Harmful Algae 37 (2014) 75–83

Contents lists available at ScienceDirect

Harmful Algae

journal homepage: www.elsevier.com/locate/hal

Metatranscriptome profiling of a harmful algal bloom

Endymion D. Cooper [a,*], Bastian Bentlage [a], Theodore R. Gibbons [a], Tsvetan R. Bachvaroff [b], Charles F. Delwiche [a,c]

[a] CMNS-Cell Biology and Molecular Genetics, 2107 Bioscience Research Building, University of Maryland, College Park, MD 20742-4407, USA
[b] Institute of Marine and Environmental Technology, University of Maryland Center for Environmental Science, 701 E Pratt St., Baltimore, MD 21202, USA
[c] Maryland Agricultural Experiment Station, AGNR, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 2 October 2013
Received in revised form 16 April 2014
Accepted 17 April 2014

Keywords:
De novo assembly
Dinoflagellate
Flow cytometry
Harmful algal bloom
Metagenomics
Metatranscriptomics
Prorocentrum minimum
Transcriptomics

ABSTRACT

Metagenomic methods provide a powerful means to investigate complex ecological phenomena. Developed originally for study of Bacteria and Archaea, the application of these methods to eukaryotic microorganisms is yet to be fully realized. Most prior environmental molecular studies of eukaryotes have relied heavily on PCR amplification with eukaryote-specific primers. Here we apply high throughput short-read sequencing of poly-A selected RNA to capture the metatranscriptome of an estuarine dinoflagellate bloom. To validate the metatranscriptome assembly process we simulated metatranscriptomic datasets using short-read sequencing data from clonal cultures of four algae of varying phylogenetic distance. We find that the proportion of chimeric transcripts reconstructed from community transcriptome sequencing is low, suggesting that metatranscriptomic sequencing can be used to accurately reconstruct the transcripts expressed by bloom-forming communities of eukaryotes. To further validate the bloom metatransciptome assembly we compared it to a transcriptomic assembly from a cultured, clonal isolate of the dominant bloom-causing alga and found that the two assemblies are highly similar. Eukaryote-wide phylogenetic analyses reveal the taxonomic composition of the bloom community, which is comprised of several dinoflagellates, ciliates, animals, and fungi. The assembled metatranscriptome reveals the functional genomic composition of a metabolically active community. Highlighting the potential power of these methods, we found that relative transcript abundance patterns suggest that the dominant dinoflagellate might be expressing toxin biosynthesis related genes at a higher level in the presence of competitors, predators and prey compared to it growing in monoculture.

1. Introduction

Communities of microscopic organisms are ubiquitous and perform an enormous range of ecologically important functions. The organisms present within a community contribute to ecological function by their individual metabolic capabilities and through interacting metabolic networks. Great examples are marine systems, where autotrophs make a significant contribution to global carbon cycling. In contrast to any positive impact they might have on atmospheric carbon, periodic blooms of dinoflagellates can have devastating ecological and economic consequences through toxin production and localized oxygen depletion (Morey et al., 2011). The processes of bloom initiation, maintenance and termination are poorly understood (Cloern et al., 2005; Anderson and Rengefors, 2006; Adolf et al., 2008; Hense, 2010). Among the contributing parameters are community interactions and resource fluxes. Response to environmental variables and interacting partners is encoded in, and controlled by, the genome. The burgeoning field of metagenomics allows the composition and function of microscopic communities to be characterized and described as a whole, without the need to isolate and characterize the individual players and to experimentally reconstruct their networks of interactions.

Although the metagenomic study of Bacteria and Archaea over the last 20 years has revealed the tremendous power of such approaches for microbial ecology, comparable studies of eukaryotes have been relatively slow to emerge (Lin et al., 2010). Application of metagenomic techniques to eukaryotes has generally focused on the discovery of taxonomic diversity through rDNA surveys of diverse communities. These efforts have demonstrated that metagenomic studies will be valuable for eukaryotic microorganisms by revealing an extraordinary array of previously unknown eukaryotic lineages whose ecological roles remain obscure (Lopez-Garcia et al., 2001; Moon-van der Staay et al., 2001; Baldauf, 2003; Stoeck et al., 2006).

* Corresponding author. Tel.: +1 3014058300.
E-mail address: endymion.dante.cooper@gmail.com (E.D. Cooper).

http://dx.doi.org/10.1016/j.hal.2014.04.016
1568-9883/

Efforts to apply genome scale sequence surveys to eukaryotic communities have been limited. Eukaryotic genomes are typically far larger than those of prokaryotes and require significantly more sequencing depth to facilitate successful assembly of their genomes. Even single organism genome sequencing remains out of reach for many lineages. Dinoflagellates, a major group of marine eukaryotes, have extremely large and unusual genomes whose content and function remain largely unknown, and whose transcriptional and translational regulation fail to obey standard patterns (Wisecaver and Hackett, 2011). Most dinoflagellate genomes are thought to be too large and too repetitive to be accessible for sequencing at reasonable cost with current technology (Shoguchi et al., 2013). For organisms known only from environmental sampling of ribosomal RNA encoding sequences, we have no basis for evaluating the size and complexity of their genomes. Consequently, environmental genome sequencing of eukaryotic communities remains out of reach.

In contrast, transcriptome sequencing provides access to a wealth of genomic information from substantially less sequence data. Single-organism shotgun transcriptome sequencing has already been highly informative for resolving phylogenetic relationships deep in the tree of life (*e.g.*, Dunn et al., 2008; Oakley et al., 2012; Timme et al., 2012; Kvist and Siddall, 2013) and is starting to reveal the origins of genomic features which have enabled major evolutionary transitions like the conquest of land by plants (Timme and Delwiche, 2010; Viaene et al., 2013). The potential value of environmental sequencing of eukaryote transcriptomes can be seen in the study of eukaryotic soil communities, focused on fungi and other microbes (Damon et al., 2012). Applications in marine systems include automated remote sampling of mixed microbial communities (Ottensen et al., 2013) and the transcriptional response of diatoms to iron availability (Marchetti et al., 2012).

For environmental transcriptome sequencing to be successful sufficient sequencing depth is required to sample the vast majority of transcripts present and to accurately describe the functional and taxonomic composition of the sample. Of currently available sequencing technologies, Illumina (Bentley et al., 2008) provides the greatest amount of data for equivalent cost (Glenn, 2011). However, Illumina sequencing favors data quantity over read length and therefore each transcript must be inferred by assembly of contiguous sequences (contigs) from multiple independent reads. A major concern in metagenomics and metatranscriptomics is the potential assembly of chimeric contigs by merging transcripts from multiple species due to regions of high sequence identity (*e.g.*, Howe et al., 2014). In order to evaluate this potential source of error we simulated metatranscriptomic datasets for simple algal communities *in silico*. Because closely related organisms are likely to share more regions of high sequence identity, we hypothesized that phylogenetic distance will have a significant effect on the amount of chimerism in metatranscriptomic assemblies. We set out to test this hypothesis using simulated datasets comprised of two species of varying phylogenetic distance. We then applied Illumina short read sequencing and *de novo* transcriptome assembly to evaluate the feasibility of characterizing the taxonomic and functional composition of an environmental sample from a major dinoflagellate bloom. We validate the environmental data using a clonal culture of the dominant organism obtained from the bloom. Using this combination of *in silico* metatranscriptome simulations and applied environmental sequencing, we demonstrate that it is feasible to accurately assemble the transcripts expressed by a community of bloom-forming eukaryotes and to achieve informative taxonomic and functional annotation of these transcripts. The approach demonstrated here can be applied through the time course of an algal bloom and has the potential to yield important insights into the dynamic nature of species interactions during harmful algal blooms.

2. Materials and methods

2.1. Sample collection, culture isolation and RNA extraction and sequencing

The composition of the sample was analyzed using a custom built CytoSense portable flow cytometer (Cytobuoy b.v., Woerden, Netherlands). Undiluted sample was pumped through the flow cuvette at a rate of approximately 0.8 $\mu l\ s^{-1}$ for 300 s. After the initial run, repeat runs were performed under identical flow rate and triggering conditions, using a combination of length of signal and maximum red fluorescence to photograph 50 particles from each of the main clusters. Additionally, video and still images were captured using a Canon EOS 5D Mark II camera mounted on a Zeiss Axioscop using either a 10× Fluor or 40× plan-NeoFluor objective. Total RNA was extracted from a 50 ml subsample of the bloom within hours of collection, using Ambion RNAqueous extraction kit (Life Technologies) following the manufacturers protocol. Single cells of the dominant algal species were isolated with a sterile micropipette, washed by five serial transfers through sterile growth medium, and grown in L1-15ppt culture medium. Cultures were grown at 18 °C under cool white fluorescent illumination, 15 h day length, until a dense clonal culture was observed. One clonal culture was selected for transcriptome sequencing. Total RNA was extracted using Nucleospin RNA Plant extraction kit (Machery-Nagel) following the manufacturers protocol. Library preparation and sequencing were carried out by the University of Maryland's Institute for Bioscience and Biotechnology Research DNA Sequencing Facility (http://www.ibbr.umd.edu/facilities/sequencing). Poly-adenylated RNA was isolated and cDNA libraries prepared using the standard TruSeq protocol (Illumina). Indexing adapters were used and each sample was multiplexed with three other libraries on a single lane of an Illumina HiSeq1000 sequencer. On average 45 million paired-end, 101 base reads were produced for each library (Table 1).

2.2. Transcript assembly from an environmental sample and comparison to a clonal isolate

2.2.1. Comparison of the read sets and assembly statistics

FastQC version 0.10.1 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) was used for initial characterization of raw read sets. Reads that did not pass the Illumina chastity filter were discarded. Truseq adapter sequences were removed using TagCleaner version 0.12 (Schmieder et al., 2010) allowing up to 17 mismatches in order to account for partial tags and sequencing error. Reads were trimmed to remove all poor quality bases (phred score <20) using a python script (https://github.com/bastodian/shed/blob/master/Python/q-trim.py). Reads shorter than 25 bases

Table 1
Comparative statistics for theenvironmental and clonal isolate datasets.

	Environmental	Clonal
Raw reads (millions of pairs)	42.2	48.9
Trimmed reads – paired (millions of pairs)	28.4	41.2
Trimmed reads – singletons (millions)	3.7	3.8
Number of contigs	172,700	167,207
Number of components	128,806	113,967
Contig length range (min–max)	201–18,679	201–10,772
Mean contig length	750	816
Median contig length	469	543
NCBI TSA accession number	GBFZ00000000	GBFP00000000
Number of unique components	24,749	10,414
Number of shared components	104,058	103,554