SEE COMMENTARY

Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses

Thomas Mock*, Manoj Pratim Samanta†, Vaughn Iverson*, Chris Berthiaume*, Matthew Robison‡, Karie Holtermann*, Colleen Durkin*, Sandra Splinter BonDurant‡, Kathryn Richmond‡, Matthew Rodesch‡, Toivo Kallas§, Edward L. Huttlin¶, Francesco Cerrina‡‖, Michael R. Sussman‡¶**, and E. Virginia Armbrust*,**

*School of Oceanography, University of Washington, Box 357940, Seattle, WA 98195; †Systemix Institute, Los Altos, CA 94024; ‡Biotechnology Center, University of Wisconsin, Madison, WI 53706; §Department of Biology and Microbiology, University of Wisconsin, Oshkosh, WI 54901; and Departments of ¶Biochemistry and ‖Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved December 3, 2007 (received for review August 22, 2007)

Formation of complex inorganic structures is widespread in nature. Diatoms create intricately patterned cell walls of inorganic silicon that are a biomimetic model for design and generation of three-dimensional silica nanostructures. To date, only relatively simple silica structures can be generated *in vitro* through manipulation of known diatom phosphoproteins (silaffins) and long-chain polyamines. Here, we report the use of genome-wide transcriptome analyses of the marine diatom *Thalassiosira pseudonana* to identify additional candidate gene products involved in the biological manipulation of silicon. Whole-genome oligonucleotide tiling arrays and tandem mass spectrometry identified transcripts for >8,000 genes, ≈3,000 of which were not previously described and included noncoding and antisense RNAs. Gene-specific expression profiles detected a set of 75 genes induced only under low concentrations of silicon but not under low concentrations of nitrogen or iron, alkaline pH, or low temperatures. Most of these induced gene products were predicted to contain secretory signals and/or transmembrane domains but displayed no homology to known proteins. Over half of these genes were newly discovered, identified only through the use of tiling arrays. Unexpectedly, a common set of 84 genes were induced by both silicon and iron limitations, suggesting that biological manipulation of silicon may share pathways in common with iron or, alternatively, that iron may serve as a required cofactor for silicon processes. These results provide insights into the transcriptional and translational basis for the biological generation of elaborate silicon nanostructures by these ecologically important microbes.

silica | transcriptome | iron | nitrogen | temperature

Marine diatoms are unicellular eukaryotic algae that generate ≈20% of the ≈100 billion metric tons of organic carbon produced through photosynthesis on Earth each year (1, 2). A distinctive feature of diatoms is their requirement for the element silicon, which they use to build cell walls composed of amorphous, hydrated silicon dioxide (silica) embedded with small amounts of organic material (3, 4). The silica-based patterns of nano- and micrometer-sized pores, spines, and other cell wall structures of diatoms are so detailed and precisely replicated that taxonomists use these features to distinguish between the estimated 10^5 to 10^6 species.

Silica patterning in diatoms is hypothesized to depend on both self-assembly processes and controlled silica polymerization (5–9), with the precipitating silica further "molded" by cytoskeletal interactions (7). Long-chain polyamines and phosphoproteins known as silaffins are the only diatom molecules thus far shown to have a direct impact on silica precipitation *in vitro*, with the resulting pore sizes of the formed structures determined by relative proportions of polyamines and silaffins (5–9). The diatom silaffins display no homology to silcateins, which initiate silica precipitation in sponges (10). Diatom transporters have been sequenced and characterized and have been shown to interact directly with silicon to actively transport silicic acid against a large concentration gradient, although the mechanism for intracellular storage of soluble silicic acid is not known (11, 12). These diatom transporters display no sequence homology to the silicon transporters recently identified in rice (13). Despite availability of whole-genome sequence for two marine diatoms (10) a molecular basis for the elaborate species-specific silica structures has remained unclear, largely because of this lack of homology to proteins involved in silicon manipulation in other organisms. Here, we describe pathways and gene products involved in silica processing in diatoms identified through the use of whole-genome expression profiling.

Results and Discussion

Whole-Genome Expression Profiling Identifies Previously Undescribed Diatom Genes. The lack of homology between proteins required for biogenic silica manipulation in different organisms has complicated identification of underlying pathways in diatoms through traditional *in silico* homology-based approaches. We instead used tiling array-based whole-genome expression profiles (14, 15) to discover potentially unreported, essential genes because these methods are not restricted by *a priori* assumptions of gene structure or location. Probes were tiled across both strands of the ≈34-megabase *Thalassiosira pseudonana* genome (10), and the resulting arrays were hybridized with RNA extracted from cells grown either in nutrient complete media or in growth-limiting concentrations of silicon (the biologically available form of silicon is silicic acid) [supporting information (SI) Fig. 5 and SI Table 1]. The tiling array data validated transcription of ≈41% (4, 653) of the 11,390 computationally predicted genes (SI Table 2). An additional 1,132 transcripts were identified that did not correspond to modeled genes (10) with few of these transcripts (<17%) predicted to encode proteins with homology (e-value $<10^{-5}$) to publicly available proteins. These previously unidentified transcripts are operationally referred to here as "unpredicted" transcriptional units (unpre-

Author contributions: T.M., M. Robison, and K.H. designed research; T.M., M.P.S., M. Robison, K.H., C.D., S.S.B., K.R., M. Rodesch, T.K., and E.L.H. performed research; M.P.S., V.I., and F.C. contributed new reagents/analytic tools; T.M., M.P.S., V.I., C.B., M. Robison, C.D., M. Rodesch, T.K., E.L.H., M.R.S., and E.V.A. analyzed data; and T.M., M.P.S., M.R.S., and E.V.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE9697).

See Commentary on page 1391.

**To whom correspondence may be addressed. E-mail: msussman@wisc.edu or armbrust@ocean.washington.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0707946105/DC1.

© 2008 by The National Academy of Sciences of the USA

ENVIRONMENTAL SCIENCES

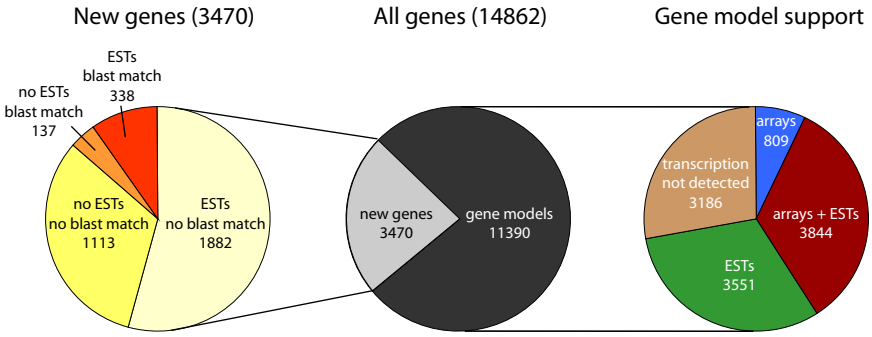


Fig. 1. Transcriptome statistics based on microarray and EST data for *T. pseudonana*. Microarrays (tiling and gene-specific) and ESTs provided transcriptional support for 8,204 gene models; 3,470 previously unreported genes were predicted.

dicted TUs) (SI Table 3) and have an average length of 1,549 bp, comparable with the average length of the computationally derived genes. Importantly, one unpredicted TU encoded a putative protein that possesses a signal peptide and displays homology to the proline-rich domain of the N terminus of tpSil1, a known silaffin required for silica precipitation (12) (SI Fig. 6). Transcription was also detected for a second putative protein carrying a domain with homology to the proline-rich domains of silaffins tpSil1 and tpSil2 [Joint Genome Institute (JGI) protein ID 9558] computationally predicted upon closure of a sequence gap (SI Fig. 6). These results suggested that additional genes required for silicon manipulation could be embedded within the previously undiscovered transcripts.

A comprehensive gene-specific expression array that occupied a single glass slide was designed to provide additional support for the unpredicted TUs discovered with the tiling arrays and to further refine which genes were specifically involved in silica biomanipulation by examining four additional growth-limiting conditions. Probes were selected that corresponded to all *in silico* predicted genes, all available EST sequences, and the previously unidentified unpredicted TUs. To generate the most extensive gene set possible, probes were also selected from nearly 16,000 additional tiling array probe clusters longer than 100 nt with hybridization signals above background but not originally defined as an unpredicted TU. The gene-specific expression array was hybridized with RNA isolated from at least four biological replicates each of cells grown under six different conditions: nutrient complete; growth limiting concentrations of nitrogen, iron, or silicon; alkaline pH of 9.4 (reduces dissolved CO_2 concentrations); and a shift to lowered temperature (SI Fig. 5 and SI Table 1).

Further support for the transcriptional data was provided by tandem mass spectrometry (MS/MS) analysis of soluble, membrane and cell-wall protein fractions of *T. pseudonana* cells grown under nutrient-complete and silicic acid-limited conditions. The 682 detected peptides were searched against a six-frame translation of the *T. pseudonana* genome, and the peptides were found to map to 349 distinct proteins (SI Table 4), with identification of 16 peptides that extended the length of computationally derived gene models.

Overall, combined analysis of tiling array, EST, gene-specific expression array, and proteomics data provided support for transcription of ≈70% of the *in silico* predicted genes (10) (SI Table 5). Those genes not supported by our studies are likely expressed under growth conditions not examined here (e.g., phosphate limitation). Significantly, the combined data provided further support that the unpredicted TUs detected via the tiling arrays corresponded to previously unidentified genes. Overall, ≈3,470 genes were identified with this combined approach, which increases by ≈30% the total number of genes (14, 860) predicted for *T. pseudonana* (Fig. 1 and SI Table 5).

Specific Induction of Genes in Response to Silicon Limitation. Seven hundred nine genes were differentially expressed by >2-fold [Bayesian t test (16, 17) $P < 0.001$] under at least one growth-limiting condition relative to nutrient-replete conditions (SI Table 6). Independent quantitative reverse-transcriptase PCR (qRT-PCR) experiments on 16 genes validated these results (SI Table 7). Included among the subset of differentially expressed genes were 51 of the genes discovered via the tiling arrays and 73 genes that encoded proteins with similarity (e-value $<10^{-5}$) only to hypothetical proteins from *Phaeodactylum tricornutum*, a second diatom for which whole-genome sequence is available (www.jgi.doe.gov). Thus, almost 20% of the differentially expressed genes detected here have so far been identified only in diatoms.

Hierarchical clustering of the differentially expressed genes identified condition-specific gene clusters and similarities in genome-wide expression between the five limiting conditions relative to control conditions (Fig. 2). A majority of the down-regulated genes encoded proteins related to photosynthetic processes such as light harvesting and electron transport (SI Table 6). Silicon limitation specifically up-regulated 75 genes, including the most highly up-regulated genes in the entire dataset (SI Table 6). Peptide support was detected for translation of 5 transcripts up-regulated and 59 transcripts down-regulated under silicon limitation (SI Table 4).

The majority of the 75 genes induced by silicon limitation encode proteins without predicted functions. For example, 24 of the 30 most highly induced transcripts encode proteins with low amino acid complexity (enriched for S, T, K, R, or P) whose only obvious features were that approximately half of them possess a secretory signal sequence and/or at least one transmembrane-spanning domain (β-sheet and/or α-helix). One highly induced transcript corresponded to a small ORF (150 bp), suggesting that this transcript might encode a small peptide or noncoding regulatory RNA. A surprising feature of the 75 transcripts induced by silicon limitation is the presence of a vertebrate-like consensus motif CANC*AUG* (18) at positions −4 to −1 upstream of predicted initiator codons (SI Table 8). Cytosine was present at positions −1 and −4 in 62% and 59% of the silicon-limitation induced genes, respectively, vs. 44% and 40% for all genes. A greater preference for adenine at position −3 (77% of the silicon-limitation-induced genes vs. 61% for all genes, SI Table 8) has been associated with regulation of translation initiation (19). The results obtained with the transcripts induced by silicon limitation are in direct contrast with transcripts differentially expressed under other limitations; transcripts induced by nitrogen limitation, for example, display no distinctive motifs upstream of the initiator codon, and a majority encode proteins with a predicted function (SI Table 9).

The microarray data were further examined for evidence of additional regulatory mechanisms associated with silicon biomanipulation. Antisense expression, a commonly detected means of posttranscriptional regulation (20–22) was identified for 385 genes, each of which was longer than 300 nt (SI Table 10). Among these were 10 genes with antisense signals that were reduced by >2-fold

SEE COMMENTARY

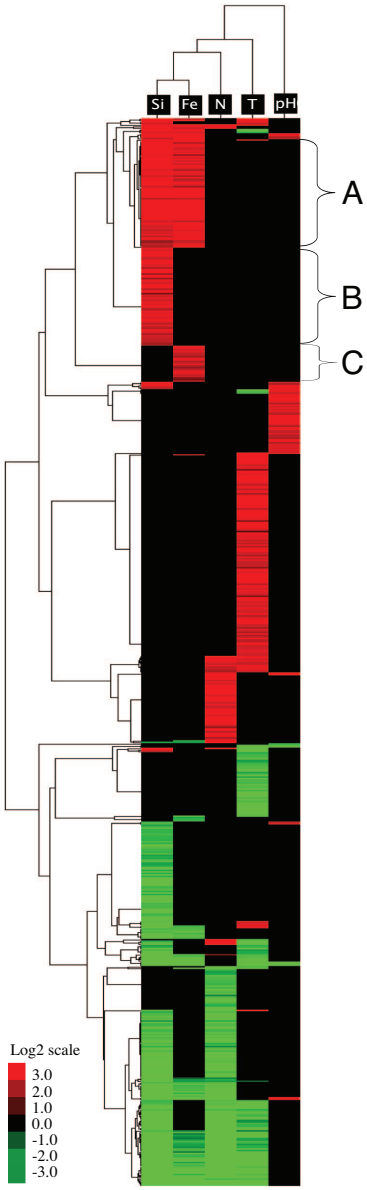


Fig. 2. Hierarchical cluster display of 709 genes that are differentially expressed (Bayesian *t* test $P < 0.001$, ≥2-fold difference in mRNA levels) under silicon (Si)-, iron (Fe)-, nitrogen (N)-, or temperature (T)-limitation, or alkaline pH (pH) relative to nutrient-replete growth. Each limitation corresponds to a single column and each gene to a single row. The color chart indicates fold change of expression by using a base 2-logarithmic scale. The color scale ranges from saturated red (log2 ratios of 3.0 and above) for up-regulated genes to saturated green (log2 ratios of −3.0 and below) for down-regulated genes; black indicates no significance. Silicon and iron limitation resulted in a cluster of coregulated genes (Cluster A) and genes up-regulated only under silicon limitation (Cluster B) and genes up-regulated only under iron limitation (Cluster C).

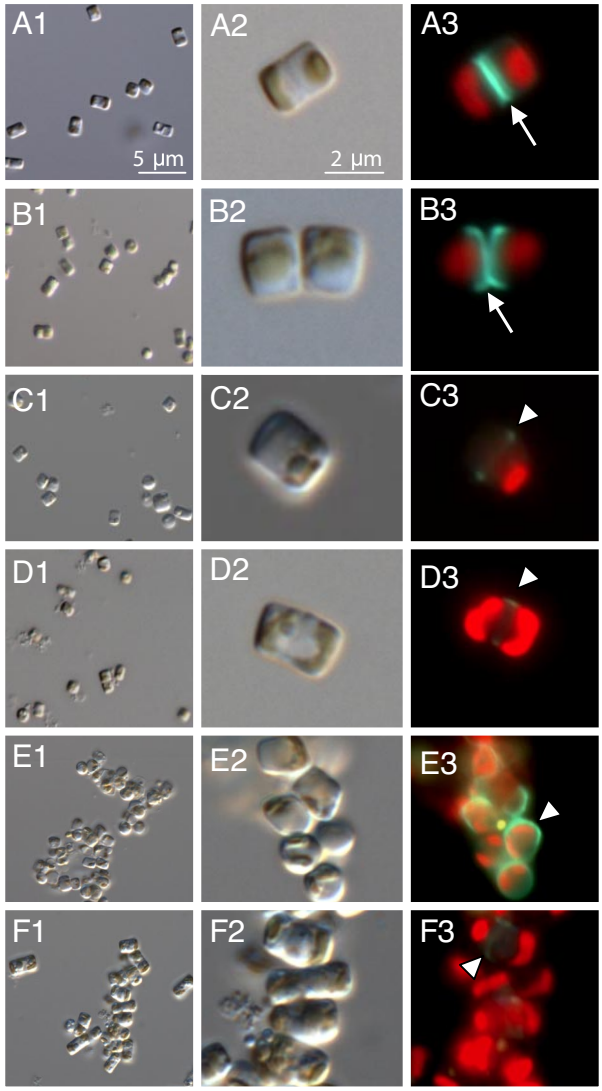
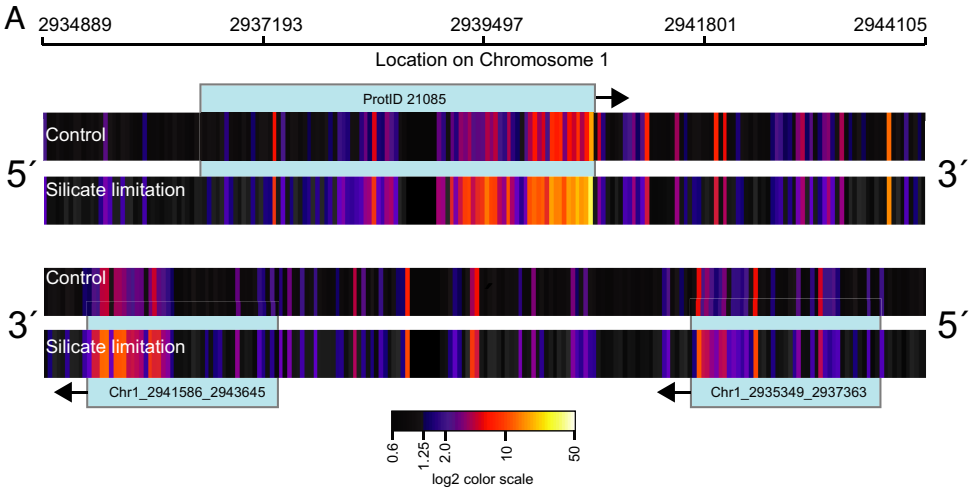


Fig. 3. Light and epifluorescent micrographs of *T. pseudonana* cells grown under nutrient-replete conditions (*A*), exposed to low temperature (*B*), growth-limiting concentrations of nitrate (*C*), alkaline pH (*D*), and growth-limiting concentrations of iron (*E*) or silicon (*F*). Cells are shown at low magnification (*Left*), and the same field of cells is shown at higher magnification and visualized under either light (*Center*) or epifluorescent (*Right*) microscopy. Red fluorescence is due to chlorophyll *a* fluorescence, and blue fluorescence results from exposure of live cells to the pH-sensitive fluorescent dye PDMPO (2-(4-pyridyl)-5{[4-dimethylaminoethyl-aminocarbamoyl)-methoxy]phenyl}oxazole) for 12 h to stain freshly deposited silica. Silica deposition at the valves is indicated by the arrow and at the girdle bands by arrowheads.

under silicon limitation, suggesting a possible regulatory role in silicon processes. A genome-wide search for possible long noncoding RNAs using conservative search criteria identified a noncoding transcript located on chromosome 1 (strand (+) position 2935331–2936177) that was up-regulated 2-fold under silicon limitation, which would be the first noncoding RNA identified associated with silicon processes.

Despite the large number of transcripts induced by silicon limitation that encode proteins with unknown functions, a number of differentially expressed genes were also detected that encode proteins reported to be involved in silicon transport and precipitation (11, 23, 24). For example, silicon limitation alone strongly (9- and 142-fold, respectively) induced transcription of two genes that encode silicon transporters (SIT1 and SIT2), with their corresponding peptides detected by mass spectrometry (SI Table 4); as described (11), the gene encoding a third silicon transporter (SIT3) was expressed at low levels under all conditions, with no evidence of differential expression. Silicon limitation specifically down-regulated genes encoding SIL3, a member of the silaffin family (12), and the putative silaffin-like protein (JGI protein ID 9558) detected with the tiling arrays (SI Fig. 6). Both results suggested that less silica should be deposited when growth is limited by silicon avail-

ENVIRONMENTAL SCIENCES



B

		Gene-Specific Arrays				Q-PCR			
Identifier	Annotation	-Si FC	p-value	-Fe FC	p-value	-Si FC	p-value	-Fe FC	p-value
Chr1_2941586_2943645	Secreted protein (signalP), leucine zipper pattern, 3xN-glycosylation and 2xN-myristoylation sites	5.9	2.2×10^{-4}	-	n.s.	3.5	3.5×10^{-2}	-4	9.9×10^{-2}
ProtID 21085 2936573 - 2940520	Conserved hypothetical protein with transmembrane spanning domain	20	9.8×10^{-10}	24	2.6×10^{-9}	26	1.0×10^{-4}	6.4	4.6×10^{-2}
Chr1_2935349_2937363	Unknown protein with transmembrane spanning domain	64	7.9×10^{-12}	45	2.0×10^{-10}	9.5	3.3×10^{-2}	3.6	4.3×10^{-2}

Fig. 4. A physical cluster composed of unknown genes potentially involved in cell-wall and silicon processes (*A*) and their expression (*B*). (*A*) Heat map of both DNA strands. Vertical bars represent tiled oligonucleotide probes (36-mer) for control and silicon limitation, with the color range indicating expression intensity on a $\log_2$ scale. Light-blue boxes indicate the putative location of genes, and black arrows indicate the direction of transcription. (*B*) Putative functional characteristics of proteins encoded by physical cluster genes and their expression based on gene-specific microarrays and quantitative (Q)RT-PCR (Q-PCR) Gene expression is given as fold change (FC) under silicon and iron limitation relative to the control growth condition. Gene-specific arrays: $n = 4$; Q-PCR: $n = 3$. *N* represents the number of biological replicates.

ability. The reduced amount of silica deposited under these conditions was readily detected by staining cells with the pH-sensitive fluorescent dye PDMPO (2-(4-pyridyl)-5{[4-dimethylaminoethyl-aminocarbamoyl)-methoxy]phenyl}oxazole), which is incorporated into newly precipitated silica (Fig. 3). The greatly reduced, but still detectable, staining of silicon-limited cells contrasted with the strong staining of nutrient-replete cells or cells growing slowly, yet still dividing, under low temperature (Fig. 3).

Induction of Common Pathways by Silicon and Iron Limitation. A striking result of the hierarchical cluster analysis was the identification of a common set of 84 genes that were up-regulated by both iron and silicon limitation, but no other treatment (Cluster A Fig. 2). Together, these two treatments accounted for approximately one-fourth of all differentially expressed genes but almost two-thirds of the differentially expressed previously unidentified genes, further emphasizing the distinctive aspects of silicon manipulation in diatoms. Three genes within this cluster are localized to a small region that spans only ≈8.1 kb on chromosome 1 (Fig. 4). Gene-specific arrays and qRT-PCR analysis confirmed that each of these adjacent genes was up-regulated under silicon limitation and that two of the three genes were also induced by iron limitation. One of the three predicted gene products possesses a signal peptide and two of the three possess transmembrane spanning domains, features consistent with targeting to the extracellular matrix. The physically close association of these similarly expressed genes suggests regulatory mechanisms similar to a bacterial operon (25) and further suggests that the resulting gene products interact (26) to carry out silicon manipulations.

Unexpectedly, genes known to be required for silica deposition were influenced by availability of iron and silicon. The gene encoding SIL1, a member of the silaffin family (12), was up-regulated 3-fold only under iron limitation, and the distinctive PDMPO staining of the iron limited cells, compared with cells limited by the other nutrients (Fig. 3), supported a possible link between iron availability and silica deposition. Both iron and silicon limitation up-regulated genes that encoded enzymes required for polyamine biosynthesis, compounds also implicated in silica deposition (5). Two genes encoding cell wall (girdle band) associated proteins (27) were up-regulated under iron and silicon limitations and the corresponding peptides were detected with mass spectrometry. A recent proteomics study identified 10 proteins hypothesized to play a direct role in diatom silica precipitation and cell-wall formation (24). Six of the genes encoding these putative cell-wall proteins were up-regulated in our study only under silicon and iron limitations, further substantiating the link between iron and silicon and cell-wall processes. Finally, both iron and silicon limitation resulted in aberrant cell morphologies with a distinctive cell aggregation phenotype, suggesting that, in addition to potentially direct effects on silica deposition, there were also changes in the extracellular matrix of cells (Fig. 3). The potential for overlapping signal transduction pathways triggered by these two nutrients comes from the observation that genes encoding six protein kinases (JGI protein ID 32738, 37322, 19048, 263081, and 33772; unpredicted TU position: chr12_857573–859126) were specifically up-regulated only by both iron and silicon limitation. Phosphorylation of silaffins has been demonstrated *in vitro* to be necessary for silica precipitation (4), and consequently, a subset of the kinases identified here may be involved in these posttranslational events.

Conclusions

This study describes a large subset of genes and proteins potentially involved in formation of the nanopatterned diatom cell wall, a resilient, but relatively light-weight structure that provides protection (resistant to 100–700 tonnes per square meter) to cells that survive by remaining afloat in the near surface ocean (28). Most diatoms have an obligate requirement for silicon, and the evidence presented here suggests that silicon processes are tightly regulated through posttranscriptional and translation initiation mechanisms. Some of the most highly induced genes detected in this study were previously undiscovered genes specifically associated with silicon limitation. An important next step will be to determine which of these gene products influence *in vitro* assays of silica precipitation.

The most surprising result, however, was the tight coupling of pathways initiated by iron and silicon bioavailability. One explanation for this result is that iron may serve as a cofactor for silicon-specific gene products. A second possibility is that iron is directly incorporated into the silica cell wall in a regulated manner. Support for this latter explanation comes from limited data indicating that the relative proportion of iron found within the silica cell wall is higher in cells maintained at extremely low concentrations of bioavailable iron ([FeIII] $<10^{-10.5}$ M) than in cells grown under moderately limiting concentrations of bioavailable iron ([FeIII] $<10^{-9.5}$ M) (29). Both possible explanations provide venues for further optimization of *in vitro* manipulations of silica precipitation by diatom molecules.

Our results also have ramifications for understanding the growth of diatoms in nature. Productivity of ≈30% of the world's oceans is limited by iron availability, and numerous field and laboratory studies have provided evidence that diatoms produce more heavily silicified cell walls in these environments (30–32). Because diatoms are responsible for ≈20% of global primary productivity (1, 2), changes in diatom sinking rates due to an increase in the amount of silica within the cell wall can influence biogeochemical cycles in the ocean (33, 34). During glacial periods, enhanced iron delivery is hypothesized to have significantly lowered atmospheric pCO_2 by reducing the silicon requirements of diatoms in the Southern Ocean, allowing the excess silicon to "leak" out of the region and fuel diatom productivity in the subtropics—the so-called Silicic Acid Leakage Hypothesis (35). The most commonly accepted explanation for the influence of iron on silica deposition is that iron limitation (through reductions of growth rate) simply lengthens the period during which silicon can be taken up and thus deposited (31, 32). Our results suggest that the relation between iron and silicon availability is more complicated and reflects an active interaction between these pathways. Development of molecular indicators of the iron/silicon nutritional status of diatoms *in situ* will enhance understanding of these important controls of marine productivity. The work described here therefore provides avenues for understanding diatom biology that have potential impacts from the nanoscale to the global scale.

Methods

Culture Work. Axenic *T. pseudonana* clone CCMP 1335, for which the whole-genome sequence is available, was used for this study. Cultures were maintained in natural seawater autoclaved and supplemented with 2× f/2 nutrients (36) at 20° ± 1°C and 100 μmol of photons $m^{-2}\cdot s^{-1}$ (24-h light). Nitrogen- and silicon-limitation experiments were conducted with sea water collected from 100 m that had been amended with 2× f/2 nutrients but without the addition of either nitrate or silicic acid. Sea water collected from ≈20 m was used for the iron-limitation experiment; 2× f/2 nutrients were added except for dissolved iron. Alkaline pH, which decreases dissolved CO_2 concentrations, was obtained by increasing the pH of 2× f/2 sea water to 8.5 by adding 1 M NaOH. In the absence of bubbling, the pH increased to a final pH of 9.4 due to photosynthetic activity. Temperature limitation entailed transferring an exponentially growing culture maintained in nutrient-replete 2× f/2 sea water at 20°C to 4°C for 24 h. All limitation experiments were conducted in parallel with growth of the nutrient-complete cultures. Cells were harvested for RNA when the growth rate began to significantly decrease compared with the control cultures (SI Fig. 5). Dissolved nutrients (phosphate, silicate, and nitrate) were measured for each treatment according to Whitledge *et al.* (37) (SI Table 1).

Tiling Arrays, RNA Work, and Gene Identification. A total of 1,308,958 36-mer probes were chosen uniformly from both strands of the 34-Mb *T. pseudonana* genome (version 2, www.jgi.doe.goc) with gaps of 10 nt between consecutive probes. Microarrays were fabricated by using a modified Maskless Array Synthesizer (MAS) as described (14, 15). Total RNA was extracted by using the Concert Plant RNA Reagent according to manufacturer's instructions for large-scale RNA isolation (Invitrogen). Total RNA was converted to double-stranded cDNA by using an oligodT primer containing the T7 RNA polymerase promoter. *In vitro* transcription for labeling of cRNA, hybridization, and washing of arrays was done as described (15). Arrays were scanned by using an Arrayworx scanner (Applied Precision).

Raw data from all 16 tiling arrays were normalized to the same scale by using a quantile normalization procedure. A preliminary set of TUs were first determined by combining neighboring probes that showed signals above a cutoff. This cutoff was computed from the hybridization levels of the random probes synthesized on the arrays, so that 85% of the random probes had signals below the level. All neighboring probes located within 100 bases and with signals above the cutoff were combined into larger clusters. In total, 15,960 clusters longer than 100 nt were determined. A set of total TUs was derived from the normalized tiling array data based on more stringent cutoff and filtering criteria. To identify the set of TUs that included the primary unpredicted TUs, a nonlinear median-of-means low-pass filter operating on a sliding window of three probes at a time was used to reduce noise and remove single probe anomalies. A set of TUs that included the secondary unpredicted TUs was derived from gene-only array data by using 14 of the gene-only arrays. Primary and secondary TUs were initially identified on the version 2 genome and then mapped to version 3 by using BLAST (38). TUs that did not overlap a gene model on the same strand were identified as unpredicted TUs. Blast (39) searches (BlastX in three reading frames; e-value cutoff 1e-5) of unpredicted TU sequence and gene model sequence were performed against the National Center for Biotechnology Information (NCBI) nonredundant protein database and also against *Arabidopsis thaliana*, *Plasmodium yoelii*, *Phytophthora ramorum*, *Phaeodactylum tricornutum*, *Tetrahymena thermophila*, *Chlamydomonas reinhardtii*, *Dictyostelium discoideum*, *Plasmodium falciparum*, and *Cyanidioschyzon merolae*. KOGs were assigned to unpredicted TU sequences and gene models based on best blast hit to the KOG database (40). InterPro IDs were assigned to unpredicted TUs by searching sense strand ORFs (100-aa cutoff) against interproscan databases by using the standalone version of InterPoScan v 12.1 (41). InterPro IDs for gene models were obtained from the Department of Energy Joint Genome Institute. A genome-wide search for possible long noncoding RNAs was performed by identifying TUs longer than 600 nt for which the longest predicted peptide corresponded to less than half the TU length. These TUs were further screened for overlap with untranslated regions of nearby coding genes.

RACE (rapid amplification of cDNA ends) experiments were conducted for five genes (JGI protein ID 20810, 269307, 42123, 25040, and 264902) identified by tiling arrays to validate TU predictions (SI Fig. 7). Full-length sequences of these genes were obtained by using the FirstChoice RLM RACE kit from Ambion according to instructions in the manual of the kit.

Gene-Specific Arrays and Differential Expression. A gene-specific array was designed with 176,320 36-mer genomic probes that included up to 68 probes from each gene and unpredicted TU. The gene-specific array was hybridized with RNA from 5 different growth conditions (SI Fig. 5). RNA sample preparation and hybridization conditions are described above. Gene Cluster 3.0 (42) was used for hierarchical cluster analysis on differentially expressed TUs. A Bayesian *t* test (16, 17) with $P < 0.001$ and ≥2-fold difference was used to identify differences in expression between treatments and controls. The program Java TreeView (http://jtreeview.sourceforge.net/) was used to generate images. Tiling and gene-specific array data are available from our web page (www.systemix.org/diatom/) and have been deposited in NCBIs Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession number GSE9697. Independent qRT-PCRs were conducted with a subset of 16 genes under silicon, iron, and nitrogen limitation as well as under the control condition to validate the gene-specific arrays (SI Table 7). Triplicate or quadruplicate 20-μl qRT-PCRs were performed for each selected gene and biological replicate. Individual reactions contained 2 μl of cDNA from the RT reaction, 10 μl of iQ SYBR green Supermix (Bio-Rad), 4.8 μl of water, and 1.6 μl of forward and reverse primers. Amplifications were conducted on an iCycler iQ Real-Time PCR Detection System (Bio-Rad) with a program of 95°C for 3 h, followed by 45 cycles of 95°C for 10 s, 60°C for 30 s, and 72°C for 50 s. Amplification efficiencies were calculated with the program LinReg PCR (43). Efficiencies for each triplicate set of reactions

were averaged together. Relative expression levels were calculated with the program Q-Gene (44) by using CT values from Bio-Rad iCycler iQ Real Time Detection System Software v.3 and averaged reaction efficiencies.

Tandem Mass Spectrometry and Proteome Analysis. Silicon limited cells were suspended and washed in buffer containing protease inhibitors and broken by agitation with glass beads in a Bead Beater apparatus largely as described by Frigeri *et al.* (24). Cell lysates were separated by centrifugation into cell-wall ($400 \times g$ and $1{,}500 \times g$), membrane ($184{,}000 \times g$), and soluble fractions. Cell-wall fractions were suspended to $\approx$1 mg of protein·ml^{-1} in 0.2–0.5 ml of 50 mM ammonium bicarbonate, 2 mM DTT buffer, and either digested directly with trypsin or first extracted with boiling 2% SDS, 10 mM DTT, 50 mM EDTA, and 1.0 M urea and then precipitated with 10% TCA, suspended in 8.0 M urea and 10 mM DTT and then diluted to 1.0 M urea in 50 mM ammonium bicarbonate buffer for trypsin digestion. Proteins from soluble and membrane fractions were digested with trypsin as described (45, 46). Peptides were extracted on Varian Spec PT C18 cartridges, suspended to $\approx$1.0 μg·μl^{-1} in 0.3% formic acid, and separated at 200 nl·min^{-1} on an Agilent HP 1100 HPLC (fused silica C18 column, 100 μM $\times$ 11 cm) coupled for MS/MS analysis to a QTOF 2 mass spectrometer (Micromass) as described (45, 46). Alternatively, the same HPLC setup was used to spot 0.2-μl samples (1,050 spots) mixed with cyano-4-hydroxycinnamic acid (4 mg·ml^{-1}) and an internal standard onto a MALDI target plate for MS/MS analysis on an ABI 4800 MALDI TOF-TOF mass spectrometer. After deisotoping and background subtraction as described (45, 46), peptides and proteins were identified via Mascot (47), with the following search conditions: MS and MS/MS search tolerances were 0.2 Da, allowing for oxidation of methionine and N-terminal acetylation as variable modifications, and tryptic digestion was assumed with up to two missed cleavages. Mascot searches were performed separately against each of two databases. The first (JGI) contained forward and reversed protein sequences derived from JGI gene models (*T. pseudonana* v3.0), whereas the second (6-RF) contained protein sequences derived from translation of the entire *T. pseudonana* v3.0 genome in all six reading frames. The second database also contained reversed forms of each amino acid sequence. Protein sequences of common contaminant proteins such as porcine trypsin and human keratins were appended to both databases. Mascot score thresholds corresponding to an estimated 1% peptide false discovery rate were established by using a reversed database strategy as described by Huttlin *et al.* (48). Thresholds were 28 and 40 for QTOF data searched against the JGI and 6-RF databases, respectively, whereas equivalent thresholds were 28 and 39 for TOF/TOF data.

ACKNOWLEDGMENTS. We thank Igor Grigoriev and Bobby Otillar from the Joint Genome Institute for useful discussions. We also thank Drs. Amy Harms and Gregory Barrett-Wilt of the University of Wisconsin Biotechnology Center Mass Spectrometry Facility (Madison, WI) for access to instrumentation and for technical advice and assistance with the proteomic analyses. This work was supported by the postdoctoral program of the German Academic Exchange Service (T.M.), the University of Wisconsin National Institutes of Health Genomic Sciences Training Grant (M.R.), the University of Wisconsin Graduate School and the National Science Foundation (M.R.S.), and the Gordon and Betty Moore Foundation (E.V.A.).

1. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) *Science* 281:237–240.
2. Hildebrand M, Wheterbee R (2003) in *Progress in Molecular and Subcellular Biology*, ed Mueller WEG (Springer, Berlin), pp 11–57.
3. Chiovitti A, Harper RE, Willis A, Bacic A, Mulvaney P, Wetherbee R (2005) *J Phycol* 41:1154–1161.
4. Hildebrand M (2005) *J Nanosci Nanotech* 5:1–12.
5. Kröger N, Deutzmann R, Sumper M (1999) *Science* 286:1129–1132.
6. Kröger N, Deutzmann R, Bergsdorf C, Sumper M (2000) *Proc Natl Acad Sci USA* 97:14133–14138.
7. Kröger N, Lorenz S, Brunner E, Sumper M (2000) *Science* 298:584–586.
8. Poulsen N, Sumper M, Kröger N (2003) *Proc Natl Acad Sci USA* 100:12075–12080.
9. Poulsen N, Berne C, Spain J, Kröger N (2007) *Angew Chem Int Ed* 46:1843–1846.
10. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou SG, Allen AE, Apt KE, Bechner M (2004) *Science* 306:79–86.
11. Thamatrakoln K, Hildebrand M (2007) *Eukaryot Cell* 6:271–279.
12. Poulsen N, Kröger N (2004) *J Biol Chem* 279:42993–42999.
13. Ma JF , Tamai K, Yamaji N, Mitani N, Konishi S, Katshuhara M, Ishiguro M, Murata Y, Yano M (2006) *Nature* 440:688–691.
14. Sing-Gasson, Green RF, Yue Y, Nelson C, Blattner F, Cerrina F, Sussman MR (1999) *Nat Biotechnol* 17:974–978.
15. Stolc V, Samanta MP, Tongprasit W, Sethi H, Liang S, Nelson DC, Hegeman A, Nelson C, Rancour D, Bednarek S (2005) *Proc Natl Acad Sci USA* 102:4453–4458.
16. Baldi P, Long AD (2001) *Bioinformatics* 17:509–519.
17. Fox RJ, Dimmic MW (2005) *BMC Bioinformatics* 7:126–136.
18. Cavener RD (1987) *Nucleic Acids Res* 15:1353–1361.
19. Cavener RD, Ray SC (1991) *Nucleic Acids Res* 19:3185–3192.
20. Bartel DP (2004) *Cell* 116:281–297.
21. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Susuki M, Kawai J (2005) *Science* 309:1564–1566.
22. Samanta MP, Tongprasit W, Sethi H, Chin C-S, Stolc V (2006) *Proc Natl Acad Sci USA* 103:4192–4197.
23. Hildebrand M, Volcani BE, Gassmann W, Schroeder JI (1997) *Nature* 385:688–689.
24. Frigeri LG, Radabaugh TR, Haynes PA, Hildebrand M (2006) *Mol Cell Proteomics* 5:182–193.
25. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) *Proc Natl Acad Sci USA* 97:6652–6657.
26. Lee JM, Sonnhammer ELL (2007) *Genome Res* 13:875–882.
27. Davis AK, Hildebrand M, Palenik BA (2005) *J Phycol* 41:577–589.
28. Hamm C, Merkel R, Springer O, Jurkojc P, Maier C, Prechtel K, Smetacek V (2003) *Nature* 421:841–843.
29. Ellwood MJ, Hunter KA (2000) *Limnol Oceanogr* 45:1517–1524.
30. Hutchins DA, Bruland KW (1998) *Nature* 393:561–564.
31. DeLa Rocha CL, Hutchins DA, Brzezinski MA, Zhang YH (2000) *Mar Ecol Prog Ser* 195:71–79.
32. Timmermans KR, van der Wagt B, de Baar HJW (2004) *Limnol Oceanogr* 49:2141–2151.
33. Boyd PW (2007) *Science* 315:612–617.
34. De Baar HJW, Boyd PW, Coale KH, Landry MR, Tsuda A, Assmy P, Bakker DCE, Bozek Y, Barber RT, Brezinski MA (2005) *J Geophys Res* 110:C09S16.
35. Brzezinski MA, Pride CJ, Franck VM, Sigman DM, Sarmiento JL, Matsumoto K, Gruber N, Rau GH, Coale KH (2002) *Geophys Res Lett* 29:1564.
36. Guillard RR, Ryther JH (1962) *Can J Microbiol* 8:229–239.
37. Whitledge TE, Malloy SC, Patton CJ, Wirick CD (1981) *BNL 51398* (Brookhaven Natl Laboratory, Upton, NY), p 216
38. Kent WJ (2002) *Genome Res* 12:656–664.
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) *Nucleic Acids Res* 25:3389–3402.
40. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankaravam UT, Rao BS, Kiryutin B, Galparin MY, Fedorova ND, Koonin EV (2001) *Nucleic Acids Res* 29:22–28.
41. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler I, Lopez R (2005) *Nucleic Acids Res* 33:116–121.
42. Hoon MJLD, Imoto S, Nolan J, Miyano S (2004) *Bioinformatics* 20:1453–1454.
43. Ramakers C, Ruijter JM, Deprez RHL, Moorman AFM (2003) *Neurosci Lett* 15:339–346.
44. Simon P (2003) *Bioinformatics* 19:1439–1440.
45. Nelson CJ, Huttlin EL, Hegeman AD, Harms AC, Sussman MR (2007) *Proteomics* 7(8):1279–1292.
46. Huttlin EL, Hegeman AD, Harms AC, Sussman MR (2007) *Mol Cell Proteom* 6(5):860–881.
47. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) *Electrophoresis* 20:3551–3567.
48. Huttlin EL, Hegeman AD, Harms AC, Sussman MR (2007) *J Proteome Res* 6:392–398.