Marine Genomics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

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

Method paper

Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*

Justin Ashworth [a,*], Serdar Turkarslan [a], Micheleen Harris [a], Mónica V. Orellana [a,b], Nitin S. Baliga [a,c,*]

[a] *Institute for Systems Biology, Seattle, WA, USA*
[b] *Polar Science Center, University of Washington, Seattle, WA, USA*
[c] *Department of Microbiology, University of Washington, Seattle, WA, USA*

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form 28 October 2015
Accepted 29 October 2015
Available online xxxx

Keywords:
Diatom
Transcriptomics
Clustering
Microarrays
Gene regulation

ABSTRACT

Diatoms are important primary producers in the ocean that thrive in diverse and dynamic environments. Their survival and success over changing conditions depend on the complex coordination of gene regulatory processes. Here we present an integrated analysis of all publicly available microarray data for the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. This resource includes shared expression patterns, gene functions, and *cis*-regulatory DNA sequence motifs in each species that are statistically coordinated over many experiments. These data illustrate the coordination of transcriptional responses in diatoms over changing environmental conditions. Responses to silicic acid depletion segregate into multiple distinctly regulated groups of genes, regulation by heat shock transcription factors (HSFs) is implicated in the response to nitrate stress, and distinctly coordinated carbon concentrating, CO_2 and pH-related responses are apparent. Fundamental features of diatom physiology are similarly coordinated between two distantly related diatom species, including the regulation of photosynthesis, cellular growth functions and lipid metabolism. These integrated data and analyses can be explored publicly (http://networks.systemsbiology.net/diatom-portal/).

1. Introduction

Diatoms are important primary producers in marine ecosystems (Armbrust, 2009; Tsuda et al., 2003), and their ability and capacity to physiologically adjust to changing ocean conditions are critical to their ecological success, both now and in the future. The heterokonts (including diatoms) are phylogenetically distinct from other clades, and represent relatively undiscovered genomic and physiological territory. Recent studies of diatom genomes (Armbrust et al., 2004; Bowler et al., 2008), transcriptomes and proteomes (Allen et al., 2008; Ashworth et al., 2013; Brembu et al., 2011; Carvalho et al., 2011; Chauton et al., 2013; Hook and Osborn, 2012; Kustka et al., 2014; Levitan et al., 2015; Mock et al., 2008; Nymark et al., 2009, 2013; Sapriel et al., 2009; Shrestha et al., 2012; Thamatrakoln et al., 2012, 2013; Valle et al., 2014) hint at complex molecular and regulatory programs that control diatom physiology and acclimation to a variety of different environmental conditions. However, the specific coordination and regulation of these molecular and adaptive processes in diatoms remain largely unknown for multiple reasons: i) the size and complexity of diatom genomes and their molecular responses to change, ii) their genetic uniqueness and lack of sufficient homology in comparison to other well-studied clades, iii) low-throughput experimental genetic approaches, and iv) insufficient data and analysis methods to identify concurrent yet distinct molecular and regulatory pathways operating simultaneously under various conditions.

The integration and analysis of data collected through many different transcriptomic experiments can be used to discover fundamentally coordinated, conditional, and distinct molecular responses that are reflective of gene regulatory processes and cannot be discovered through individual experiments (Brooks et al., 2014; Danziger et al., 2014; Reiss et al., 2006). Data-driven approaches for the discovery of molecular coordination may be particularly useful in the case of diatoms, given the size and novelty of their genomes, transcriptomes and proteomes. To systematically discover and identify modules of putatively coordinated and functionally related diatom genes, we aggregated all available microarray expression data for the model diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, and performed hierarchical clustering (Eisen et al., 1998) and motif-guided biclustering (Reiss et al., 2006) over many experimental conditions. Highlights of this analysis in the context of specific conditions and functions are discussed herein, and the complete results have been made available for further use and exploration in a web portal at the following url: http://networks.systemsbiology.net/diatom-portal/.

[*] Corresponding authors.
E-mail addresses: ashwortj@uw.edu (J. Ashworth), nitin.baliga@systemsbiology.org (N.S. Baliga).

http://dx.doi.org/10.1016/j.margen.2015.10.011
1874-7787/© 2015 Elsevier B.V. All rights reserved.

2. Data sources and methods

2.1. Integrated diatom transcriptomic dataset

Transcriptome-wide microarray expression data *T. pseudonana* used in this analysis included: silica, iron, and nitrogen limitation, low temperature and elevated pH (Mock et al., 2008), exposure to pollutant and mutagen benzo[*a*]pyrene (Carvalho et al., 2011), iron starvation (Thamatrakoln et al., 2012), silica re-supplementation (Shrestha et al., 2012), diel growth from exponential to stationary phase (Ashworth et al., 2013), and growth at moderate and elevated CO_2 levels under moderate and elevated light (Gene Expression Omnibus (GEO) accession #GSE57737). Transcriptome-wide microarray expression data for *P. tricornutum* used in this analysis included: silica limitation (Sapriel et al., 2009), acclimation to high light (Nymark et al., 2009), exposure to cadmium (Brembu et al., 2011), acclimation to light and dark cycles (Chauton et al., 2013), exposure to a panel of pollutants (Hook and Osborn, 2012), darkness and re-illumination (Nymark et al., 2013), exposure to red, blue and green light (Valle et al., 2014). All public microarray data were downloaded from GEO, converted to $\log_2$ expression ratios vs. within-experiment control samples, and aggregated by common transcript identifiers for each species. The normalization performed within these independent experimental datasets was preserved in order to reflect previous findings; further approaches to normalize these data were explored in conjunction with various distance metrics and hierarchical clustering methods, as described below in the section: Co-Expression Clustering.

2.2. Genome information, protein annotations, functional enrichment analysis and predictions of orthology

The genomes, gene, transcript and protein models, and functional annotations for each diatom species (*T. pseudonana*: version 3; *P. tricornutum*: version 2) were downloaded from the Joint Genome Institute (JGI) Genome Portal (Grigoriev et al., 2011). Existing protein functional annotations were supplemented with significant matches to the Conserved Domain Database (Marchler-Bauer et al., 2015). The enrichment of GO terms (biological process) across all co-expression clusters was computed from existing JGI GO annotations using a hypergeometric test and multiple hypothesis correction by the Benjamini–Hochberg method (Benjamini and Hochberg, 1995). Putative orthologous proteins between diatom species were inferred using a pairwise reciprocal best BLASTp (Altschul et al., 1990) approach with an E-value cutoff of 1×10^{-10}. Significantly orthologous co-expression clusters between species were identified by computing the hypergeometric probability of drawing intersections of orthologous genes (Kalinka, 2013) and multiple hypothesis correction (Benjamini and Hochberg, 1995).

2.3. Co-expression clustering

The co-expression of transcripts across all included experiments was assessed using a Pearson correlation distance metric over all corresponding $\log_2$ expression ratios for each pair of transcripts. Fast agglomerative hierarchical clustering (Müllner, 2013) was used to generate a hierarchical tree of correlated transcripts, and ten thousand-fold multi-scale bootstrap resampling (Suzuki and Shimodaira, 2006) was used to estimate the significance and robustness of hierarchical subclusters to simulated noise and variation in the data. For efficiency on large transcriptomes, all pairwise distance calculations, hierarchical clustering using the fastcluster c++ library (Müllner, 2013), and performance-critical functions of the pvclust R package (Suzuki and Shimodaira, 2006) were computed using compiled c++ code that is available for download. The false discovery rate of cluster-level changes in expression over categorical conditions was computed based on empirical distributions ($n \geq 10{,}000$) of clusters composed of randomly-selected transcripts for each cluster size over each condition, with multiple hypothesis correction to account for total number of clusters. Significantly shared DNA sequence motifs in the upstream promoter regions of genes co-occurring in hierarchical co-expression clusters were detected using MEME (Bailey and Elkan, 1994) and compared to databases of known motifs using TOMTOM (Gupta et al., 2007), both of which are part of the MEME software suite (Bailey et al., 2009).

2.4. DNA motif-guided biclustering

The simultaneous clustering of genes based on i) co-expression over subsets of experimental conditions, ii) shared upstream non-coding DNA sequence patterns (possible *cis*-regulatory motifs), and iii) known gene and protein associations (Szklarczyk et al., 2011) was performed using version 4.9.21 of the R package cMonkey (Reiss et al., 2006). For biclustering using the cMonkey algorithm, the data were row-normalized as described in (Reiss et al., 2006). The bottom twenty-five percent least variant transcripts over all conditions were excluded from this analysis for computational and algorithmic efficiency. DNA sequences up to four hundred base pairs upstream of gene model start sites were iteratively searched using MEME (Bailey et al., 2009) for significantly occurring DNA sequence motifs in the putative promoters of co-expressed genes. The cMonkey algorithm was run for two thousand iterations to generate biclusters with a mean of thirty genes per cluster.

3. Results

3.1. Pan-transcriptomic clustering identifies separately coordinated groups of genes within singular condition-specific responses

Through integrative, whole-genome multi-experiment clustering (Fig. 1), numerous distinctly and significantly co-expressed sets of functionally-related transcripts were discovered in *T. pseudonana* and *P. tricornutum* (Table 1, Table S1). In addition, lists of transcripts that were significantly affected in individual previous experiments were segregated into separate and distinctly coordinated groups of genes. For example, transcripts whose expression increased in silica-limited *T. pseudonana* diatoms (Mock et al., 2008) represent at least three distinct transcriptional and physiological response mechanisms (Fig. S1). Three distinct clusters of co-expressed genes contain the top five transcripts (protein ids 268895, 9619, 21665, 3250, 10363) whose expression increased exclusively under silica limitation in previous experiments, in addition to several other silica-sensitive transcripts (Mock

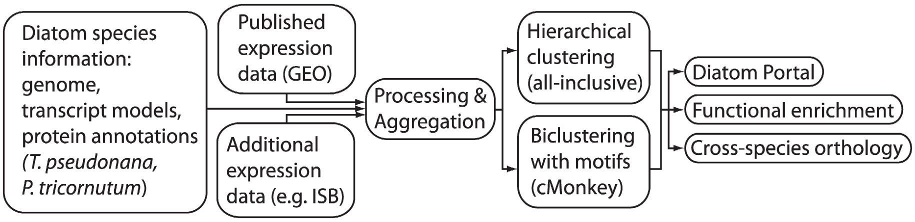
Diatom species information: genome, transcript models, protein annotations (T. pseudonana, P. tricornutum)
Published expression data (GEO)
Additional expression data (e.g. ISB)
Processing & Aggregation
Hierarchical clustering (all-inclusive)
Biclustering with motifs (cMonkey)
Diatom Portal
Functional enrichment
Cross-species orthology

Fig. 1. Pan-transcriptomic discovery of co-expressed functional modules in diatoms.

Please cite this article as: Ashworth, J., et al., Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms *Thalassiosira pseudonana* and ..., Mar. Genomics (2015), http://dx.doi.org/10.1016/j.margen.2015.10.011