santa Cruz wharf: Vacuum	179 643	80%	20%	17 988	15 869
Santa Cruz wharf: ESP	160 364	82%	18%	14 702	12 850
Station M1: Peristaltic Pump	118595	41%	26%	20981	18677
Station M1: ESP	203574	42%	11%	41295	37086
5am rRNA-subtracted	248 016	33%	4%	82 387	69 157
5am Unsubtracted	298 380	91%	7%	11 802	10 089
10am rRNA-subtracted	102 024	40%	17%	25 197	22 250
10am Unsubtracted	149 186	82%	27%	9 979	8 612
6pm rRNA-subtracted	238 635	38%	17%	54 040	46 253
6pm Unsubtracted	232 248	83%	12%	15 694	13 156
10pm rRNA-subtracted	235 339	35%	13%	52 069	43 956
10pm Unsubtracted	202 650	82%	24%	10 701	8 890
DNA[5]	1 535 834	0.4%	0.97%	1 035 676	956 510

848 [1]Total number of sequence reads passing quality filters
849 [2]Percentage of total pyrosequencing reads with significant (bit score > 50) BLASTN hits to
850   prokaryotic and eukaryotic rRNA (16S, 18S, 23S, 28S, 5S)
851 [3]Percentage on non-rRNA reads identified as artificial replicates (99% identity, 1bp length
852   difference) and removed
853 [4]Non-replicate, non-rRNA reads with significant (bit score > 50) BLASTX hits to proteins in the
854   NCBI non-redundant (nr) or KEGG Genes databases.
855 [5]Metagenomic dataset, sequenced using GS FLX Titanium chemistry rather than GS FLX.
856

856 **Table 3.** NCBI-nr reference genes with significantly different abundances in metatranscriptomes
857   generated from vacuum filtered vs. ESP-filtered and preserved replicate samples[1]

NCBI-nr Reference Gene	Vacuum[2]	ESP[2]	*p*-value
AAM48736 antenna pigment protein, alpha chain [uncultured marine proteobacterium]	N.D.	0.22%	6.1×10^{-08}
ZP_01224258 hypothetical protein [*gamma proteobacterium* HTCC2207]	0.25%	0.67%	4.1×10^{-05}
ZP_01447883 branched-chain amino acid ABC transporter, periplasmic substrate-binding protein [*Rhodobacterales* sp. HTCC2255]	0.54%	0.17%	5.4×10^{-05}
ZP_03559919 50S ribosomal subunit protein L3 [*Glaciecola* sp. HTCC2999]	0.12%	N.D.	8.5×10^{-03}
ZP_03559922 50S ribosomal protein L2 [*Glaciecola* sp. HTCC2999]	0.10%	N.D.	0.041
ZP_01447418 glutamate synthase large subunit [*Rhodobacterales* sp. HTCC2255]	0.14%	0.01%	0.043

858 [1]Reference genes with FDR-corrected *p*-value > 0.05.
859 [2]The percent of unique, non-rRNA reads with significant NCBI-nr database hits mapping to each
860   reference gene in vacuum-filtered and the ESP-processed and preserved samples.
861

Table 4. Comparison of metatranscriptomic profiles from manually-processed and ESP-collected and preserved samples.

DS compared[1]		NCBI-nr taxa[2]			WI[3]	NCBI-nr references[4]			sig. diff. refs[5]	% hits in sig. diff. refs[6]	
DS 1	DS 2	DS1	DS2	Both		DS1	DS2	Total		DS1	DS2
SC Vac	SC ESP	785	666	448	0.89	10723	9386	17284	6	1.2%	1.1%
M1 PP	M1 ESP	1280	1613	1028	0.82	13754	25211	35036	28	1.6%	1.3%

[1]Datasets used in pariwise comparisons (SC Vac and ESP: Santa Cruz wharf, vacuum-filtered and flash frozen sample and ESP-processed and preserved sample, stored on board ESP for a 30-day deployment in Monterey Bay; M1 PP and ESP: Monterey Bay water collected by CTD cast, filtered immediately by peristaltic pump or processed and preserved by ESP. Additional pairwise comparisons are listed in Table S2.

[2]Number of taxa (NCBI-nr taxonomy ID's) with one or more uniquely assigned sequences (reads with exactly one top-scoring database match).

[3]Whittaker's index of association (Whittaker 1952) for NCBI-nr taxon counts.

[4]Number of NCBI-nr reference genes with one or more mapped reads

[5]NCBI-nr reference genes with significantly different abundances in the two datasets.

[6]Percent of sequences with NCBI-nr hits that map to reference genes with significantly different abundances.

875 **Table 5. Percent of transcripts assigned to different KEGG functional categories.**

Category	Bulk Community[1]				HTCC2255[2]				*Polaribacter*[3]			
	5am	10am	6pm	10pm	5am	10am	6pm	10pm	5am	10am	6pm	10pm
Unassigned	51	49	48	54	31	30	33	40	56	54	51	60
Metabolism	30	35	32	29	35	41	37	31	27	32	29	27
Genetic information processing	18	13	17	14	23	20	20	15	17	15	21	13
Environmental information processing	4.4	4.5	4.9	4.7	15	14	16	19	3.0	2.3	3.2	2.6
Cellular processes	2.3	2.6	2.6	3.3	1.9	2.4	2.4	2.5	2.0	2.3	1.9	2.4

876 [1]Percent of sequences with significant hits in the KEGG genes database.
877 [2]Percent of total sequences assigned to *Rhodobacterales* sp. HTCC2255 or to either of the
878 sequenced *Polaribacter* genomes.
879

879   **Figure Legends**

880   **Figure 1.   Size fractionation of total RNA extracted from ESP-collected and preserved**

881   **samples.** All samples were diluted to approximately equal concentrations prior to analysis to

882   facilitate comparison of RNA quality. **A**. Total RNA extracted from replicate surface water

883   samples collected and preserved using the ESP and stored at room temperature under conditions

884   that simulate a deployed instrument (high humidity, N2 atmosphere). **B**. Total RNA extracted

885   from samples collected throughout a single deployment of the ESP at Monterey Bay station M0.

886

887   **Figure 2. Metatranscriptomic analysis of ESP-collected and preserved samples.**   The

888   abundance of NCBI-nr reference genes is shown for a sample collected by ESP and retained on

889   the instrument for a 29-day deployment in Monterey Bay and a replicate sample collected by

890   vacuum filtration and flash frozen vaccum-filtered. For visualization purposes, reference genes

891   with 0 assigned sequence reads were counted as 0.5. Reference genes with significantly different

892   abundances in the two data sets (FDR-corrected p-value <0.05) are shown in red. Accession

893   numbers and p-values of significantly different reference genes are listed in Table 3.

894

895   **Figure 3.   Relative abundance of major taxonomic groups in metatranscriptomic and**

896   **metagenomic samples.** Sequences were assigned to the NCBI taxonomy using the MEGAN

897   program (Huson et al 2007), (bitscore > 50, top 10% of hits). Taxonomic analyses of small

898   (SSU) and large (LSU) subunit rRNA sequences are based on unsubtracted RNA samples only.

899   Coding sequence (CDS) taxonomy generated from the combined non-replicate, nonrRNA

900   fraction of both unsubtracted and subtracted RNA. Groups representing > 1% of assignable

901   sequences in one or more samples are shown, those representing < 1% of sequences in all

902   samples are included in the 'other' category, and those assigned at lower taxonomic levels are

903   not shown.

904

905   **Figure 4.   Percent identity histograms for sequences assigned to *Rhodobacterales* sp.**

906   ***HTCC2255 or *Polaribacter*.** A global percent identity (percent amino acid similarity * fraction

907   of read covered) was calculated for all significant (bitscore >50) BLASTX hits in the NCBI-nr

908   database for metatranscriptomic libraries from Station M0, and the number of sequences aligning

909   at each percent identity determined. **A.** Sequences for which *Rhodobacterales* sp. HTCC2255 is

910   among the top-scoring hits in the NCBI-nr database. To show the specificity with which these

911   sequences were mapped, additional database hits passing the bitscore threshold were examined,

912   and percent identity histograms generated for the two most frequently-identified taxa,

913   *Rhodobacterales* spp. HTCC2150 and HTCC2083. **B.** Sequences with at least one top hit to

914   either of the two *Polaribacter* reference genomes. Significant alignments of *Polaribacter-*

915   assigned reads to *Flavobacteriales* sp. ALC-1 and *Kordia algicida* OT-1 are also shown.

916
917   **Figure 5. Relative abundances of KEGG pathways in metatranscriptomic datasets.** The 10

918   most abundant KEGG pathways in the bulk community, plus pathways within the 10 most

919   abundant pathways for either *Rhodobacterales* sp. HTCC2255 or *Polaribacter* are shown in

920   order of descending abundance in the total community. Percent of sequences with significant

921   hits to the KEGG database (community) or percent of total sequences assigned to specific taxa

922   shown.   Error bars represent 95% confidence limits.   Photosynthesis signal in the (non-

923   photosynthetic) HTCC2255 and *Polaribacter* bins is due to the assignment of F0F1 ATP
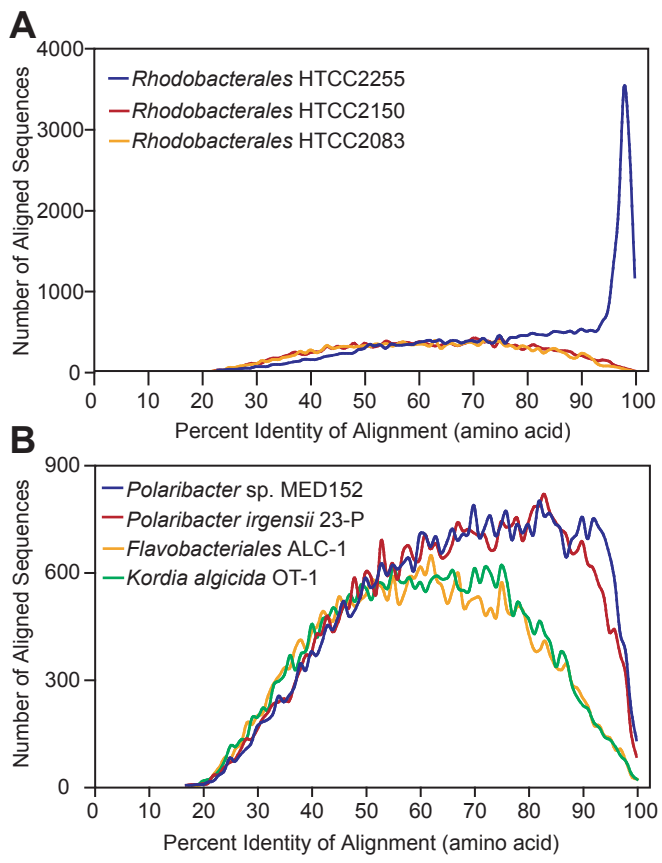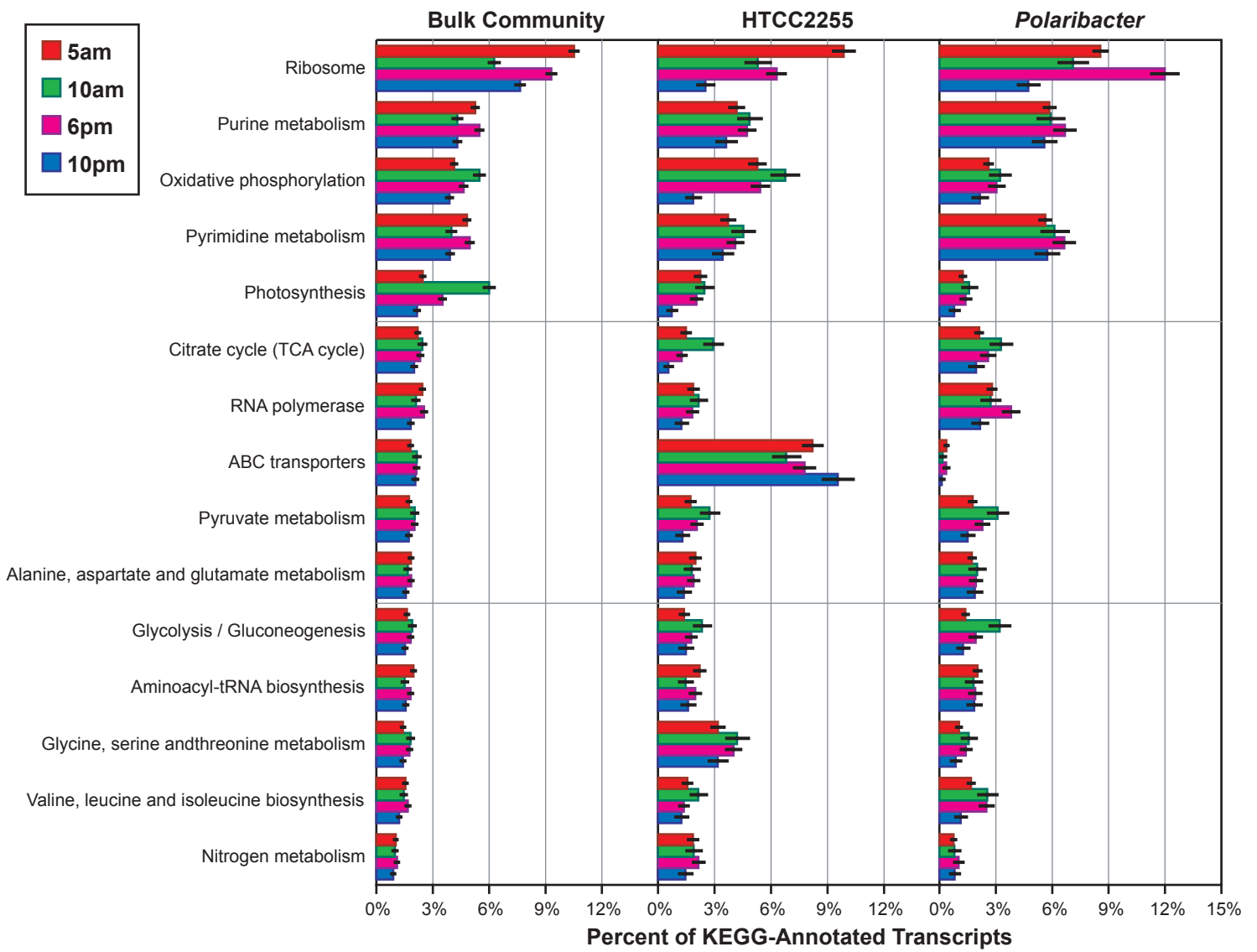
924   Synthase genes to this category.
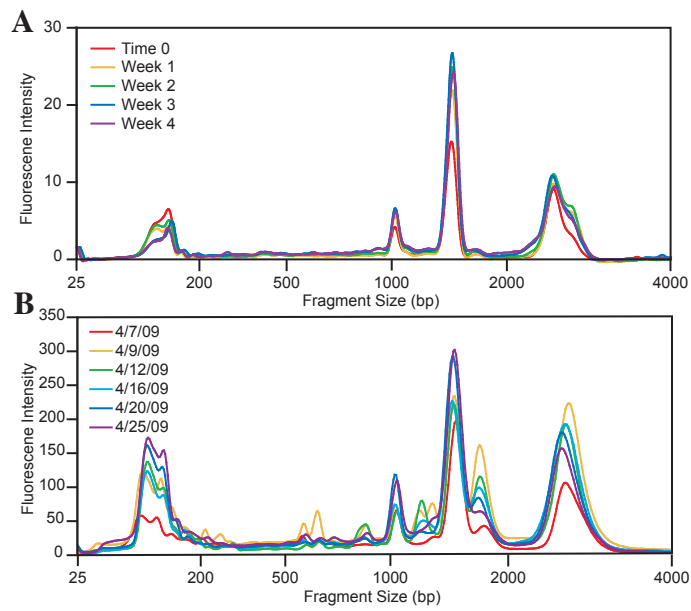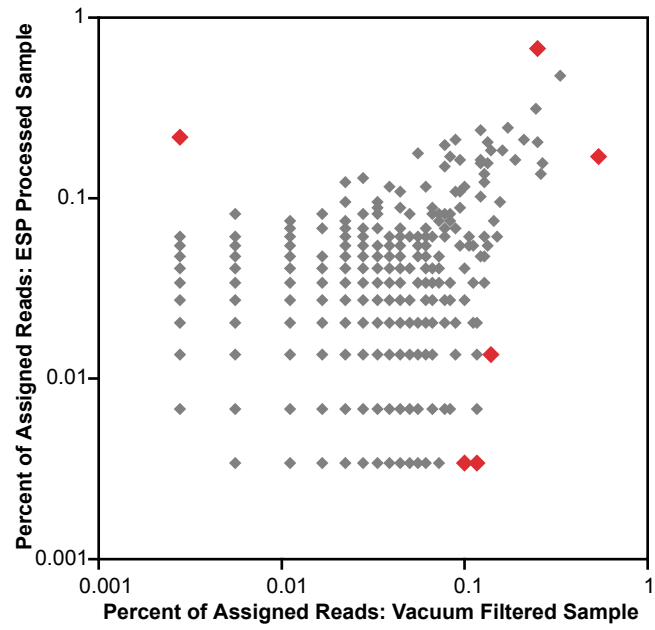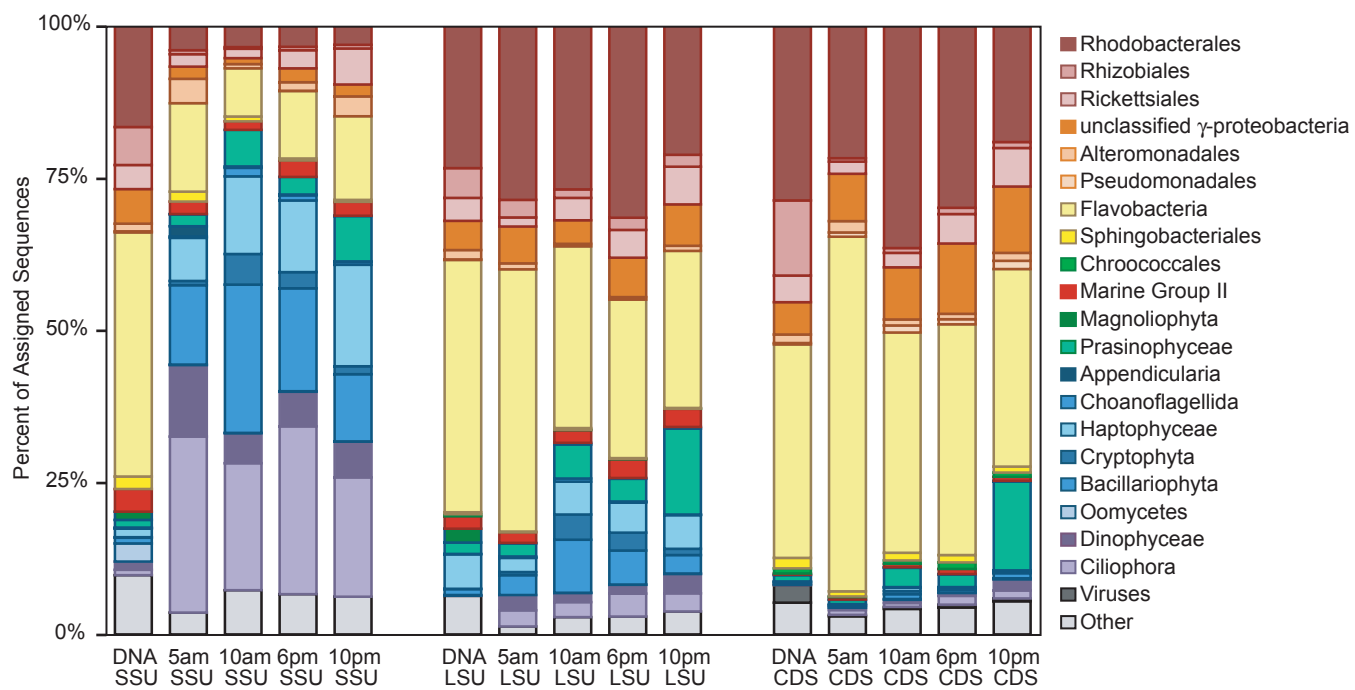
Figure 1.

Figure 2.

Figure 3.

Figure 4.

Figure 5.