Published online 23 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D1001–D1005***
doi:10.1093/nar/gkn905

Update of the Diatom EST Database: a new tool for digital transcriptomics

Uma Maheswari[1], Thomas Mock[2], E. Virginia Armbrust[2] and Chris Bowler[1,3,*]

[1]CNRS UMR8186, Department of Biology, Ecole Normale Supérieure, Paris, France, [2]School of Oceanography, University of Washington, Seattle, WA 98195, USA and [3]Stazione Zoologica 'Anton Dohrn', Villa Comunale, I-80121 Naples, Italy

Received September 15, 2008; Revised October 24, 2008; Accepted October 28, 2008

ABSTRACT

The Diatom Expressed Sequence Tag (EST) Database was constructed to provide integral access to ESTs from these ecologically and evolutionarily interesting microalgae. It has now been updated with 130 000 *Phaeodactylum tricornutum* ESTs from 16 cDNA libraries and 77 000 *Thalassiosira pseudonana* ESTs from seven libraries, derived from cells grown in different nutrient and stress regimes. The updated relational database incorporates results from statistical analyses such as log-likelihood ratios and hierarchical clustering, which help to identify differentially expressed genes under different conditions, and allow similarities in gene expression in different libraries to be investigated in a functional context. The database also incorporates links to the recently sequenced genomes of *P. tricornutum* and *T. pseudonana*, enabling an easy cross-talk between the expression pattern of diatom orthologs and the genome browsers. These improvements will facilitate exploration of diatom responses to conditions of ecological relevance and will aid gene function identification of diatom-specific genes and *in silico* gene prediction in this largely unexplored class of eukaryotes. The updated Diatom EST Database is available at http://www.biologie.ens.fr/diatomics/EST3.

INTRODUCTION

Diatoms are globally distributed, eukaryotic brown microalgae that participate in various biogeochemical cycles and play key roles in maintaining the ecological balance of the earth. They are major contributors to global primary production and CO_2 sequestration (1,2), and are also receiving attention as a potential source of biofuels (3). They fall within the heterokont branch of the eukaryotic tree (4) and are believed to have evolved from a secondary endosymbiotic process (5–7). The molecular and cellular biology of diatoms is dramatically underexplored. Previous Expressed Sequence Tag (EST) studies (8,9) together with the first whole genome sequences from diatoms, *Thalassiosira pseudonana* (10) and *Phaeodactylum tricornutum* (11), have shown that less than 50% of diatom genes can be assigned a putative function using homology-based methods, due to the lack of genomic information from well studied taxonomically related organisms. Similar observations were also made in a pilot study of ESTs derived from the polar diatom *Fragilariopsis cylindrus* grown at low temperature (12). Our earlier diatom EST database (9) enabled comparative studies of eukaryotic algal genomes and revealed some interesting differences in genes involved in basic cell metabolism (13,14). It also aided the study of key signalling and regulatory pathways (15), silica metabolism (16,17), nitrogen metabolism (18) and carbohydrate metabolism (19).

Furthermore, elucidation of the functions of diatom-specific genes can be facilitated by identifying conditions in which they are expressed. Non normalized EST libraries made from cells grown in different growth conditions can therefore provide a good dataset for comparative, functional as well as phylogenetic studies. For example, comparative study of the mRNAs expressed under different conditions can provide a systematic exploration of the molecular adaptations of a cell by differential gene expression. As a case in point, EST collections derived from cells grown under different conditions have proven to be a good tool for transcriptomics studies and genome annotation in the green alga *Chlamydomonas reinhardtii* (20–24). By comparing the expression profiles from more than one growth condition, differential gene expression studies can

*To whom correspondence should be addressed. Tel: +33 1 44323525; Fax: +33 1 44323935; Email: cbowler@biologie.ens.fr
Present address:
Thomas Mock, School of Environmental Sciences, University of East Anglia, Norwich, UK.

therefore provide a useful means to explore diatom gene function and genome annotation.

In this update we describe EST collections derived from diatom cells grown under different conditions and statistical methods used to explore gene expression. This digital gene expression database contains more than 200 000 ESTs from the two recently sequenced diatom genomes, *T. pseudonana* (10) and *P. tricornutum* (11). *T. pseudonana* is a centric diatom and has been a model organism for physiological studies of widely distributed species belonging to the order Thalassiosirales. *P. tricornutum* is a pennate diatom for which a range of reverse genetics tools have been generated (25), therefore making it a good model for functional genomic studies. The sequenced diatoms revealed many interesting features of diatom genes and metabolic pathways, although comparative studies also revealed a high level of molecular divergence (11,15). Bearing in mind these striking differences, the updates in the Diatom EST Database described here provide key insights into differential gene expression in diatoms grown in a range of ecologically relevant conditions.

DATA SOURCES AND DATABASE CONSTRUCTION

The Diatom EST Database was initially made with 12 136 ESTs from *P. tricornutum* and 15 174 ESTs from *T. pseudonana*, each obtained from a single growth condition (9). These libraries were expanded with 120 411 ESTs from *P. tricornutum* and 61 913 ESTs from *T. pseudonana* obtained from cells grown in 15 and 6 additional growth conditions, respectively. The new sets of ESTs were subjected to preliminary analysis such as vector clipping, quality control, etc. (9) and sequence assembly and redundancy checking was then done in two steps. First, the ESTs were clustered together with the predicted gene models from their respective genomes (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html and http://genome.jgi-psf.org/Phatr2/Phatr2.home.html). We were able to assign 120 575 ESTs to 8944 of the 10 402 gene models in *P. tricornutum* and 43 114 ESTs to 7268 of the 11 776 gene models in *T. pseudonana* using the BLASTN programme (cut-off *e*-value 10^{-10}) (26). These 8944 and 7268 transcriptional units (TUs) with predicted gene models were directly added to the non-redundant transcript sets with new sequence identifiers containing 'G' as a prefix along with the gene model identifier, e.g., G10065 for gene model 10065. The number of ESTs clustering to each gene model gives the redundancy or cluster size of the transcript. Secondly, transcripts which did not have a predicted gene model (11 513 ESTs from *P. tricornutum* and 18 073 ESTs from *T. pseudonana*), mainly due to the fact that ESTs from only a few libraries were used for training the gene prediction programmes (11), were subjected to analysis by CAP3 (27). Sequences with greater than 95% identity over a region longer than 30 base pairs were clustered using this programme and we thus obtained 1330 contigs and 2096 singletons for *P. tricornutum* and 1769 contigs and 2039 singletons for *T. pseudonana*. These were added to the non-redundant transcript set with sequence identifiers starting with 'C' for contigs and 'S' for the singletons. Adding the TUs with gene models to the contigs and singletons obtained from CAP3, we counted 12 370 non-redundant TUs in *P. tricornutum* and 11 076 TUs in *T. pseudonana*. Among the non-redundant TUs which do not have a predicted gene model, we found only 612 TUs in *P. tricornutum* and 1 083 TUs in *T. pseudonana* that do not align in their respective genomes, likely because of remaining gaps in the genome sequences.

The contribution of ESTs from different libraries to the cluster size of each TU gives the abundance of each expressed transcript across different libraries. The counts were normalized to the library size by converting the counts to frequencies, which allows a statistical comparison to be made of expression levels of transcripts in different conditions. Specifically, the log-likelihood ratio was calculated for each contig (28) to statistically validate whether a difference in frequency across different libraries was random or due to differential expression. The database schematized in Figure 1 provides access to frequency distribution plots (Figure 1E) and log-likelihood ratios (*R*-values) for each TU, which are catalogued by library (Figure 1C) as well as across libraries (Figure 1D and H). Figure 1E shows an example of a TU with high R-value (i.e. a gene that is strongly differentially expressed in the conditions tested). By cataloguing the TUs based on their R-values we were then able to identify transcripts that are differentially expressed under each condition. For example, transcripts expressed during iron limitation served as a useful starting point to explore the molecular response of *P. tricornutum* to life at low iron concentrations (29), providing experimental validation of our statistical methods. TUs were also subjected to hierarchical clustering (30) to identify transcripts with similar expression profiles in the different conditions. These analyses together with relevant functional information were visualized using Java Treeview (31). Figure 1F shows a screen shot of hierarchical clustering (30) of *P. tricornutum* contigs.

The updated dataset and the accompanying results are stored in upgraded servers with the Linux Debian 'etch' platform in DELL1850 hosting the relational database PostgreSQL 8.3 and DELL1855 with the web server Apache 2.0 and PHP 5. The relational database was migrated to postgreSQL for faster access and to enable the dynamic clustering of expression data. The new web interface is also linked to the gene models on the JGI diatom genome browsers (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html and http://genome.jgi-psf.org/Phatr2/Phatr2.home.html), which enables the user to have direct access to annotation and gene structure for each TU (Figure 1G).

DATABASE CONTENTS AND WEB INTERFACE

The database provides access to details of each cDNA library and corresponding growth conditions (Figure 1A). The raw sequences are catalogued by library and each raw sequence table gives access to DNA sequence, length and BLAST output. These tables also

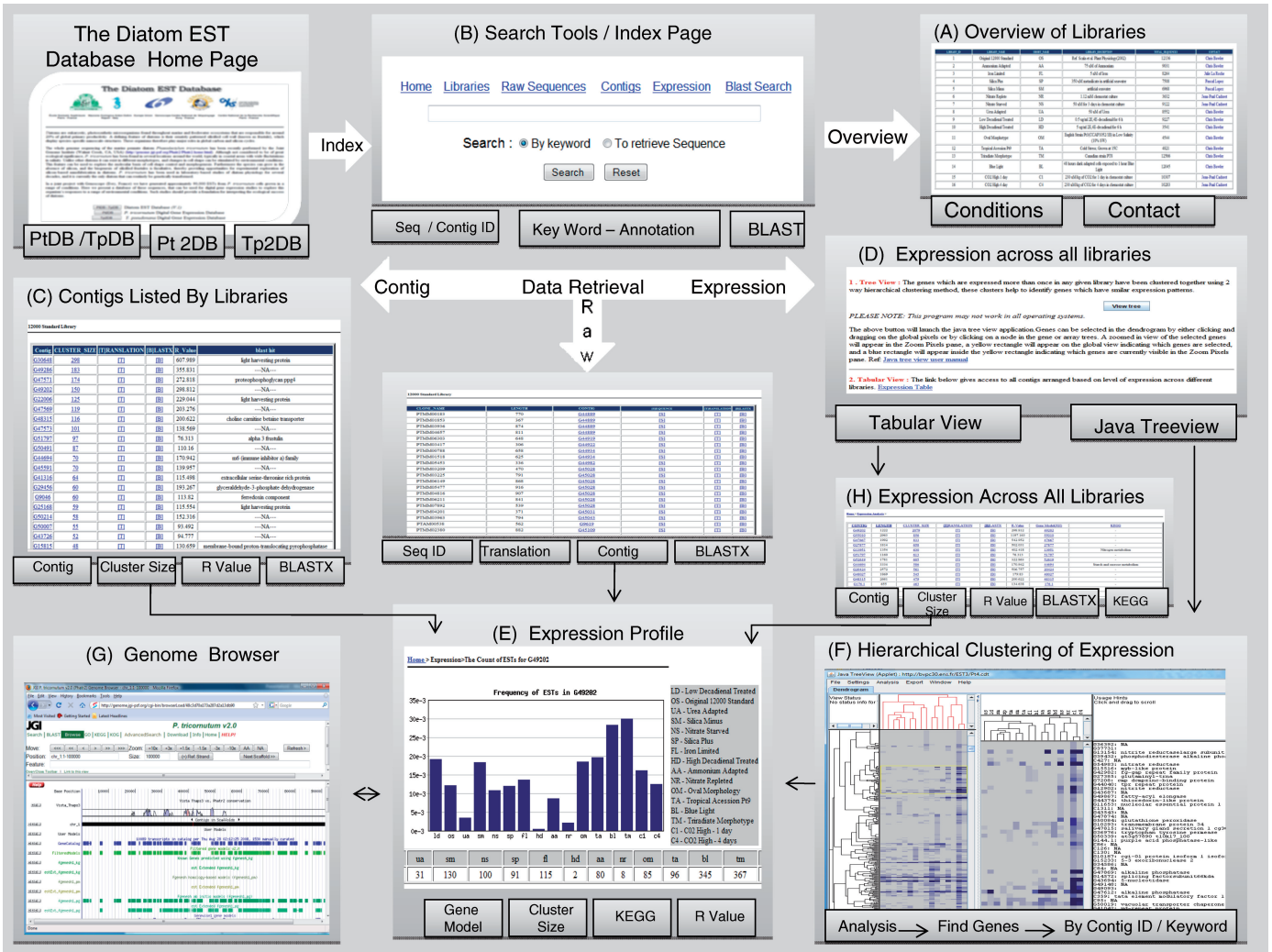

Figure 1. Overview of the updated Diatom EST Database.

provide links to the TU that each sequence belongs to. The contig tables give access to the TU of each library, catalogued based on the abundance of ESTs in each condition (Figure 1C). The cluster size of each TU is linked to the dynamically generated frequency plot (Figure 1E), which enables comparison of expression levels in the other libraries. This table also shows *R*-values and the best BLAST results.

The expression of each TU across all the libraries can be accessed by two different methods (Figure 1D), either in tabular form (Figure 1H) or as a hierarchical cluster visualized using Java Treeview (Figure 1F). The tabular view gives access to all TUs expressed more than once in any given condition and they are catalogued based on cluster size, which is again linked to each frequency plot (Figure 1E). This table also provides a link to the ortholog if present in the other diatom and its expression profile, as well as the corresponding gene models hyperlinked to the genome databases hosted at JGI, providing access to further functional annotation and visualization of neighbouring genes (Figure 1G). The Java Treeview visualizes the two-way hierarchical clustering of all the transcripts which are expressed more than once, helping to identify libraries that cluster together and transcripts with similar expression patterns. The annotations for each TU are hyperlinked to the frequency plots and to the JGI genome browsers.

The new web interface is inspired by Google, having a simplified, self-explanatory look and easy retrieval of data. The database is queryable by keyword, based on annotation from homology search methods and the TU identifier, and sequence retrieval is possible by using either the sequence identifier or TU identifier. Homology searches, using BLAST against each library and the total non-redundant sets are also available via the web interface.

FUTURE DIRECTIONS

The diatom genomic repository is rapidly expanding with several sequencing projects. For example, the genomes of two additional pennate diatoms, *Pseudo-nitzschia multiseries* and *F. cylindrus*, are currently nearing completion at JGI, together with accompanying EST collections.

The database analysis and pipeline described here are semi-automated and can easily incorporate these and other data sets from diatoms and related species. Pilot microarray projects in *T. pseudonana* and *P. tricornutum* have already provided experimental validation for this EST-based digital transcriptomics database under some conditions (29,32) and possibilities to link microarray based studies to the existing database are currently being explored, as is the incorporation of transcriptomics data from massively parallel sequencing platforms. Reverse genetics studies are providing additional experimental validation for the expression, localization and functions of individual TUs (33) and so information derived from the database can also be used to train the gene prediction programmes to improve *in silico* gene annotation in diatoms and related organisms.

AVAILABILITY

The Diatom EST database is freely available on the web at http://www.biologie.ens.fr/diatomics/EST3. The *P. tricornutum* ESTs have been submitted to the NCBI dbEST (Genbank accession numbers CD374840–CD384835, BI306757–BI307753, CD374840–CD384835, BI306757–BI307753, CT868744–CT950687 and CU69 5349–CU740080). Requests for bulk queries of the expression data and to house EST data from other diatoms can be addressed to Dr Chris Bowler.

ACKNOWLEDGEMENTS

P. tricornutum ESTs were generated and sequenced by Genoscope (Evry, Paris) and the *T. pseudonana* ESTs were sequenced by JGI, USA. We are grateful to Pierre Vincens and Jean-Pierre Roux for managing the server and the software, to Andrew Allen and Kamel Jabbari for their help and suggestions, Igor V. Grigoriev and Alan Kuo at JGI for providing links to gene models, and to Alok J. Saldanha for his help to integrate the Java Treeview in the database. *T. pseudonana* cultures were grown with the help of Karie Holtermann. The contact information of the people responsible for each *P.tricornutum* library can be obtained from the database.

FUNDING

Partial funding for the Diatom EST Database was obtained from the EU-funded Diatomics (LSHG-CT-2004-512035) and Marine Genomics Europe projects (GOCE-CT-2004-505403) and the Agence Nationale de la Recherche. *P. tricornutum* ESTs were funded by Genoscope (Evry, Paris). Generation of *T. pseudonana* ESTs was funded by a Gordon and Betty Moore Foundation Marine Microbiology Investigator Award (EVA). Funding for open access charge: Centre National de la Recherche Scientifique.

Conflict of interest statement. None declared.

REFERENCES

1. Field,C.B., Behrenfeld,M.J., Randerson,J.T. and Falkowski,P. (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, **281**, 237–240.
2. Irigoien,X., Huisman,J. and Harris,R.P. (2004) Global biodiversity patterns of marine phytoplankton and zooplankton. *Nature*, **429**, 863–867.
3. Chisti,Y. (2008) Biodiesel from microalgae beats bioethanol. *Trends Biotechnol.*, **26**, 126–131.
4. Baldauf,S.L. (2008) An overview of the phylogeny and diversity of eukaryotes. *J. Syst. Evol.*, **46**, 263–273.
5. Yoon,H.S., Hackett,J.D., Ciniglia,C., Pinto,G. and Bhattacharya,D. (2004) A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.*, **21**, 809–818.
6. Patron,N.J., Rogers,M.B. and Keeling,P.J. (2004) Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. *Eukaryot. Cell*, **3**, 1169–1175.
7. Li,S., Nosenko,T., Hackett,J.D. and Bhattacharya,D. (2006) Phylogenomic analysis identifies red algal genes of endosymbiotic origin in the chromalveolates. *Mol. Biol. Evol.*, **23**, 663–674.
8. Scala,S., Carels,N., Falciatore,A., Chiusano,M.L. and Bowler,C. (2002) Genome properties of the diatom *Phaeodactylum tricornutum*. *Plant Physiol.*, **129**, 993–1002.
9. Maheswari,U., Montsant,A., Goll,J., Krishnaswamy,S., Rajyashri,K.R., Patell,V.M. and Bowler,C. (2005) The Diatom EST Database. *Nucleic Acids Res.*, **33**, D344–D347.
10. Armbrust,E.V., Berges,J.B., Bowler,C., Green,B.R., Martinez,D., Putnam,N.H., Zhou,S., Allen,A.E., Apt,K.E., Bechner,M. *et al.* (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science*, **306**, 79–86.
11. Bowler,C., Allen,A.E., Badger,J.H., Grimwood,J., Jabbari,K., Kuo,A., Maheswari,U., Martens,C., Maumus,F., Otillar,R.P. *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–244.
12. Mock,T., Krell,A., Glockner,G., Kolukisaoglu,U. and Valentin,K. (2006) Analysis of expressed sequence tags (ESTs) from the polar diatom *Fragilariopsis cylindrus*. *J. Phycol.*, **42**, 78–85.
13. Montsant,A., Jabbari,K., Maheswari,U. and Bowler,C. (2005) Comparative genomics of the pennate diatom *Phaeodactylum tricornutum*. *Plant Physiol.*, **137**, 500–513.
14. Herve,C., Tonon,T., Collen,J., Corre,E. and Boyen,C. (2006) NADPH oxidases in Eukaryotes: red algae provide new hints! *Curr. Genet.*, **49**, 190–204.
15. Montsant,A., Allen,A.E., Coesel,S., De Martino,A., Falciatore,A., Mangogna,M., Siaut,M., Heijde,M., Jabbari,K., Maheswari,U. *et al.* (2007) Identification and comparative genomic analysis of signaling and regulatory components in the diatom Thalassiosira pseudonana. *J. Phycol.*, **43**, 585–604.
16. Montsant,A., Maheswari,U., Bowler,C. and Lopez,P.J. (2005) Diatomics: Toward diatom functional genomics. *J. Nanosci. Nanotechnol.*, **5**, 5–14.
17. Lopez,P.J., Descles,J., Allen,A.E. and Bowler,C. (2005) Prospects in diatom research. *Curr. Opin. Biotechnol.*, **16**, 180–186.
18. Allen,A.E., Vardi,A. and Bowler,C. (2006) An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr. Opin. Plant Biol.*, **9**, 264–273.
19. Kroth,P.G., Chiovitti,A., Gruber,A., Martin-Jezequel,V., Mock,T., Parker,M.S., Stanley,M.S., Kaplan,A., Caron,L., Weber,T. *et al.* (2008) A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PLoS ONE*, **3**, e1426.
20. Asamizu,E., Nakamura,Y., Miura,K., Fukuzawa,H., Fujiwara,S., Hirono,M., Iwamoto,K., Matsuda,Y., Minagawa,J., Shimogawara,K. *et al.* (2004) Establishment of publicly available cDNA material and information resource of Chlamydomonas reinhardtii (Chlorophyta) to facilitate gene function analysis. *Phycologia*, **43**, 722–726.
21. Fukuzawa,H. (2007) Genome and transcriptome analyses of Chlamydomonas reinhardtii: Isolation and functional analyses of genes for carbon-concentrating mechanism. *Genes & Genetic Systems*, **82**, 514.

22. Miura,K., Yamano,T., Yoshioka,S., Kohinata,T., Inoue,Y., Taniguchi,F., Asamizu,E., Nakamura,Y., Tabata,S., Yamato,K.T. *et al.* (2004) Expression profiling-based identification of CO_2-responsive genes regulated by CCM1 controlling a carbon-concentrating mechanism in Chlamydomonas reinhardtii. *Plant Physiol.*, **135**, 1595–1607.
23. Jain,M., Shrager,J., Harris,E.H., Halbrook,R., Grossman,A.R., Hauser,C. and Vallon,O. (2007) EST assembly supported by a draft genome sequence: an analysis of the Chlamydomonas reinhardtii transcriptome. *Nucleic Acids Res.*, **35**, 2074–2083.
24. Liang,C., Liu,Y.S., Liu,L., Davis,A.C., Shen,Y.J. and Li,Q.S.Q. (2008) Expressed sequence tags with cDNA termini: Previously overlooked resources for gene annotation and transcriptome exploration in Chlamydomonas reinhardtii. *Genetics*, **179**, 83–93.
25. Siaut,M., Heijde,M., Mangogna,M., Montsant,A., Coesel,S., Allen,A., Manfredonia,A., Falciatore,A. and Bowler,C. (2007) Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene*, **406**, 23–35.
26. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Evol.*, **215**, 403–410.
27. Huang,X. and Madan,A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
28. Stekel,D.J., Git,Y. and Falciani,F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Res.*, **10**, 2055–2061.
29. Allen,A.E., Laroche,J., Maheswari,U., Lommer,M., Schauer,N., Lopez,P.J., Finazzi,G., Fernie,A.R. and Bowler,C. (2008) Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation. *Proc. Natl Acad. Sci. USA*, **105**, 10438–10443.
30. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
31. Saldanha,A.J. (2004) Java Treeview-extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.
32. Mock,T., Samanta,M.P., Iverson,V., Berthiaume,C., Robison,M., Holtermann,K., Durkin,C., BonDurant,S.S., Richmond,K., Rodesch,M. *et al.* (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.
33. Vardi,A., Bidie,K.D., Kwityn,C., Hirsh,D.J., Thompson,S.M., Callow,J.A., Falkowski,P. and Bowler,C. (2008) A diatom gene regulating nitric-oxide signaling and susceptibility to diatom-derived aldehydes. *Curr. Biol.*, **18**, 895–899.