

## Creation of a pilot metatranscriptome library from eukaryotic plankton of a eutrophic bay (Tampa Bay, Florida)

David E. John\*, Brian L. Zielinski, and John H. Paul

University of South Florida, College of Marine Science, St. Petersburg, FL, USA

### Abstract

Analysis of the suite of genes expressed by natural populations of phytoplankton can potentially elucidate valuable information about the types, cellular activity, and biogeochemical impacts of organisms present in the marine environment. Here we describe the construction of a pilot metatranscriptome library created from eukaryotic planktonic organisms in Tampa Bay, FL, USA. RNA from cells greater than 2  $\mu\text{m}$  was extracted and purified, and poly(A) tailed mRNA was concentrated and amplified using linear amplification chemistry. Amplified RNA was converted to double-stranded cDNA using reverse transcriptase and DNA polymerase I (Klenow fragment) and cloned, and 232 clones were sequenced. Sequences with significant GenBank BLAST homology revealed genes related to photosynthesis and nutrient acquisition, along with general cell functions. In total, 27% of sequenced transcript clones contained significant homology to GenBank sequences, 2% of the total were putatively derived from ribosomal RNA, and 1% were most similar to sequences originating from prokaryotes. About 70% of the identified transcripts were putatively derived from eukaryotic phytoplankton, including diatoms, chlorophytes, and dinoflagellates. Although small in scale, this study provides the basis for future efforts to characterize the metatranscriptome of marine phytoplankton populations.

### Introduction

The biogeochemical cycling of globally important elements is intrinsically controlled by the expression of genes encoding the reactions catalyzed by the organisms (principally microbes) participating in these cycles. The sum of all gene transcripts in an organism is termed the transcriptome. The application of transcriptome analysis to natural environments (metatranscriptomics) affords an understanding of the expression of genes of the ambient microbial community in situ. The first glimpse at an environmental transcriptome was the work of Poretsky et al. (2005), who built primarily prokaryotic mRNA libraries derived from two aquatic sites: a tidal salt marsh creek (southeastern US) and a hypersaline soda lake (Mono Lake, CA). Four hundred clones were analyzed from this library, and among them were found transcripts encoding sulfur oxidation, acquisition of Cl compounds, and polyamine degradation. Although this was only a glimpse of the gene expression profile of these environments, it provided the proof of concept necessary for metatranscriptomics.

Other prokaryotic transcriptome research has focused on open-ocean microbial populations and the use of pyrosequencing (Frias-Lopez et al. 2008). Pyrosequencing enables

vastly larger libraries of relatively short reads; Frias-Lopez et al. evaluated 128,324 cDNA reads with an average length of 114 nucleotides (nt). Subsurface water (240 L) from the oligotrophic Pacific Ocean near Hawaii was processed to extract total RNA. To examine prokaryotic transcripts, poly(A) tails were linked to the community RNA extract and linear amplification was performed using polyadenylation-dependent amplification methods with a commercial kit (Ambion MessageAmp). Amplified RNA was then reverse-transcribed to cDNA for pyrosequencing. Whereas the *Escherichia coli* poly(A) polymerase used enables preferential adenylation of bacterial mRNA over ribosomal RNA (rRNA), a majority (53%) of the sequences obtained corresponded to rRNA. Of the remainder, 12% (7275 sequences) matched known protein sequences from the NCBI nonredundant protein database (nr). The class of protein-encoding genes most frequently identified were those corresponding to microbial phototrophy, such as carbon fixation (ribulose-1,5-bisphosphate carboxylase/oxygenase [RuBisCO] and glutamine synthase), light-harvesting proteins, photosynthesis reaction centers, and bacterial proteorhodopsin. *Prochlorococcus*-derived transcripts appeared to be highly represented.

A metagenomic cDNA library targeting eukaryotic transcripts has also been reported (Grant et al. 2006). Three samples were analyzed, two from geothermal spring algal mats and one from an activated sewage sludge sample. The algal

\*Corresponding author: E-mail: djohn@marine.usf.edu

mat samples were not enriched for poly(A) RNA, whereas the sample from sewage sludge was split with a portion of the RNA enriched for poly(A)-tailed transcripts using oligo(dT) magnetic bead capture techniques. Total or poly(A)-enriched RNA was reverse transcribed and subsequently amplified using polymerase chain reaction (PCR), before ligation into lambda vectors. Fifty-three clones from the algal mat samples, 24 from sludge total RNA, and 23 from poly(A)-selected sludge RNA were sequenced and analyzed. Sequences from algal mat samples revealed six possible proteins, whereas 60% of the sequences were putatively derived from rRNA. From the activated sludge samples, the total RNA preparation resulted in five rRNA sequences (20%), whereas only one of 23 clones from poly(A)-enriched RNA represented rRNA. Sequences with homology to known proteins indicated a diverse assemblage of genes with primarily eukaryotic origins. A large fraction of the algal mat sequences were derived from prokaryotic organisms, which the authors speculate could have been due to mispriming of the reverse transcription or polyadenylation of prokaryotic RNA, along with a very low population of eukaryotic microbes. Although the organisms were not from a marine source, this article does introduce the use of metatranscriptomics for eukaryotic organisms in environmental samples.

We are principally interested in understanding the suite of genes expressed by eukaryotic phytoplankton. Tremendous advances have been made in understanding the transcriptomes of marine diatoms in culture (Mock et al. 2006, 2008; Scala and Bowler 2001). These studies have provided insight on environmental adaptation of these organisms and specifically on the functioning of silica metabolism and deposition, a prominent oceanic biogeochemical process. Additional efforts at describing gene expression in cultured phytoplankton have been reported. Dyhrman et al. (2006) have employed serial analysis of gene expression (SAGE) with *Emiliana huxleyi* to identify differentially expressed transcripts in cultures under nitrogen or phosphorus starvation. Likewise, Erdner and Anderson (2006) have evaluated the expression profile of the toxic dinoflagellate *Alexandrium fundyense* under nitrate- and phosphate-limited conditions using massively parallel signature sequencing (MPSS). Both these techniques claim to provide quantitative gene expression profiling while employing extremely short (21-nt or 17-bp, respectively) sequence tags. Further investigation of differentially expressed tags can be performed using RT-PCR methods.

Our objective was to establish a method for capturing the metatranscriptome of the eukaryotic plankton population in aquatic/marine samples. The basis for such methodology is to eventually enable ecological descriptions based on the suite of genes expressed in the organisms active in a given marine region (e.g., oceanic or coastal river plumes). Although Grant et al. (2006) present a method for obtaining elements of a eukaryotic metatranscriptome from algal mats and activated sewage sludge, their interest was chiefly in collecting full-length transcripts for potential biotechnology applications,

thus their use of long-distance PCR for amplification of the transcript pool. However, the linear RNA amplification technique first established by Eberwine et al. (1992, Vangelder et al. 1990) is the more common and established method for unbiased amplification of cDNA signal for expression profiling, generally by microarray hybridization (Feldman et al. 2002, Kacharina et al. 1999, Li et al. 2004, Pabon et al. 2001, Polacek et al. 2003) but also in work on metatranscriptomes of prokaryotic populations, including the research discussed above (Frias-Lopez et al. 2008, Moreno-Paz and Parro 2006). Linear amplification of RNA enabled by T7 RNA polymerase has been shown to introduce minimal, if any, bias for whole transcriptome profiling while allowing the detection of lower-abundance transcripts (Feldman et al. 2002, Li et al. 2004, Polacek et al. 2003). Conversely, analysis by Poretsky et al. on the use of PCR in constructing their metatranscriptomic libraries indicated instances of bias in terms of selective amplification and unequal capture of transcripts. Also, the exponential amplification of PCR can result in lower abundance templates in mixed pools being underrepresented in the final amplicon pool (the Monte-Carlo effect, Karrer et al. 1995) and is not used for expression profiling. Furthermore, considering the differences in sample characteristics between algal mats, sewage sludge, and the marine plankton community (biomass concentrations, nature of the organisms, organic matter concentrations), validation of techniques for sample collection and eukaryotic metatranscriptome creation specifically from filtered seawater organisms is desirable.

For this research, we sought to apply techniques for eukaryotic gene expression profiling to environmental plankton samples. We desired to test the effectiveness of mRNA enrichment with poly(A) capture techniques and poly(A) tail-based linear RNA amplification for obtaining eukaryotic-specific cDNA clones in a library. Our two primary questions were (1) whether we could adequately sample for eukaryotic organisms over the more-abundant prokaryotic fraction and (2) what proportion of the sequences identified would be protein-coding versus rRNA, to gauge the potential efficiency of future, larger-scale efforts for marine plankton community transcriptome profiling and the ecological information it could provide. Here we provide a method for the production of eukaryotic transcriptomes of planktonic microbial populations. We present a pilot study of one such eukaryotic transcriptome of a surface water sample from a eutrophic harbor, and demonstrate that 70% of the identifiable transcripts were putatively of eukaryotic phytoplankton origin.

## Materials and procedures

**RNA extraction and purification**—Water from Tampa Bay, FL, USA, was collected from the surface adjacent to our facility (latitude 27.7616, longitude -82.6326) by a dip bucket, strained through a 50- $\mu$ m mesh filter, and immediately brought into the lab for processing. The sample was collected 1 day after a bloom of the nontoxic dinoflagellate, *Peridinium*

(syn: *Protoperdinium*) *quinquecorne*, and thus represented a declining bloom sample. Sampling occurred in mid-August at 1000 local time. Cells were collected on 47-mm, 2.0- $\mu$ m-pore polycarbonate filters (#11111; Whatman plc) by passing approximately 180 mL bay water through each of seven filters, for a total of 1.25 L filtered. Water was filtered by gentle vacuum pressure (200 mmHg); the multiple filters were used to reduce the length of time any given cell collection resided on a filter, thereby minimizing premature lysis of fragile cells or changes to expression profile. The filtration was completed in approximately 15 min. Filters were split between two 50-mL polypropylene centrifuge tubes (three or four in each) and vortexed at maximum speed for 1 min on a Vortex Genie 2 benchtop vortexer (Fisher Scientific) with 40 mL additional bay water to liberate cells from filters. After removing filters, cell suspensions were centrifuged at 7000*g* for 10 min at 15°C in a benchtop centrifuge.

The supernatant was carefully removed, leaving 2 mL overlying the cell pellet. Microscopic examination revealed that cells were not visible in the supernatant. The cell pellet was resuspended in 6 mL Qiagen RLT buffer with 10  $\mu$ L/mL  $\beta$ -mercaptoethanol (Qiagen RNeasy kit and Sigma-Aldrich, respectively) and transferred to a 15-mL conical centrifuge tube with 2 g 200- $\mu$ m low-protein-binding zirconium grinding beads (OPS Diagnostics). This was vortexed at maximum speed for 3 min to lyse cells and liberate RNA. Microscopic examination at this point revealed that some cells remained apparently intact after vortexing. From this point, the sample was split to ensure we could both recover RNA from recalcitrant cells and prevent undo shearing due to bead beating of RNA from more-fragile cells that had already lysed, our aim being to reduce potential bias toward either group. Thus, the suspension was transferred to another 15-mL centrifuge tube to separate from the grinding beads, and centrifuged at 10,000*g* for 5 min at 4°C to pellet unlysed cells. The supernatant (7 mL) was removed and purified with RNeasy midi columns; lysate was mixed with 4.9 mL 100% ethanol and split between two columns (6 mL each), requiring two rounds of centrifugation following kit protocols to pass the volume through the columns for RNA binding.

Concurrently, the pelleted cells were resuspended in 1.5 mL RLT buffer plus  $\beta$ -mercaptoethanol by vortexing and transferred to a 2-mL bead-beating tube with about 0.2 mL zirconium beads. This was homogenized on a Biospec bead beater on “homogenize” (maximum) setting for 2 min, placed on ice for 5 min, and centrifuged for 2 min at 16,100*g* at room temperature. The 1.5-mL supernatant was mixed with 1.1 mL 100% ethanol and passed through an RNeasy midi column according to kit protocols. The two “fractions” (fragile cells lysed by vortexing with lysis buffer and grinding beads and recalcitrant cells requiring bead beating) constituted sufficient volumes such that processing in multiple RNeasy columns was necessary. All columns were washed with 4 mL buffer RW1 (once) and 2.5 mL buffer RPE (twice); RNA was eluted in 150

$\mu$ L RNase-free water for each of the two vortexed-only sample columns and 250  $\mu$ L for the single bead-beaten sample column. Eluate volumes were passed through each respective column twice. Eluates from the two vortexed-only columns were combined (300  $\mu$ L), and the two RNA preparations were measured for quantity and purity on a Nanodrop ND-1000 spectrophotometer (Thermo Scientific). Both fractions contained similar concentrations of RNA (45–55 ng/ $\mu$ L) and equal 260/280 nm absorbance ratios of 2.1. Because both preparations contained a substantial quantity of RNA, purified RNA from the two fractions were at this point combined (total 550  $\mu$ L) to ensure coverage of the entire plankton community in the sample with a single library. In addition, our total RNA recovered was low for the recommendations of RNA quantity to be used in poly(A) enrichment (below).

Enrichment for eukaryotic mRNA was performed using Oligotex poly(A) capture techniques (Qiagen). Kit protocols for the Oligotex purification methods were followed for 500- $\mu$ L volumes, except that incubation at room temperature to allow binding of poly(A) moieties to resin beads proceeded for 15 min rather than the suggested 10 min, with gentle agitation every 5 min to enhance binding, as our total RNA concentration was below the minimum advised for 500  $\mu$ L starting volume (recommended quantity is a minimum of 250  $\mu$ g, we had 27  $\mu$ g; RNA quantities are a persistent challenge for environmental metatranscriptomics). Also, elution was performed with 30  $\mu$ L buffer OEB at 70°C twice, using the same volume passed through the column with reheating in a heat block to 70°C for 1 min between elution centrifugations to maximize recovered RNA concentration. The eluted poly(A) mRNA was again analyzed in the Nanodrop spectrophotometer, revealing that 720 ng poly(A)-enriched RNA had been recovered.

*Amplification of community mRNA*—Immediately after purifying poly(A)-tailed mRNA, the RNA was amplified using Ambion’s MessageAmp II aRNA amplification kit. Following kit protocols, the RNA was processed to cDNA (first- and second-strand reactions), with a T7 RNA polymerase binding site added via the T7 oligo(dT) primer, and frozen overnight at –20°C. The following day, the cDNA was purified using kit components and protocols, quantity of DNA was analyzed on the Nanodrop spectrophotometer, and the in-vitro transcription (40  $\mu$ L reaction) to create amplified antisense RNA was performed overnight (14 h at 37°C), again according to kit protocols. RNA was then purified and eluted in 2 $\times$  75  $\mu$ L (150  $\mu$ L total) RNase-free water, heated to 55°C. Yield was again measured using the Nanodrop, reflecting a concentration of 1.15  $\mu$ g/ $\mu$ L or 173  $\mu$ g total.

*Reverse-transcription of amplified RNA and cloning*—Reverse-transcription was performed using a Superscript III reverse transcription (RT) kit (Invitrogen) with 25  $\mu$ g amplified antisense RNA (aRNA) at a concentration of 1.15  $\mu$ g/ $\mu$ L. RNA was first heat-denatured in a mixture of 25  $\mu$ L RNA, 8  $\mu$ L random hexamers (400 ng or 16 ng/ $\mu$ g RNA), 5  $\mu$ L dNTPs, and 12  $\mu$ L

H<sub>2</sub>O and incubated at 65°C for 5 min, then immediately placed in ice for 5 min. A separate RT mix was made, scaled up 5× to accommodate the 25 µg RNA: 10 µL 10× reaction buffer, 20 µL MgCl<sub>2</sub>, 10 µL 0.1 M dithiothreitol, 5 µL RNase-Out (RNase inhibitor), and 5 µL Superscript III reverse transcriptase (200 U/µL). The two mixtures were combined and incubated for 10 min at room temperature to allow hexamers to bind, then transferred to 50°C for 2 h. The reaction was heated to 85°C for 5 min to inactivate reverse transcriptase, cooled to room temperature, and spun to collect volume to bottom; 5 µL RNase H was added, and the reaction was incubated at 37°C for 45 min then frozen at -20°C.

The second-strand synthesis was performed in the same reaction mixture as first-strand synthesis; reaction components were 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 5 mM MgCl<sub>2</sub>, 10 mM DTT, and 333 µM each dNTP. To 27 µL of this first-strand reaction (the equivalent of 6.75 µg starting aRNA template) was added 5 units DNA polymerase I (Klenow fragment; Promega), 1 µL additional dNTP mix (10 mM each, from Invitrogen Superscript III reaction kit), 1.2 µL 10 mg/mL bovine serum albumin (for a reaction concentration of 400 µg/mL; Promega), and 0.5 µL random hexamers (for a final reaction concentration of 8 ng/µL). This second-strand reaction was incubated for 90 min at 25°C, then heat-inactivated at 75°C for 15 min. The DNA was purified with DNA Clean and Concentrate (Zymo) with the addition of 3 volumes binding buffer (90 µL) to facilitate column binding. DNA was eluted in 50 µL nuclease-free water heated to 50°C. To ensure that only double-stranded DNA was quantified, DNA was analyzed using Hoechst 33258 (Paul and Myers 1982), which indicated that 1.5 µg was recovered (30 ng/µL). The eluate was also evaluated in the Nanodrop ND1000 and by Ribogreen RNA assay (Invitrogen), which revealed a considerable quantity remaining of either RNA or single-stranded DNA (measured at 115 ng/µL by Ribogreen; Nanodrop measured 112 ng/µL based on the A260 absorbance correlation for double-stranded DNA).

Because RNA has the potential to interfere with DNA phosphorylation required for blunt-end cloning, an RNase digestion was incorporated into the procedure. Phosphorylation was performed using T4 polynucleotide kinase (Promega), which is accompanied by a reaction buffer concentrate. An initial RNase digestion was performed with 4 µL kinase 10× buffer, 2 µL 0.1 mM ATP (Promega), 2 µL RNase One (Promega), 10 µL eluted DNA (300 ng double-stranded cDNA), and 22 µL nuclease-free water for a final reaction volume of 40 µL. The digestion was incubated for 1 h at 37°C, and directly to it was added 1 µL polynucleotide kinase, followed by another 1-h incubation at 37°C. The reaction was halted by the addition of 2 µL 0.5 M EDTA. DNA was purified on with the Zymo Clean and Concentrate columns, with three wash steps rather than the suggested two, and eluted in 25 µL nuclease-free water.

Cloning was performed with the pSmart-CloneSmart HCamp cloning vector and kit (Lucigen) as follows. Phosphorylated cDNA (6.5 µL; 78 ng DNA equivalent) was mixed with

2.5 µL 4× pSmart vector mix and 1 µL ligase (both from kit) and incubated at 25°C for 2 h. Cloned vector was transformed via electroporation into Lucigen "E Cloni" 10G Supreme Duo cells with a Bio-Rad Micropulser on Ec1 setting as directed in cloning kit protocols; 1 µL ligation mixture was used to transform cells. Recovered cells were spread on YT plates with 100 µg/mL ampicillin and grown overnight at 37°C.

Individual colonies were picked and screened by PCR for adequately sized inserts. Screening PCR was performed with Promega *Taq* polymerase using the pSmart sequencing primers SR2 and SL1. PCR amplicons indicating an insert size greater than 200 base pairs (bp) were retained and sent for sequencing.

Two hundred thirty-two cDNA clones from the metatranscriptome were sequenced (average query size 350 bp). BLAST searching (NCBI GenBank, using *blastn* and *blastx*) revealed that 64 of these showed homology to previously identified sequences with BLAST scores  $\geq 40$  bits, as used in Frias-Lopez et al. (2008) for establishing significance of hits. These 64 sequences were submitted to GenBank; 59 that were homologous to protein coding sequences (identified by translated BLAST search) were submitted to the EST database with accession numbers GH571923 through GH571981, and five that were homologous to ribosomal RNA sequences were submitted to the nonredundant nucleotide database with accession numbers FJ643611 through FJ643615.

### Assessment

*RNA purification and amplification*—From the filtered seawater volume of 1.25 L, a total of 27 µg RNA was recovered as measured by spectrophotometry. Because DNase digestion was not included in the RNeasy purification protocol, a portion of this signal may have been DNA, although the portion is likely to have been very small as the RNeasy methods strongly favor recovery of RNA. The RNeasy literature states that potential DNA contamination is generally only a concern for sensitive techniques such as RT-PCR, wherein a proportionally small fraction of DNA could be amplified by specific primers, thereby skewing results. Two treatments were used to lyse cells and liberate RNA: vortexing the filters with lysis buffer, which is likely to recover RNA from delicate cells, and a bead-beating homogenization step of the remaining particulate matter, which should enable recovery of additional RNA from cells with more resistant cell walls. Quantification of RNA from the two treatments revealed that each resulted in isolation of approximately equal amounts of total RNA, 13.4 µg RNA from the fraction that was only vortexed and 13.9 µg from the fraction homogenized by bead-beating. Again, to fully capture the metatranscriptome from both recalcitrant cells requiring bead-beating and more fragile cells lysed by vortexing in RLT lysis buffer, we combined these two fractions to create a single clone library. Although homogenizing the entire lysis buffer/cell suspension volume via bead-beating would have captured all the RNA without dividing into fractions, our concern was that this might unduly shear RNA already liberated

from the easily-lysed cells. Yet we did not want to lose the substantial RNA (approximately half of the total, as it was revealed) contained within cells that required more intense treatment to lyse. It may be that concerns over shearing were unfounded; RNA integrity of the fractions was not analyzed for this pilot project, as our concern was primarily if we could adequately capture eukaryotic protein coding sequences.

Regardless of quantities of total RNA and any contaminating DNA, we selected for messenger RNA for downstream processes via isolation of poly(A)-tailed mRNA using oligo(dT)-coated beads (the Qiagen Oligotex process). From the 27 µg total RNA processed, 720 ng was recovered using the Oligotex protocol, representing 2.7% of the total RNA. This should have been rather restricted to eukaryotic origin; however, as we observed from sequencing of clones, a small fraction was putatively of prokaryotic origin or rRNA (approximately 1% and 2%, respectively; Table 1 and Fig. 1). This mRNA was amplified using chemistry wherein an oligo(dT) reverse-transcription primer complementary to the 3' poly(A) tail is linked to a T7 phage RNA polymerase promoter sequence; hybridization to the poly(A) tail and reverse transcription forms a 5' extension with the T7 promoter on the antisense cDNA strand. The reverse transcription primer was anchored such that it should anneal to the beginning of the poly(A) tail. By use of T7 RNA polymerase, a starting mass of 192 ng purified mRNA was amplified to 173 µg antisense RNA (aRNA). The MessageAmp protocol thus resulted in an approximately 900-fold amplification of community RNA. If all 720 ng of poly(A) RNA had been amplified using the MessageAmp technique, we could have theoretically generated a total of 648 µg amplified RNA.

Conversion of the antisense RNA to cDNA for blunt-end cloning used reverse-transcriptase for first-strand creation, based on random hexamer primers, and DNA polymerase I (Klenow fragment) and additional random hexamer primers for second-strand synthesis. For the first-strand synthesis, 25 µg aRNA was used, and a fraction of this reaction (27%) equating to 6.75 µg of the original template was used for second-strand synthesis. Quantification of double-stranded cDNA using Hoechst 33258 revealed that from a processed mass of 6.75 µg equivalent aRNA, approximately 1.5 µg cDNA was obtained after second-strand synthesis. This equates to a conversion efficiency of about 11%, accounting for the molecular weight difference between RNA and DNA. Nonetheless, the amount of DNA recovered was sufficient to produce a large number of clones for sequencing, from the equivalent of only 78 ng RNase-digested, phosphorylated, and purified DNA. If all 720 ng poly(A)-enriched RNA were amplified using the MessageAmp procedure (resulting in 648 µg) and converted to cDNA, even at an efficiency of only 11%, the theoretical yield created for downstream processes would be 71.3 µg.

A total of 912 clones were screened by PCR, and 367 (40%) of these contained an insert. The pSMART vector system does not enable blue-white colony screening, so 60% of the transformants screened did not have vector that ligated an insert

from the cDNA. Of those that did contain an insert, 266 (29.1% of the total clones screened) contained an insert of at least 200 bp. As this was a pilot project, only a few hundred clones (232) were sent for sequencing. We estimated based on volumes plated and the number of colonies grown that our protocol could produce 40,000 transformants per electroporation transformation reaction. Based on PCR colony screening, we would estimate that about 12,000 clones (~30%) would contain inserts greater than 200 bp. Based on the sequencing results, we can estimate that 25.4% (59 of 232 for our submissions) of those would contain putative protein-coding sequences with BLAST scores  $\geq 40$  bits, whereas 2.1% were homologous to rRNA sequences. Thus, we can extrapolate that each transformation could produce approximately 3050 sequences with homology to known proteins. If 78 ng of cDNA were phosphorylated and ligated, and because only 1 µL of the 10 µL ligation mixture is used for each transformation, 10 transformations could be produced from every 78 ng of cDNA. Thus, extrapolating our theoretical yield of clones to the maximum, if all 720 ng of poly(A)-enriched RNA were amplified and converted to cDNA, which could produce up to 71 µg cDNA, this could produce up to 9100 transformation reactions, resulting in a potential yield of 27.755 million sequences with homology to known protein-coding sequences! Clearly such a theoretical maximum would not be achievable due to other limiting resources (funding, personnel, time, etc.), but lack of available nucleic acid material should not be among them.

*Clone library analysis*—Sequenced cDNA (metatranscriptome) clones were analyzed via translated (amino acid) and nucleotide BLAST searches against the NCBI GenBank database (Altschul et al. 1990). BLAST results of either type that returned a score of 40 bits or higher were identified as hits for further characterization. Of the 232 clones sent for sequencing, 64 (27.5%) returned search results meeting this threshold. The average query size of acceptable hits was 293 nucleotides, and the median was 243.5 nucleotides. The average BLAST score from these was 93.8 bits, and the geometric mean of expectation values was  $3.54 \times 10^{-15}$ . Table 1 displays hits that corresponded to protein or hypothetical protein genes, along with the organism of the highest-ranking database listing. Listings labeled as “similar” indicate that the highest-ranking hit did not specify a putative protein product (only “predicted protein”), so the organism of this hit is given with the best match or consensus among other, better-described proteins. Omitted from this table are five hits to rRNA gene sequences, four related to mitochondria of the dinoflagellate *Karlodinium micrum*, and one to the mitochondria of the calenoid copepod *Labidocera jollae*.

Table 2 summarizes the putative functions of genes identified from the metatranscriptome. About half (48%) are related to intracellular biochemical metabolism, such as protein and nucleic acid modification or metabolism (synthesis and degradation), and ribosomal protein-related genes. Also identified

**Table 1.** Listing of protein coding genes identified by translated BLAST (BLASTX) searches from the Tampa Bay metatranscriptome.

Accession no.	Predicted protein homolog	Organism of closest hit	Score, bits	Identities, %	Homolog accession no.
GHS71923	RuBisCO small subunit <i>N</i> -methyltransferase I	<i>Phaeodactylum tricornutum</i> (diatom)	366	71	XP_002185762
GHS71924	Sim. phosphoglycerate kinase precursor	<i>Thalassiosira pseudonana</i> (diatom)	330	84	AC164945
GHS71925	Ubiquitin-activating enzyme E1, protein 3	<i>Phaeodactylum tricornutum</i> (diatom)	315	70	XP_002178207
GHS71926	Elongation factor 1 alpha long form	<i>Reticulomyxa filosa</i> (amoeba)	224	58	ACF24599
GHS71927	Glyceraldehyde-3-phosphate dehydrogenase precursor	<i>Odontella sinensis</i> (diatom)	210	85	AAF34326
GHS71928	Zeaxanthin epoxidase	<i>Phaeodactylum tricornutum</i> (diatom)	160	70	XP_002178367
GHS71929	Caltractin	<i>Karlodinium micrum</i> (dinoflagellate)	155	100	ABV22245
GHS71930	Proteorhodopsin	Uncultured marine bacterium HF10_19P19	139	62	ABL60988
GHS71931	Sim. deoxyribonuclease tatD	<i>Tribolium castaneum</i> (beetle)	132	49	XP_970061
GHS71932	Extrinsic protein in photosystem II	<i>Chaetoceros gracilis</i> (diatom)	128	81	BAG85209
GHS71933	Extrinsic protein in photosystem II	<i>Chaetoceros gracilis</i> (diatom)	126	83	BAG85209
GHS71934	Sim. hydrolase alpha/beta fold family	<i>Phaeodactylum tricornutum</i> (diatom)	116	76	XP_002183464
GHS71935	Sim. <i>N</i> -acetyl-gamma-glutamyl-phosphate reductase	<i>Phaeodactylum tricornutum</i> (diatom)	114	72	XP_002185924
GHS71936	Ribosomal protein S15a	<i>Ostreococcus lucimarinus</i> (prasinophyte)	110	86	XP_001415819
GHS71937	Mitochon. solute carrier/adenosylmethionine transporter	<i>Phaeodactylum tricornutum</i> (diatom)	106	43	XP_002180544
GHS71938	Silicon transporter protein	<i>Skeletonema japonicum</i> (diatom)	104	96	ABF50464
GHS71939	DNA replication factor C complex subunit 3	<i>Theileria parva</i> (protozoa)	95.9	40	XP_764015
GHS71940	Sim. 60S ribosomal protein L10	<i>Physcomitrella patens</i> (moss)	93.2	62	EDQ83353
GHS71941	Sim. MPBQ/MSBQ transferase	<i>Ostreococcus lucimarinus</i> (prasinophyte)	88.2	84	XP_001416301
GHS71942	ADP-ribosylation factor	<i>Phaeodactylum tricornutum</i> (diatom)	87.8	97	XP_002177519
GHS71943	Acidic ribosomal protein P2	<i>Chlamydomonas reinhardtii</i> (chlorophycean)	86.7	74	XP_001694856
GHS71944	Light-harvesting chlorophyll a-c binding protein	<i>Phaeodactylum tricornutum</i> (diatom)	86.3	47	XP_002184763
GHS71945	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	83.6	78	XP_002182330
GHS71946	Dienelactone hydrolase family protein	<i>Aspergillus fumigatus</i> (mold)	80.5	44	XP_750459
GHS71947	Sim. Mg-dependent phosphatase (ISS)	<i>Ostreococcus lucimarinus</i> (prasinophyte)	75.5	46	XP_001422082
GHS71948	Fucoxanthin chlorophyll a/c protein	<i>Phaeodactylum tricornutum</i> (diatom)	73.6	56	XP_002183291
GHS71949	Putative methionine sulfoxide reductase	<i>Phaeodactylum tricornutum</i> (diatom)	73.2	85	XP_002180919
GHS71950	Sim. proteasome beta-subunit	<i>Phaeodactylum tricornutum</i> (diatom)	73.2	49	XP_002178278
GHS71951	Put. serine/threonine protein kinase domain	<i>Xenopus laevis</i> (frog)	69.7	40	NP_001089194
GHS71952	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	69.3	67	XP_002180802
GHS71953	Sim. pyrroline-5-carboxylate reductase	<i>Phaeodactylum tricornutum</i> (diatom)	68.9	46	XP_002184346
GHS71954	ATP-dependent RNA helicase	<i>Schizosaccharomyces pombe</i> (yeast)	66.6	71	XP_002175680
GHS71955	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	64.0	55	XP_002179046
GHS71956	Sim. cob(II)alamin adenosyltransferase	<i>Ostreococcus lucimarinus</i> (prasinophyte)	60.5	54	XP_001422511
GHS71957	Sim. signal recognition particle subunit 9	<i>Physcomitrella patens</i> (moss)	58.9	45	EDQ69379
GHS71958	Sim. protein kinase domain protein	<i>Thalassiosira pseudonana</i> (diatom)	57.8	43	AC164862
GHS71959	Sim. 40S ribosomal protein S24	<i>Strongylocentrotus purpuratus</i> (sea urchin)	53.1	60	XP_797339
GHS71960	Trypsin	<i>Aedes aegypti</i> (mosquito)	52.0	91	XP_001660676
GHS71961	Proteasome alpha subunit	<i>Phaeodactylum tricornutum</i> (diatom)	52.0	80	XP_002185881
GHS71962	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	51.2	68	XP_002182254

continued

Table 1. Continued

Accession no.	Predicted protein homolog	Organism of closest hit	Score, bits	Identities, %	Homolog accession no.
GH571963	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	48.5	37	XP_002181515
GH571964	Sim. divalent cation transporter	<i>Phaeodactylum tricornutum</i> (diatom)	47.4	89	XP_002177316
GH571965	Titin b	<i>Culex quinquefasciatus</i> (mosquito)	47.0	46	XP_001844102
GH571966	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	47.0	62	XP_002180802
GH571967	Elongation factor-like protein	<i>Skeletonema costatum</i> (diatom)	46.6	100	BAG30808
GH571968	Sim. ABC transporter protein	<i>Pediculus humanus corporis</i> (louse)	46.4	35	EEB12306
GH571969	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	46.2	45	XP_002177532
GH571970	Ribosomal protein L18a	<i>Chlamydomonas reinhardtii</i> (chlorophycean)	45.8	67	XP_001689743
GH571971	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	45.4	70	XP_002178113
GH571972	Protein kinase domain containing protein	<i>Tetrahymena thermophila</i> (ciliate)	43.9	65	XP_001013155
GH571973	Delta carbonic anhydrase	<i>Thalassiosira weissflogii</i> (diatom)	42.7	86	AAV39532
GH571974	RuBisCO, small subunit	<i>Ostreococcus tauri</i> (prasinophyte)	41.6	69	CAL58651
GH571975	Nonribosomal peptide synthetase 10	<i>Cochliobolus heterostrophus</i> (fungus)	41.6	62	AAV09992
GH571976	Sim. putative transcriptional accessory protein	<i>Burkholderia pseudomalleri</i> (bacterium)	41.2	43	ZP_03452285
GH571977	Sim. LRP16 protein (ADP-ribose binding domain)	<i>Halobacterium</i> sp. NRC-1 (bacterium)	41.2	36	NP_280675
GH571978	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	40.8	79	XP_002176682
GH571979	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	40.8	61	XP_002183690
GH571980	Aspartyl beta-hydroxylase (ISS)	<i>Ostreococcus tauri</i> (prasinophyte)	40.4	60	CAL52009
GH571981	Predicted protein	<i>Phaeodactylum tricornutum</i> (diatom)	40.0	88	XP_002183376

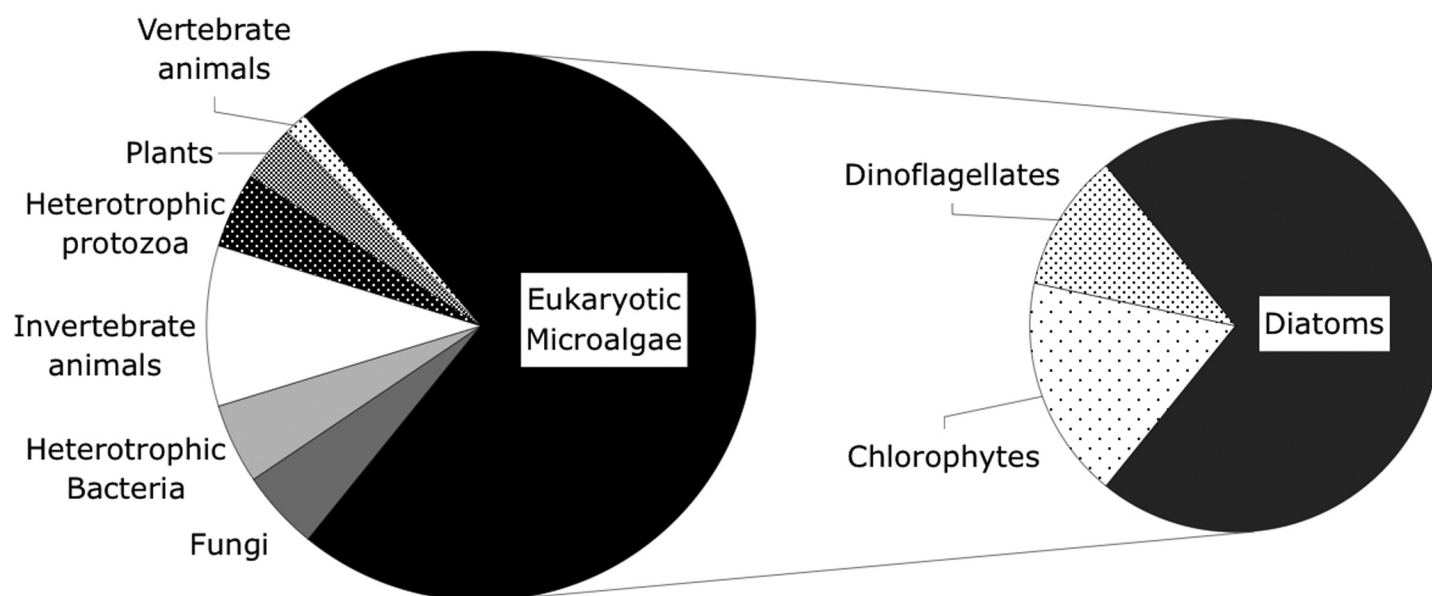
Accession numbers of the sequences derived from our library are GH571923 through GH571981 in the EST database. Corresponding organisms, BLAST scores, identity percentages (of amino acid sequences over homologous areas only, not entire query), and GenBank accession numbers of the homolog are listed. The predicted protein homolog is the closest identifiable protein to the best match, in cases where the top entry was listed as a hypothetical or unknown protein. Sim. = similar to, indicating that another entry was used as a reference to the gene identity rather than that from the best hit indicated by the organism and accession numbers.

were important transcripts related to nutrient acquisition or metabolism of autotrophs (8%), such as a silicon transporter and carbonic anhydrase, and light harvesting and photosynthetic pigment associated genes (8%). Eleven of the 64 were homologous to predicted proteins such that we could not identify a putative function; all these were hits to the newly sequenced genome of the diatom *Phaeodactylum tricornutum*.

The phylogenetic associations of resulting sequences are summarized in Fig. 1. More than 70% of the hits were putatively derived from eukaryotic algae, namely diatoms, chlorophytes, and dinoflagellates. Smaller proportions of the sequences corresponded to other types of organisms: invertebrate animals, fungi, heterotrophic protozoa, heterotrophic bacteria, plants, and vertebrate animals. The source organism most commonly identified for homologous sequences to our metatranscriptome library (25 of 64 hits) was the diatom *Phaeodactylum tricornutum* CCAP 1055/1, whose genome and predicted protein-coding sequences were only recently updated to GenBank (Bowler et al. 2008). It is important to note that for many eukaryotic genes, sequences may be very similar among seemingly distantly related organisms, and the GenBank database is heavily weighted in favor of those organisms most frequently studied for molecular genetics or whose genomes have been sequenced. Many sequences we identified likely derived from organisms whose particular gene has not been characterized. It is possible that a significant fraction of the remaining 182 sequences that did not indicate homology to known organisms' genes may have been sequences from organisms that have simply not been characterized or submitted to GenBank.

## Discussion

This is the first report of a metatranscriptome from eukaryotic marine plankton. In conducting this experiment as a proof-of-concept exercise, we attempted to establish sample processing and molecular biology techniques that will enable future, more intensive constructions of transcriptome profiles from marine phytoplankton samples. In targeting eukaryotic organisms, we circumvented the significant challenge of some previously reported efforts toward obtaining environmental metatranscriptomes from prokaryotic organisms (Frias-Lopez et al. 2008, Poretsky et al. 2005). The presence of poly(A) tails on eukaryotic mRNAs allowed for selection



**Fig. 1.** Taxonomic associations of metatranscriptome sequences from Tampa Bay. Associations are based on the identity of the best matching hit from BLAST searches and categorized as shown. About 70% of the sequences were putatively derived from eukaryotic phytoplankton, which could be further divided as shown in the subchart on right.

**Table 2.** Putative functions of the characterized protein products from transcriptome sequences.

Putative gene function	Fraction, %
Protein synthesis/modification/degradation	19
Unknown	17
Nucleic acid synthesis/modification/degradation	8
rRNA gene	8
Ribosomal protein	8
Nutrient acquisition/metabolism	8
Light harvesting and photosynthesis	8
Amino acid metabolism	6
Mitochondrial metabolism	3
Vitamin/cofactor synthesis	3
Lipid metabolism	3
Ester/xenobiotic metabolism	2
Motility	2
Proteorhodopsin	2
Immune system	2
Intracellular vesicle	2
Membrane transport	2

General function associations were determined as necessary from the protein family (pfam), conserved domain, and COG databases of the NCBI. General cellular functions such as protein and nucleic acid synthesis, modification, or degradation and ribosome-related genes were most frequently observed.

of these molecules using well-established, commercially available means. The selection method applied was quite successful, in that only five of 232 sequenced clones (2.2%) corresponded

to rRNA (all eukaryotic organisms) and only three of 232 (1.3%) were most closely related to prokaryotic RNA, whereas 25.4% were homologous to known eukaryotic protein coding sequences. Because mRNA makes up only a small portion of total cellular RNA, most being in the form of rRNA, the presence of a relatively large percentage of eukaryotic mRNA sequences in our library indicates adequate selection via both the Oligotex mRNA selection kit and amplification of poly(A) RNA using the oligo(dT) primer in the MessageAmp procedure.

Approximately 70% of scored hits were related to phytoplankton, specifically diatoms, chlorophytes, and dinoflagellates. Gene sequences uncovered in the pilot metatranscriptome reveal a diverse assemblage of cellular activities. Transcripts related to a number of general cellular processes such as protein, amino acid, and nucleic acid synthesis, modification, and degradation were observed, as would be expected. Also significant are the numerous protein groups related to autotrophic/ photosynthetic growth and biogeochemical function, such as a silicon transporter, vitamin or cofactor metabolism, carbon acquisition/fixation (carbonic anhydrase and RuBisCO small subunit), light harvesting or pigment proteins, and phosphatases. Naturally, all these functional gene groups were represented by sequences from phytoplankton. The identification of a putative diene lactone hydrolase gene was interesting. Proteins of this class enable degradation of chlorocatechol compounds, a class of organic molecules often known to be central intermediates in the biodegradation pathway of a large array of chlorinated compounds including polychlorinated aromatic compounds (PCBs) (Guan et al. 2000, Neilson 1990). Its finding from a putative fungi in this sample is perhaps not surprising

due to the large amount of urban surface water runoff entering Tampa Bay. The other genes putatively derived from fungi are interesting as well, since little is known about the role of fungi in the marine environment. However, marine fungi are often the target of research on natural products potentially useful in pharmacology (Kamat et al. 2008, Kasetrathat et al. 2008, Mayer and Gustafson 2008) and possibly bioremediation, as suggested by incidence of the dienelactone hydrolase-like transcript.

In the context of other metatranscriptome studies, the study by Frias-Lopez et al., although focused on prokaryotes, is most similar to the work presented here. Genes they identified as most highly expressed in the oligotrophic sample site (Hawaii Ocean Time Series) were related to photoautotrophy and photoheterotrophy. Transcripts of photosystem I and II genes, RuBisCO large subunit, and light-harvesting protein genes were among the most frequently identified, as well as glutamine synthase and proteorhodopsin (light-activated ion transport, also identified from our library). Likewise, transcripts involved with autotrophy, such as inorganic nutrient uptake, as well as photosystem and photosynthetic pigment proteins, were found in our library. Because of the small size of our library, any transcripts found are likely to be more highly expressed in the environmental population. We did not find any RuBisCO large-subunit gene transcripts (*rbcL*) in the eukaryotic transcriptome, but since this is a plastid gene, it is not typically poly(A)-tailed and thus would not be expected in abundance in our library. We did identify a RuBisCO small-subunit (*rbcS*) transcript. However, *rbcS* in some organisms, particularly green algae and plants, is nuclear-encoded (Ellis 1981); the *rbcS* transcript we identified was most closely related to the green algae *Ostreococcus tauri* (a prasinophyte). Transcripts from additional classes of proteins identified from our library were also found to be highly expressed in the Frias-Lopez et al. marine microbial metatranscriptome, such as protein synthesis genes (including ribosomal proteins), nucleic acid modification, and transport protein genes.

Regarding the proportion of rRNA transcripts obtained, the challenges of isolating protein-coding transcripts from prokaryotes is apparent upon comparing the number we identified, 2% of the total sequences obtained, versus the 53% rRNA sequences identified in the Frias-Lopez et al. study. In the eukaryotic metatranscriptomes of Grant et al. (2006), the scale of their sequencing was even smaller than ours, but one important point is the recovery of rRNA, which from the samples they processed without poly(A) enrichment was 60% from the algal mats and 20% from the activated sewage sludge. However, poly(A) enrichment of a fraction of the sludge sample reduced their rRNA recovery to 4% (one in 23 sequences). It is evident that although poly(A) enrichment of eukaryotic mRNA can dramatically reduce the amount of rRNA retained and eventually cloned/sequenced, some will unavoidably pass through the purification and poly(A)

tail-based amplification. This will only slightly reduce the efficiency of protein-coding sequence recovery from eukaryotic metatranscriptome libraries.

The identification of such a diverse group of sequences from this small library suggests that much larger metatranscriptomic sequencing projects will be quite informative of gene expression activity from phytoplankton communities. Characterization of actively transcribed genes would be an important complement to metagenomic studies, which identify the complement of genes present in an environment. The use of larger-scale metatranscriptome techniques could allow future targeting of function-related activity for molecular biological studies, without a priori knowledge of important functional gene groups or sequence variants present in the sample. The pairing of microarray technology with basic information uncovered by environmental transcriptomics could be particularly worthwhile.

### Comments and recommendations

Based on the results of sequencing transcripts cloned using the methods described here, we believe that successful capturing of the eukaryotic plankton metatranscriptome could be enabled from other marine samples. We obtained only 2% rRNA sequences, and only three of the 64 sequences with good homology to known genes/coding sequences were ostensibly derived from prokaryotic organisms. Of these, only the proteorhodopsin would unequivocally be derived from prokaryotes. The remaining two were relatively poor hits to different sequences (oddly, both with scores of 41.2 bits and expectation values of 0.029; verification revealed this is not an error and they are not homologous to each other). In both cases, other, slightly less robust, homologies were found to protein sequences of various eukaryotic and other prokaryotic organisms, many of which were uncharacterized predicted/hypothetical protein coding sequences. It seems likely that these two clones from our library are derived from an organism that is poorly characterized or poorly represented in GenBank, and their identity as prokaryotic sequences is uncertain. Thus we feel comfortable that the methods used for RNA extraction, poly(A) enrichment and poly(A)-based amplification successfully selected against prokaryotic organisms. We did not perform a DNase digestion, and there is no indication from our results that this compromised the selection of eukaryotic mRNA sequences. However, to fully rule out the possibility of capturing genomic sequences rather than mRNA, a DNase digestion of purified RNA would be advised. Also, during the course of sample processing, the additional sample water used to wash captured cells from the filters was not 0.2  $\mu\text{m}$  filtered; this volume represented approximately 6% of the total filtered. Although some prokaryotes were likely already entrained on the 2- $\mu\text{m}$  filters and larger organisms captured thereon, using unfiltered water likely introduced additional prokaryotic cells to the sample, and to minimize their potential impact, rinsing water should be filtered.

The methods employed in this pilot project for capturing transcriptomic sequences used cloning in a pSMART vector to create a library. Although we estimated that potentially millions of protein-coding sequences could be obtained from eukaryotic plankton with the use of linear RNA amplification technology, the construction and screening of larger libraries would be quite cumbersome. If larger efforts were to be undertaken with cloning, the use of a vector enabling blue-white colony screening would be strongly advised. Although the pSMART system we employed claims that blue-white screening is not necessary due to results indicating >99% of transformants should contain plasmid with insert, we found that our ratio was only 40% (of transformants screened that carried a cloned insert at all), while only 29% carried an insert above our minimum desired size of 200 bp. The result was the screening of many more clones than were actually sequenced. Thus, the use of  $\beta$ -galactosidase screening or the like would improve efficiency of clone selection. Alternatively, using a greater mass of cDNA (we used 78 ng) in the ligation reaction may improve cloning efficiency.

The use of cloning for large-scale projects would probably be unadvisable in light of the development of pyrosequencing technology (e.g., Cheung et al. 2006, Frias-Lopez et al. 2008). The size of positive hits in our transcriptomic library averaged 293 nt (median 243.5 nt); we intentionally selected clones with an insert of at least 200 bp, although in some cases poor-quality ends or other quality issues resulted in BLAST query sequences significantly shorter (the shortest query with a significant match was 83 nt, to the *Karodinium micrum* LSU rRNA gene). Whereas these average sizes are larger than obtainable with previous pyrosequencing methods (which were on average 150–200 bp reads with the GS 20 technology), newer technology is enabling greater read lengths of an average 250 bp using the GS FLX instrument and possibly soon 500 bp (Blow 2008). For characterizing environmental transcript data sets, longer reads would be critical; read length is shown to be important for establishing relatedness to sequences with lesser degrees of homology (Wommack et al. 2008). Thus with the ability of future pyrosequencing instruments to return read lengths of 500 bp, the greater throughput and lower cost per read of this technology would provide a great advantage over Sanger sequencing of cloned environmental transcripts. For future use in pyrosequencing, the yield of cDNA produced in this project would be more than adequate. We calculated that if all the poly(A)-enriched RNA we obtained were amplified and converted to cDNA, we could obtain a total of 71  $\mu$ g double-stranded cDNA; current pyrosequencing methods typically require at least 5  $\mu$ g template.

Although we have demonstrated the feasibility of characterizing gene transcripts from eukaryotic plankton populations, our work should be viewed as proof of concept. Future efforts, eventually employing pyrosequencing for much larger datasets, will hopefully enable more complete descriptions of the metatranscriptome from varying marine environments.

This study provides a basis for enhanced development of environmental transcriptomic methods.

## References

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
- Blow, N. 2008. Metagenomics: Exploring unseen communities. *Nature* 453:687-690.
- Bowler, C. and others 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456:239-244.
- Cheung, F., B. J. Haas, S. M. D. Goldberg, G. D. May, Y. L. Xiao, and C. D. Town. 2006. Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *Bmc Genomics* 7.
- Dyrhrman, S. T., S. T. Haley, S. R. Birkeland, L. L. Wurch, M. J. Cipriano, and A. G. McArthur. 2006. Long serial analysis of gene expression for gene discovery and transcriptome profiling in the widespread marine coccolithophore *Emiliania huxleyi*. *Appl. Environ. Microbiol.* 72:252-260.
- Eberwine, J. and others 1992. Analysis of gene-expression in single live neurons. *Proc. Natl. Acad. Sci. USA* 89:3010-3014.
- Ellis, R. J. 1981. Chloroplast proteins - Synthesis, transport, and assembly. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 32:111-137.
- Erdner, D. L., and D. M. Anderson. 2006. Global transcriptional profiling of the toxic dinoflagellate *Alexandrium fundyense* using massively parallel signature sequencing. *Bmc Genomics* 7.
- Feldman, A. L. and others 2002. Advantages of mRNA amplification for microarray analysis. *Biotechniques* 33:906-912, 914.
- Frias-Lopez, J. and others 2008. Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. USA* 105:3805-3810.
- Grant, S. and others 2006. Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. *Appl. Environ. Microbiol.* 72:135-143.
- Guan, X. and others 2000. Chlorocatechol detection based on a *clc* operon/reporter gene system. *Analytical Chemistry* 72:2423-2427.
- Kacharina, J. E., P. B. Crino, and J. Eberwine. 1999. Preparation of cDNA from single cells and subcellular regions, p. 3-18. *Cdna Preparation and Characterization. Methods in Enzymology.*
- Kamat, T., C. Rodrigues, and C. G. Naik. 2008. Marine-derived fungi as a source of proteases. *Indian J. Mar. Sci.* 37:326-328.
- Karrer, E. E. and others 1995. In situ isolation of mRNA from individual plant cells: creation of cell-specific cDNA libraries. *Proc Natl Acad Sci U S A* 92:3814-3818.

- Kasetrathat, C., N. Ngamrojanavanich, S. Wiyakrutta, C. Mahidol, S. Ruchirawat, and P. Kittakoop. 2008. Cytotoxic and antiplasmodial substances from marine-derived fungi, *Nodulisporium* sp and CRI247-01. *Phytochemistry* 69:2621-2626.
- Li, Y. and others 2004. Systematic comparison of the fidelity of aRNA, mRNA and T-RNA on gene expression profiling using cDNA microarray. *J. Biotechnol.* 107:19-28.
- Mayer, A. M. S., and K. R. Gustafson. 2008. Marine pharmacology in 2005-2006: Antitumour and cytotoxic compounds. *Eur. J. Cancer* 44:2357-2387.
- Mock, T., A. Krell, G. Glockner, U. Kolukisaoglu, and K. Valentin. 2006. Analysis of expressed sequence tags (ests) from the polar diatom *fragilariopsis cylindrus*. *J. Phycol.* 42:78-85.
- Mock, T. and others 2008. Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. *Proc. Natl. Acad. Sci. USA* 105:1579-1584.
- Moreno-Paz, M., and V. Parro. 2006. Amplification of low quantity bacterial RNA for microarray studies: time-course analysis of *Leptospirillum ferrooxidans* under nitrogen-fixing conditions. *Environ. Microbiol.* 8:1064-1073.
- Neilson, A. H. 1990. The biodegradation of halogenated organic compounds - A review. *Journal of Applied Bacteriology* 69:445-470.
- Pabon, C. and others 2001. Optimized T7 amplification system for microarray analysis. *Biotechniques* 31:874-879.
- Paul, J. H., and B. Myers. 1982. Fluorometric determination of DNA in aquatic microorganisms by use of Hoechst-33258. *Appl. Environ. Microbiol.* 43:1393-1399.
- Polacek, D. C. and others 2003. Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA. *Physiol. Genomics* 13:147-156.
- Poretsky, R. S. and others 2005. Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* 71:4121-4126.
- Scala, S., and C. Bowler. 2001. Molecular insights into the novel aspects of diatom biology. *Cell. Mol. Life Sci.* 58:1666-1673.
- Vangelder, R. N., M. E. Vonzastrow, A. Yool, W. C. Dement, J. D. Barchas, and J. H. Eberwine. 1990. Amplified RNA synthesized from limited quantities of heterogenous cDNA. *Proc. Natl. Acad. Sci. USA* 87:1663-1667.
- Wommack, K. E., J. Bhavsar, and J. Ravel. 2008. Metagenomics: read length matters. *Appl. Environ. Microbiol.* 74:1453-1463.

*Submitted 4 September 2008*

*Revised 27 January 2009*

*Accepted 18 February 2009*