Positive Selection within a Diatom Species Acts on Putative Protein Interactions and Transcriptional Regulation

Julie A. Koester,†,‡,§ Willie J. Swanson,‡ and E. Virginia Armbrust*,†

†School of Oceanography, University of Washington
‡Department of Genome Sciences, University of Washington
§Department of Geography and Environment, Mount Allison University
*Corresponding author: E-mail: armbrust@uw.edu.
Associate editor: Todd Oakley

Abstract

Diatoms are the most species-rich group of microalgae, and their contribution to marine primary production is important on a global scale. Diatoms can form dense blooms through rapid asexual reproduction; mutations acquired and propagated during blooms likely provide the genetic, and thus phenotypic, variability upon which natural selection may act. Positive selection was tested using genome and transcriptome-wide pair-wise comparisons of homologs in three genera of diatoms (Pseudo-nitzschia, Ditylum, and Thalassiosira) that represent decreasing phylogenetic distances. The signal of positive selection was greatest between two strains of Thalassiosira pseudonana. Further testing among seven strains of T. pseudonana yielded 809 candidate genes of positive selection, which are 7% of the protein-coding genes. Orphan genes and genes encoding protein-binding domains and transcriptional regulators were enriched within the set of positively selected genes relative to the genome as a whole. Positively selected genes were linked to the potential selective pressures of nutrient limitation and sea surface temperature based on analysis of gene expression profiles and identification of positively selected genes in subsets of strains from locations with similar environmental conditions. The identification of positively selected genes presents an opportunity to test new hypotheses in natural populations and the laboratory that integrate selected genotypes in T. pseudonana with their associated phenotypes and selective forces.

Key words: positive selection, natural selection, diatom, evolution.

Introduction

Diatoms are evolutionarily the youngest members of the phytoplankton, originating near the boundary of the Permian and Triassic periods ~250 Ma, a time of mass extinction in the world ocean (Raup 1979; Sorhannus 2007). Radiation of diatoms into marine, freshwater, soil, and ice ecosystems is evidence of genetic specialization and an exemplary ability to adapt. The number of marine diatom species has increased during the last 65 My, reducing the global silicic acid concentration due to the incorporation of precipitated silica into the diatom cell walls (Siever 1991; Falkowski et al. 2004). The biogeochemical cycles of carbon and silicon are coupled by diatoms and to a lesser extent by other siliceous organisms. Diatoms control the biological portion of the silicon cycle in today’s oceans and contribute substantially to primary production and the export of carbon to the deep sea (Maliva et al. 1989; Tréguer and Pondaven 2000).

Diatoms are the most species-rich group of phytoplankton in the ocean (Kooistra et al. 2007), and they appear to adapt rapidly to local conditions. Marine species with circum-global distributions have regionally, genetically structured populations (Rynearson and Armbrust 2004; Casteleyn et al. 2010), despite wide distribution via ocean currents. Growth rates of diatoms occur on the same time scale as environmental fluctuations within different marine ecosystems. Estuarine and coastal species are adapted to daily fluctuations in photosynthetically available light that ranges from extremely low to stressful levels as the turbidity of the water column and their position in it changes; in contrast, open ocean species live in consistently clear water (Lavaud et al. 2007). Over the course of a week, coastal species can experience transitions from cool, high-nutrient, upwelling conditions to warmer, low-nutrient, downwelling states (Austin and Barth 2002). In the nutrient-poor open ocean, normally low diatom abundances increase dramatically in response to ephemeral, cold-core eddies that bring nutrients to the surface, allowing diatoms to grow quickly into blooms that can persist for up to a month (Benitez-Nelson et al. 2007).

Diatoms have a high capacity to accrue mutations upon which selection may act. They are diploid and can theoretically accommodate a high proportion of recessive mutations. Their life history is dominated by asexual reproduction, during which cells divide at least once per day during bloom conditions (Furnas 1990). Sexual generation times are ~2 years in marine diatoms (D’Alelio et al. 2010; Holtermann et al. 2010); thus, genetic variation increases and surviving mutations are passed along relatively quickly. The mutational load carried by diatoms is evident in the sequenced genomes of Thalassiosira pseudonana and Phaeodactylum tricornutum. The amount of genetic diversity accrued between T. pseudonana and Pha. tricornutum in 90 My is similar to that accrued between mammals and fish in 550 My (Bowler et al. 2008); however, when normalized to
The two cryptic species of Ditylum brightwellii are differentia
ted by genome size and likely diverged recently because the
larger genome-sized species is thus far found only in the
northeastern Pacific Ocean (Koester et al. 2010). By rigorously
testing genes within seven strains of T. pseudonana using a
phylogenetic framework, we identified a large suite of posi-
tively selected genes and explored links between the genes,
potential phenotypes, and environmental selective forces
using profiles of gene expression and branch-site models of
selection among different strains. We used the $d_{NS}^{A}$/d_{SS}^{A}
metric instead of population-based polymorphism methods beca
se preliminary observations of the data revealed polymorphic
differences between strains, which were sampled from distant
locations and across decades suggesting that, similar to other
planktonic diatoms, there is population structure within
T. pseudonana.

Materials and Methods

Sequence Data and Identification of Homologous
Genes

We used transcriptomic and genomic data to detect positive
selection in three diatom genera, Pseudo-nitzschia, Ditylum,
and Thalassiosira (supplementary table S1, Supplementary
Material online: database depositions). The transcriptomes
of the three species of Pseudo-nitzschia were sequenced and
quality curated previous to this study (supplementary
table S1, Supplementary Material online). For the two species
of Ditylum, we extracted total RNA from exponentially grow-
ing cultures using Plant RNA reagent (Invitrogen, Life
Technologies), selected mRNA with the MicroPoly(A)Purist
Kit (Ambion; Life Technologies), and reverse transcribed the
mRNA using the SuperScript Double-Stranded cDNA
synthesis Kit (Invitrogen, Life Technologies). Pyrosequencing
was performed in the Schuster Laboratory (University Park,
PA). We used an automated pipeline that integrates cus-
tom scripts with extant bioinformatic tools to remove ribo-
somal sequence with BLAST, and Lucy to trim poly-A tails
and low-quality sequence with scores <14. Quality-curated
reads were assembled with CABOG default settings
(Miller et al. 2008).

We identified homologous genes among transcripts of the
Pseudo-nitzschia species and between the two Ditylum
species using another pipeline. In brief, the pipeline retrieved
all possible open reading frames (ORFs) of at least 25 amino
cids in length for each contig using getorf (EMBOSS). ORFs
were compared with protein databases including those of
several phytoplankton and the nonredundant (nr) protein
database of NCBI using BLASTp at an e-value cutoff of $10^{-3}$
(Altschul et al. 1997). Contigs with the lowest e value were
selected for further analysis. If a contig did not have a protein
match, the longest ORF was chosen. Putative homologs be
tween diatom species were paired using a best reciprocal
blastp between the transcripts of each species at an e-value
cutoff of $10^{-10}$. Homologs were aligned on the translated
protein sequence using CLUSTALW2 (Larkin et al. 2007)
and converted to the original DNA sequence with revtrans.py
(Wernersson and Pedersen 2003). Aligned homologs were

generation time, the number of neutral substitutions per
generation is equivalent between the pairs. In addition, the
genome of T. pseudonana has, on average, one polymorphism
per 150 bases (Armbrust et al. 2004), which is an order of
magnitude more than humans (Sachidanandam et al. 2001)
but an order of magnitude less than the purple sea urchin
(Sodergren et al. 2006).

Gene variants associated with adaptive phenotypes are
positively selected and identified as the protein-coding
genes in which the ratio of nonsynonymous substitutions
at nonsynonymous sites to synonymous substitutions at syn-
nonymous sites ($d_{NS}$/d_{SS}) is greater than one among homologs of
two or more organisms. Positively selected genes and their
encoded proteins share traits with other divergent and diver-
ging genes such as those under relaxed purifying selection.
Rates of protein sequence divergence tend to be higher in
lineage-specific, or “young,” proteins, than in proteins with
deep evolutionary histories (Alba and Castresana 2005; Wolf
et al. 2003). Aligned homologs were

Materials and Methods

Sequence Data and Identification of Homologous
Genes

We used transcriptomic and genomic data to detect positive
selection in three diatom genera, Pseudo-nitzschia, Ditylum,
and Thalassiosira (supplementary table S1, Supplementary
Material online: database depositions). The transcriptomes
of the three species of Pseudo-nitzschia were sequenced and
quality curated previous to this study (supplementary
table S1, Supplementary Material online). For the two species
of Ditylum, we extracted total RNA from exponentially grow-
ing cultures using Plant RNA reagent (Invitrogen, Life
Technologies), selected mRNA with the MicroPoly(A)Purist
Kit (Ambion; Life Technologies), and reverse transcribed the
mRNA using the SuperScript Double-Stranded cDNA
synthesis Kit (Invitrogen, Life Technologies). Pyrosequencing
was performed in the Schuster Laboratory (University Park,
PA). We used an automated pipeline that integrates cus-
tom scripts with extant bioinformatic tools to remove ribo-
somal sequence with BLAST, and Lucy to trim poly-A tails
and low-quality sequence with scores <14. Quality-curated
reads were assembled with CABOG default settings
(Miller et al. 2008).

We identified homologous genes among transcripts of the
Pseudo-nitzschia species and between the two Ditylum
species using another pipeline. In brief, the pipeline retrieved
all possible open reading frames (ORFs) of at least 25 amino
cids in length for each contig using getorf (EMBOSS). ORFs
were compared with protein databases including those of
several phytoplankton and the nonredundant (nr) protein
database of NCBI using BLASTp at an e-value cutoff of $10^{-3}$
(Altschul et al. 1997). Contigs with the lowest e value were
selected for further analysis. If a contig did not have a protein
match, the longest ORF was chosen. Putative homologs be
tween diatom species were paired using a best reciprocal
blastp between the transcripts of each species at an e-value
cutoff of $10^{-10}$. Homologs were aligned on the translated
protein sequence using CLUSTALW2 (Larkin et al. 2007)
and converted to the original DNA sequence with revtrans.py
(Wernersson and Pedersen 2003). Aligned homologs were

Materials and Methods

Sequence Data and Identification of Homologous
Genes

We used transcriptomic and genomic data to detect positive
selection in three diatom genera, Pseudo-nitzschia, Ditylum,
and Thalassiosira (supplementary table S1, Supplementary
Material online: database depositions). The transcriptomes
of the three species of Pseudo-nitzschia were sequenced and
quality curated previous to this study (supplementary
table S1, Supplementary Material online). For the two species
of Ditylum, we extracted total RNA from exponentially grow-
ing cultures using Plant RNA reagent (Invitrogen, Life
Technologies), selected mRNA with the MicroPoly(A)Purist
Kit (Ambion; Life Technologies), and reverse transcribed the
mRNA using the SuperScript Double-Stranded cDNA
synthesis Kit (Invitrogen, Life Technologies). Pyrosequencing
was performed in the Schuster Laboratory (University Park,
PA). We used an automated pipeline that integrates cus-
tom scripts with extant bioinformatic tools to remove ribo-
somal sequence with BLAST, and Lucy to trim poly-A tails
and low-quality sequence with scores <14. Quality-curated
reads were assembled with CABOG default settings
(Miller et al. 2008).

We identified homologous genes among transcripts of the
Pseudo-nitzschia species and between the two Ditylum
species using another pipeline. In brief, the pipeline retrieved
all possible open reading frames (ORFs) of at least 25 amino
cids in length for each contig using getorf (EMBOSS). ORFs
were compared with protein databases including those of
several phytoplankton and the nonredundant (nr) protein
database of NCBI using BLASTp at an e-value cutoff of $10^{-3}$
(Altschul et al. 1997). Contigs with the lowest e value were
selected for further analysis. If a contig did not have a protein
match, the longest ORF was chosen. Putative homologs be
between diatom species were paired using a best reciprocal
blastp between the transcripts of each species at an e-value
cutoff of $10^{-10}$. Homologs were aligned on the translated
protein sequence using CLUSTALW2 (Larkin et al. 2007)
and converted to the original DNA sequence with revtrans.py
(Wernersson and Pedersen 2003). Aligned homologs were

Materials and Methods

Sequence Data and Identification of Homologous
Genes

We used transcriptomic and genomic data to detect positive
selection in three diatom genera, Pseudo-nitzschia, Ditylum,
and Thalassiosira (supplementary table S1, Supplementary
Material online: database depositions). The transcriptomes
of the three species of Pseudo-nitzschia were sequenced and
quality curated previous to this study (supplementary
table S1, Supplementary Material online). For the two species
of Ditylum, we extracted total RNA from exponentially grow-
ing cultures using Plant RNA reagent (Invitrogen, Life
Technologies), selected mRNA with the MicroPoly(A)Purist
Kit (Ambion; Life Technologies), and reverse transcribed the
mRNA using the SuperScript Double-Stranded cDNA
synthesis Kit (Invitrogen, Life Technologies). Pyrosequencing
was performed in the Schuster Laboratory (University Park,
PA). We used an automated pipeline that integrates cus-
tom scripts with extant bioinformatic tools to remove ribo-
somal sequence with BLAST, and Lucy to trim poly-A tails
and low-quality sequence with scores <14. Quality-curated
reads were assembled with CABOG default settings
(Miller et al. 2008).

We identified homologous genes among transcripts of the
Pseudo-nitzschia species and between the two Ditylum
species using another pipeline. In brief, the pipeline retrieved
all possible open reading frames (ORFs) of at least 25 amino
cids in length for each contig using getorf (EMBOSS). ORFs
were compared with protein databases including those of
several phytoplankton and the nonredundant (nr) protein
database of NCBI using BLASTp at an e-value cutoff of $10^{-3}$
(Altschul et al. 1997). Contigs with the lowest e value were
selected for further analysis. If a contig did not have a protein
match, the longest ORF was chosen. Putative homologs be
between diatom species were paired using a best reciprocal
blastp between the transcripts of each species at an e-value
cutoff of $10^{-10}$. Homologs were aligned on the translated
protein sequence using CLUSTALW2 (Larkin et al. 2007)
and converted to the original DNA sequence with revtrans.py
(Wernersson and Pedersen 2003). Aligned homologs were
trimmed to blunt ends with at least three identical amino acids at each end and converted to PHYLIP format for analysis of positive selection.

Seven strains of Thalassiosira were sequenced with an ABI SOLiD and mapped with BWA 0.5.9 (parameters: bwa aln -k 2 -l 18 -n .001) to the reference strain (CCMP1335) previously sequenced using the Sanger method (Armbrust et al. 2004). SOLiD-sequenced reads were trimmed based on quality (Iverson et al. 2012); reads less than 24 bp that did not meet the quality threshold were discarded. Alignments of the homologous genes consist of the majority consensus sequence of each strain with the introns removed. The T. pseudonana υ. 3.0 gene models (http://genomejgi.doe.gov/Thaps3/Thaps3.home.html) defined the start, stop, and intron boundaries.

Pair-Wise Tests for Detecting Positive Selection at Increasing Phylogenetic Distance

We used five pair-wise tests (table 1) to determine the phylogenetic distance that best detects positive selection in diatoms. Gene sets from two strains of Thalassiosira, two species of Ditylum, and three species pairs of Pseudo-nitzschia were each tested. We calculated species divergence from 18S ribosomal DNA (rDNA) sequences aligned with CLUSTALW2 and trimmed to the length of the shortest partial sequence, which belonged to Pseudo-nitzschia multistriata. Percent divergence was calculated in BioEdit (Hall 1999).

Phylogenetic analysis by maximum likelihood (PAML version 4.4 [Yang 2007]) was used to analyze each set of paired genes (table 1) with the program codeml in runmode = −2 (pairwise), model = 0, NSites = 0 (one dk,ds value), and fix_omega = 0 (estimates omega) and cleandata = 1, such that sites with ambiguity codes were removed from analysis. We calculated dK,ds for synonymous substitutions (ds > 1.0) and used the ds,ds value to evaluate at what phylogenetic distance positive selection could be detected (dk,ds > 1.0).

Testing for Positive Selection among Seven Strains of T. pseudonana

We further tested the genes with dk,ds ≥ 0.5 in the pair-wise analysis of T. pseudonana for positive selection among seven genetically distinct strains because approximately 80% of genes with a dk,ds ≥ 0.5 are under positive selection (Swanson et al. 2004). A coalescent tree was generated from the individual gene trees (Supplementary Fig. S1, Supplementary Material online). Gene trees and the coalescent tree were constructed with RaXML 7.2.5 (GTRPROTGAMMAWAG [Stamatakis 2006]) and PhyloNet (unrooted minimum coalescence [Than and Nakhleh 2009]), respectively. Individual genes were tested for positive selection by comparing how well the data fit a null model (M8a, nearly neutral) versus a selection model (M8). Both models allow omega (dk,ds for the tree of seven strains) to vary at different sites along the gene alignment. In the neutral model M8a, omega is distributed between zero and one, whereas in the selection model M8, omega may also be greater than one. We ran null and selection models three and four times, respectively, to evaluate convergence of the likelihood estimates. Any genes that did not converge within the selection model were removed from further analysis. We calculated the likelihood ratio tests (Irt) for each gene using the formula: Irt = −2(ln(likelihood)\textsubscript{NULL} - ln(likelihood)\textsubscript{SELECTION}). The test statistic generally follows a χ² distribution, therefore nominal significance (P value) was determined by integrating the right-hand side of the χ² distribution for one degree of freedom (number of parameters: M8a = 16, M8 = 17) by the Irt statistic. The statistical significance for multiple tests was adjusted using a Bonferroni correction and a false discovery rate (FDR) of 0.01 with Q value

Table 1. Phylogenetic Divergence and Rates of Synonymous (ds) and Nonsynonymous (dn) Substitutions in Homologs of Three Diatom Genera Using Pair-Wise Tests.

<table>
<thead>
<tr>
<th>Species</th>
<th>18S div. (%)</th>
<th>Seq (No.)</th>
<th>Paired Genes (No.)</th>
<th>Mean Align. Length (bp)</th>
<th>Saturated Pairs ds ≥ 1.0</th>
<th>Unsaturated Pairs ds &lt; 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No.</td>
<td>Percentage</td>
</tr>
<tr>
<td>Thalassiosira pseudonana (NY)</td>
<td>0</td>
<td>11,390</td>
<td>11,355</td>
<td>1,488</td>
<td>9</td>
<td>0.08</td>
</tr>
<tr>
<td>T. pseudonana (Wales)</td>
<td>11,390</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ditylum brightwellii 1</td>
<td>0</td>
<td>3,910</td>
<td>113</td>
<td>204</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>D. brightwellii 2</td>
<td>477</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudo-nitzschia multistriata</td>
<td>0.60</td>
<td>16,512</td>
<td>2,653</td>
<td>483</td>
<td>2,359</td>
<td>89</td>
</tr>
<tr>
<td>P. multiseries</td>
<td>16,535</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. australis</td>
<td>1.03</td>
<td>920</td>
<td>277</td>
<td>324</td>
<td>176</td>
<td>64</td>
</tr>
<tr>
<td>P. multistriata</td>
<td>16,512</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. australis</td>
<td>1.12</td>
<td>920</td>
<td>404</td>
<td>336</td>
<td>266</td>
<td>66</td>
</tr>
<tr>
<td>P. multiseries</td>
<td>16,535</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of starting sequences. T. pseudonana: JGI version 3.0 gene model transcripts; others: BLASTp-annotated or the longest ORF.

*a* Accession: AM235384 and P. multistriata.

*b* Accession: AM235384 and P. multistriata.

*c* Accession: AM235384 and P. multistriata.
(parameters: lambda = range 0.0 to 0.9 by 0.05; $\pi_0$ method = bootstrapped; Storey and Tibshirani 2003).

Testing for Functional Enrichment and Orphan Status of Genes under Positive Selection

We submitted the T. pseudonana version 3.0 genes to InterProScan (Zdobnov and Apweiler 2001) and performed BLASTp against the nr database at NCBI to assess functional annotations for each gene. Orphans are defined as genes whose translated proteins either have no matches or matches with an e-value of $10^{-5}$ or greater from the BLAST-searched databases. Search databases included genomic and transcriptomic sequence from five additional diatoms, two oomycetes, two cryptophytes, three prasinophytes, one haptophyte, one amoeba, one ciliate, one green alga, and the nr. Gene ontology (GO) terms were extracted from the InterProScan results to test for significant associations of GO terms within the set of positively selected genes using GOSTATS (Falcon and Gentleman 2007). We performed conditional, hypergeometric tests for over-representation of GO terms separately for each of the GO ontologies, molecular function, cell component, and biological process. A Bonferroni correction and an FDR of 0.05 were used to adjust P values for multiple tests.

Expression of Positively Selected Genes in T. pseudonana (Strain CCMP1335)

We used T. pseudonana transcription data from Mock et al. (2008) to test the hypothesis that positively selected genes have lower levels of expression than neutral or purified genes and to identify growth conditions under which positively selected genes are coexpressed. The relative level of expression was determined for each gene by taking the median fluorescence of the aggregated (previously quantile normalized) probes and replicates of each experimental condition. We tested differences in the distributions of gene expression between positively selected genes and genes evolving neutrally and under purifying selection for each experimental condition using one-sided Mann–Whitney tests (wilcox.text R v2.12.1). Correlations between gene expression and the number of exons per gene, mature transcript length, and G + C content in the transcript were investigated using descriptive statistics and the linear model implemented in R. Two-way hierarchical clustering (Cluster 3.0 [Eisen et al. 1998]) with the city blocks distance algorithm was used to group both genes and experimental conditions by similarity of expression patterns to identify genes that were coexpressed under specific conditions and thus potentially coevolving.

Testing Positive Selection along Specified Lineages of T. pseudonana

Four different branch-site selection models were used to test for positively selected genes in subsets of strains grouped by environmental conditions at the site each strain was collected. Two hypotheses addressed seasonal temperature variability. We placed strains in one of two groups based on the variability in seasonal sea surface temperature (SST) where they were collected. Differences in seasonal SST were plotted and calculated using the Smith and Reynolds climatology from the NCEP NOMADS Meteorological Data Server for SST between January and July 1970–2000 at the location each strain was collected (http://www.emc.ncep.noaa.gov/research/cmb/sst_analysis/#_cch2_1007146782, last accessed November 1, 2012). Two hypotheses explored the potential for ecosystem type and geographic isolation to promote positive selection using the open ocean strain (CCMP1014) and the Adriatic strain (RcTP), respectively. The null branch-site model A1 fixes omega to 1.0 on the branch(es) of interest, whereas the selection model A estimates a distribution of omega values across the gene on the chosen branch(es).

We also tested the amount of time that strains have been in culture as a selective force using a branch model. Omega values were estimated for strains based on the decade in which they were collected. Branch models estimate one omega value per designated set of lineages for the entire gene. The null model M0 fixes omega to 1.0 and the selection model (model = 2) estimates omega for each set of lineages. Genes in which the lrt is significant and omega > 1.0 for the specified branches are positively selected.

The branch-site and branch models used the same tree topology and statistics as described earlier with the exception that individual model parameters differed. Branches of the tree were annotated to support each of the different hypotheses being tested. The background omega was applied to internal branches unless both tips originating from a node were being tested with the same omega, then that internal branch would be assigned to the alternative, estimated omega.

Results

Pair-Wise Tests for Positive Selection

We determined the best phylogenetic distance at which to detect positive selection in diatoms through pair-wise comparisons of homologs within and between species. The phylogenetic distance, or percent divergence of the 18S rDNA sequence, ranges from zero between both the Thalassiosira strains and Ditylum species to 1.12% between P. australis and P. multiseries (table 1). The proportion of gene pairs saturated for synonymous substitutions is greatest in the interspecies comparisons of Pseudo-nitzschia, accounting for 89% of those tested between P. multistriata and P. multiseries (table 1). We did not use gene pairs saturated for synonymous substitutions to detect positive selection because the synonymous substitution rate is uncertain. Rates of synonymous and nonsynonymous substitutions between homologs decrease with decreasing phylogenetic distance. The rates of synonymous substitution between homologs of the sister species of Ditylum and strains of Thalassiosira are one and two orders of magnitude less, respectively, than the interspecies comparisons of Pseudo-nitzschia (table 1). The length of gene alignments does not appear to affect the estimates of the rates of substitution.

The majority of homologous pairs that are not saturated for synonymous substitutions are under strong purifying
selection with $d_N/d_S \leq 0.1$ in all comparisons except the Thalassiosira strains (fig. 1). Most Pseudo-nitzschia homologs have a $d_N/d_S < 0.4$, consistent with purifying and relaxed purifying selection and only a few have a $d_N/d_S > 1.0$, indicating that they may be subject to positive selection. Intermediate values of $d_N/d_S$ for the homologs of both Ditylum and Thalassiosira suggest that many are experiencing relaxed purifying selection. Approximately 10% and 20% of the Ditylum and Thalassiosira homologs, respectively, have $d_N/d_S > 1.0$. We did not perform lrt s of statistical significance for this initial survey of the genes.

Positive Selection among Seven Strains of T. pseudonana

The genome of the T. pseudonana reference strain (CCMP1335) includes 11,390 gene models of which 11,355 were tested in the pair-wise survey—35 genes had internal stop codons in one or the other strain and were not tested. In the test of the two T. pseudonana strains, 3,565 genes had a $d_N/d_S \geq 0.5$; therefore, we tested these genes more rigorously among seven strains to determine the statistical support for positive selection. The maximum likelihood approach used here incorporates a phylogenetic model allowing omega ($d_N/d_S$ for a gene tree) to vary across the alignment of each gene. Maximum likelihood analysis identified 2,035 genes with a nominal $P$ value $\leq 0.05$ (supplementary table S2, Supplementary Material online). After correcting for multiple tests, 809 (Bonferroni, $P$ value $< 1.5 \times 10^{-5}$) and 1,784 (FDR $= 0.01$) genes emerged as strong candidates of positive selection representing 7% and 16% of the protein-coding genes, respectively (supplementary table S2, Supplementary Material online). The set of 809 genes is a subset of 1,784 genes and is referred to as the positively selected set.

Protein functions encoded by the 809 positively selected genes were assessed using the GO data structure and InterPro for greater resolution of information concerning protein domains. GO terms are currently assigned to 329 (41%) of the 809 positively selected genes and 5,875 (52%) of the 11,390 total coding genes. We performed enrichment analysis for genes with GO terms. Six GO terms, including 146 unique genes, are over-represented in the positively selected set of genes relative to the distribution of GO terms for all protein-coding genes. The majority (112) of over-represented genes encode proteins involved in protein–protein interactions (table 2). One-third (34%) of the 329 positively selected genes with GO terms are represented by GO:0005515, the term for protein binding, in contrast to 18% of all genes with GO terms. Genes encoding biosynthetic and metabolic regulatory proteins are also over-represented within the positively selected set (table 2). Transcription factors dominate the regulatory genes with the majority (25 of 28) represented by GO:0009889; 8.5% of positively selected genes versus 4.3% of all genes with GO terms are represented by this term (supplementary table S3, Supplementary Material online). Protein domains with the greatest representation are WD40, zinc fingers, tetratricopeptides, PDZ, and domains associated with heat shock functions (supplementary table S3, Supplementary Material online). Orphan genes, for which no homologs were found in other organisms, represent 24% (191 genes) of the positively selected genes and 15% (1,718 genes) of all coding genes.

Expression of Positively Selected Genes in T. pseudonana (Strain CCMP1335)

We analyzed relative mRNA expression of T. pseudonana CCMP1335 from a publicly available data set with respect to the positively selected genes. Thalassiosira pseudonana had been limited for growth by silicic acid, iron, nitrate, carbon dioxide, or low temperature (4°C) or maintained under nutrient replete conditions (Mock et al. 2008). The data set of expressed genes comprised 8,996 genes, of which 636 are in the positively selected set of 809. The positively selected genes are expressed at lower levels than those under neutral and purifying selection (fig. 2). In each experimental treatment, the highest value of expression is an order of magnitude less for positively selected genes than genes evolving neutrally or under purifying selection (fig. 2). In addition, the frequency distributions of gene expression for positively selected and neutral and purified genes are significantly different (Mann–Whitney; fig. 2). These results are not affected by the number of exons per gene, mature transcript length, or G+C content within the mature transcript (data not shown).

Two-way hierarchical clustering of the expression of the 636 positively selected genes highlights similarities and differences among groups of genes with respect to the experimental treatments (fig. 3). Two groups of genes are of special interest because they have very low and specific patterns of

---

**Fig. 1.** Frequency distributions of the $d_N/d_S$ of homologs between diatom pairs for genes that are not saturated for silent substitutions (i.e., for genes where $d_S < 1.0$). Number of genes analyzed is in parentheses.
Table 2. Functional enrichment in 809a positively selected genes of *Thalassiosira pseudonana*.

<table>
<thead>
<tr>
<th>Ontology</th>
<th>P-cutb</th>
<th>GO ID</th>
<th>P</th>
<th>Odds Ratio</th>
<th>Exp²</th>
<th>Obsd</th>
<th>Totale</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>18</td>
<td>GO:0005515</td>
<td>6.05E-12</td>
<td>2.43</td>
<td>62</td>
<td>112</td>
<td>1,098</td>
<td>Protein binding**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0009982</td>
<td>0.001353</td>
<td>4.34</td>
<td>2.23</td>
<td>8</td>
<td>39</td>
<td>Pseudouridine synthase activity</td>
</tr>
<tr>
<td>CC</td>
<td>6</td>
<td>GO:0005634</td>
<td>0.007638</td>
<td>1.89</td>
<td>16</td>
<td>26</td>
<td>313</td>
<td>Nucleus</td>
</tr>
<tr>
<td>BP</td>
<td>23</td>
<td>GO:0009889</td>
<td>2.69E-05</td>
<td>2.63</td>
<td>13</td>
<td>29</td>
<td>282</td>
<td>Regulation of biosynthetic process***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0060255</td>
<td>3.53E-05</td>
<td>2.54</td>
<td>14</td>
<td>30</td>
<td>301</td>
<td>Regulation of macromolecule metabolic process**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:009304</td>
<td>4.20E-05</td>
<td>2.19</td>
<td>24</td>
<td>43</td>
<td>530</td>
<td>Nucleic acid metabolic process***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0051171</td>
<td>7.57E-05</td>
<td>2.46</td>
<td>14</td>
<td>29</td>
<td>298</td>
<td>Regulation of nitrogen compound metabolic process***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0031323</td>
<td>0.000126</td>
<td>2.34</td>
<td>15</td>
<td>30</td>
<td>322</td>
<td>Regulation of cellular metabolic process***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0080090</td>
<td>0.000194</td>
<td>2.31</td>
<td>15</td>
<td>29</td>
<td>314</td>
<td>Regulation of primary metabolic process**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:001522</td>
<td>0.00022</td>
<td>5.90</td>
<td>2</td>
<td>8</td>
<td>37</td>
<td>Pseudouridine synthesis**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:006355</td>
<td>0.00039</td>
<td>2.48</td>
<td>10</td>
<td>22</td>
<td>219</td>
<td>Regulation of transcription, DNA dependent**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0032774</td>
<td>0.000819</td>
<td>2.33</td>
<td>11</td>
<td>22</td>
<td>231</td>
<td>RNA biosynthetic process**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0050789</td>
<td>0.000878</td>
<td>1.89</td>
<td>24</td>
<td>39</td>
<td>513</td>
<td>Regulation of biological process**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0034641</td>
<td>0.001622</td>
<td>1.64</td>
<td>47</td>
<td>65</td>
<td>1,017</td>
<td>Cellular nitrogen compound metabolic process**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0016567</td>
<td>0.001646</td>
<td>10.51</td>
<td>1</td>
<td>4</td>
<td>12</td>
<td>Protein ubiquitination**</td>
</tr>
</tbody>
</table>

NOTE.—Ontologies: MF, molecular function; BP, biological process; CC, cellular component.

aThree hundred twenty nine of 809 genes were annotated by GO terms.

bP-cut: the number of GO IDs with P values < 0.05 in the hypergeometric tests for each GO ID.

cExp: expected number of genes for the GO term in positively selected set if the function is not enriched.

dObs: observed number of genes in the positively selected set of genes.

eTotal: the total number of genes in the *T. pseudonana* genome annotated for a specific GO term.

*Significant with Bonferroni corrections per Ontology (MF = 1.4E-4; CC = 3.7E-4; BP = 9.6E-5).

**Significant with FDR of 0.05 using q-value correction.

Fig. 2. Boxplots and statistics of the expression of positively selected (gray) and neutral and purified (white) genes of *Thalassiosira pseudonana* (CCMP1335) grown in a nutrient replete control (ctrl), nutrient limitation (Si, Fe, NO₃, and CO₂), and a 4°C cold treatment. Boxplots: box = 1st and 3rd quartiles, whiskers = 1.5 × interquartile range, notch = median of expression as log₁₀(median aggregated probe intensity per gene). Mann–Whitney tests (N = 636 selected genes, N = 8,357 neutral and purified genes; *significance at P ≤ 0.01).
expression, respectively. The first is a group of 69 genes with consistently low expression across all treatments; some of these genes encode proteins putatively associated with sexual reproduction (fig. 3 and supplementary table S4, Supplementary Material online). The second group contains five genes, four of which are differentially upregulated in the silicic acid and iron-limited conditions relative to the control. One gene encodes a transcription factor, three appear to encode extracellular proteins including one with a chitin-binding domain, and the fifth has a transmembrane helix and possibly binds lectins (fig. 4).

Positive Selection along Specified Lineages of T. Pseudonana

The seven different T. pseudonana strains were grouped according to similarities in environmental conditions at the locations where they were collected to test whether those conditions promoted positive selection. Four strains were collected from locations where the SST is relatively stable year round, with a summer–winter difference of no more than 6°C (table 3). These strains have 5-fold more genes under positive selection than strains isolated from regions where the SST is seasonally more variable and the fluctuations are up to 13–15°C (table 3). Forty-one of the 121 positively selected genes associated with stable SSTs have GO terms. Eighteen (44%) of the 41 genes interact with other proteins, GO:0005515, and are statistically enriched after Bonferroni correction (GOSTATS: P value: 0.00012). Nine of the 22 positively selected genes associated with the more variable SSTs have GO terms; four of the encoded proteins have protein-binding activity, but none are related to heat stress.

The strain from the Pacific Gyre is the only representative from the open ocean, which is characterized by lower nutrient concentrations than coastal and estuarine areas. Two genes are positively selected in the Pacific Gyre strain, and neither has a GO term. The population of T. pseudonana from which the northern Adriatic Sea strain was collected is potentially geographically isolated from other populations because there are three hydrogeographic and genetic barriers for other species between it and the Atlantic Ocean. There are no positively selected genes within the Adriatic Sea strain (table 3).

Six of the seven T. pseudonana strains have been in culture for more than 25 years; therefore, we investigated the influences of culturing using a branch model that tested for positive selection upon strains based on the decade in which they were isolated (table 3). Time in culture does not appear to promote positive selection. There are 468 genes with a nominal P value ≤ 0.05, meaning that estimated omegas of the specified branches were significantly different from one, but they could be significantly greater or less than one. Twenty-five of the 468 genes were significant at a Bonferroni corrected value (P < 1.5 × 10⁻⁵), but none had an omega ≥ 1.0.

Discussion

Phylogenetic Distance at Which Positive Selection Is Best Detected in Diatoms

The greatest number of genes with a signal of positive selection occurs within a species, between two strains of T. pseudonana. Thalassiosira pseudonana diverged from D. confervacea, its sister species, ~2 Ma, based on a molecular clock applied to the divergence of 18S rDNA (Sorhannus

**Fig. 3.** Relative expression for 636 of 809 positively selected genes in Thalassiosira pseudonana (CCMP1335) grown under nutrient limitation (Si, Fe, NO₃, and CO₂), a nutrient replete control, and a 4°C cold treatment. Each gene is represented by one row in the heatmap. The similarity relationship of gene expression is represented by the dendrogram to the left. The upper dendrogram represents the similarity of responses among treatments. (A) Genes associated with sexual reproduction. (B) Putative cell wall-associated genes detailed in figure 4. Data are from Mock et al. 2008; log₁₀(median aggregated probe intensity per gene) is plotted.
The 18S rDNA sequences of *T. pseudonana* and *D. confervacea* differ by 0.4% over 1,758 base pairs (from accession numbers used in Sorhannus [2007]). Comparisons of diatoms with similar or equal sequence divergence should be tractable for identifying positively selected genes. Positive selection may also be detected at greater phylogenetic distances for particular genes. Many sexual reproduction genes evolve rapidly, and those involved in gamete recognition are implicated in maintaining reproductive isolation (Lyon and Vacquier 1999). The putative diatom gamete recognition gene, *SIG1*, is positively selected among four species of *Thalassiosira* (Sorhannus and Pond 2006). Selecting the appropriate phylogenetic distance to detect positively selected genes is, therefore, also contingent upon the functions of the proteins they encode.

The homologs detected between *Pseudo-nitzschia* species are currently subject to strong purifying selection, suggesting that their functions are conserved in other groups of organisms as well. Interestingly, the distribution of *dN*/*dS* in *Pseudo-nitzschia* homologs is of the same shape, and the rates of nonsynonymous mutations per codon are in the same range as genes shared by mice and humans that also have homologs in plants and yeast and have deep evolutionary histories of hundreds of millions of years (Alba and Castresana 2005). The apparent difference in the time required to accrue similar amounts of genetic variation,
10 My for _Pseudo-nitzschia_ species versus 100s of million years for metazoans and plants, is likely due to the high frequency of asexual reproduction in diatoms that provides the opportunity for mutation on a daily basis.

The cryptic sister species of _Ditylum_ have a 2-fold difference in genome size hypothesized to result from whole-genome duplication (Koester et al. 2010). Positively selected genes may be associated with speciation and niche differentiation reflecting the two species different adaptive strategies. The proportion of genes indicated to be under positive selection in _Ditylum_ was intermediate to the _Pseudo-nitzschia_ species and _Thalassiosira_ strain pairs. This may be because the _Ditylum_ species diverged only recently, and too little time has passed to identify selective substitutions.

**Candidate Genes under Positive Selection among Seven Strains of _T. pseudonana_**

Seven percent of the 11,390 known genes in _T. pseudonana_ are strong candidates for positive selection. This estimate is statistically conservative because it relies upon a Bonferroni correction rather than an FDR. One potential caveat is that this estimate may include both intra- and interstrain divergence if unidentified gene duplications were collapsed onto a single locus during genome mapping and assembly. In these instances, the signal of positive selection is accurate, but instead of applying to a single-locus alignment including sequence from seven strains, the alignment contains the consensus sequence of multiple loci for one or more strains.

Different organisms accrue positively selected genes at different rates, highlighting differences in selective pressures and mechanisms retaining beneficial mutations within populations. Both _Thalassiosira pseudonana_ and humans have ~7% of their loci under positive selection (Biswas and Akey 2006), but the length of time that humans have been diverging from their last common ancestor is two to three times longer than that between _T. pseudonana_ and _D. confervacea_. Two coral species _Acropora millepora_ and _A. palmata_ also have ~7% of their genomes under positive selection (Voolstra et al. 2011), but they diverged from one another 10–12 Ma (van Oppen et al. 2001). Approximating sexual generation times for diatoms, acroporid corals, and humans as 2, 4, and 20 years, respectively, these groups have accrued the same relative number of positively selected genes in $1.0 \times 10^6$, $2.8 \times 10^6$, and $3.0 \times 10^5$ generations. The intermediate proportion of positively selected genes within _T. pseudonana_ might be due to large population sizes of diatoms, within population allelic variability, gene flow between populations, and the strength and periodicity of selective forces (Lynch et al. 1991).

Versatility within the regulatory networks of gene expression provides a mechanism for cells to react quickly to environmental fluctuations (Li and Chen 2010). Positive selection in a single transcription factor potentially affects the expression of many genes at once. Transcription factors comprise ~2% of the protein-coding genes in _T. pseudonana_ (Rayko et al. 2010) and are the most specific group of overrepresented proteins encoded by the positively selected genes (table 2 and supplementary table S3, Supplementary Material online). Ten percent of the 258 transcription factors are positively selected in _T. pseudonana_ suggesting that differential gene regulation is an important component of adaptation for these strains. Positively selected genes were identified within the families of basic leucine zipper (bZip), heat shock, Myb, and zinc finger transcription factors. Transcription factors are also over-represented within the positively selected genes among human populations (Bustamante et al. 2005) and between two species of the nematode _Caenorhabditis_ (Castillo-Davis et al. 2004) but not between two closely related corals (Voolstra et al. 2011).

Responding to the molecular signals of biotic and abiotic stressors is within the purview of both bZip and heat shock transcription factors (HSFs); genes from both families also function under normal homeostatic physiologies. These two families are of special interest in diatoms because the size of membership differs from other organisms. There are 10-fold fewer bZip transcription factors in diatoms and their phylogenetic group, the stramenopiles, than in plants and animals (Montsant et al. 2007; Rayko et al. 2010). Yet, 20% of the genes in the bZip transcription family are under positive selection in _T. pseudonana_. Diatoms have an expanded family of HSFs, currently estimated to contain 94 members in contrast to the one to four HSFs found in yeast and metazoans (Montsant et al. 2007; Fujimoto and Nakai 2010; Rayko et al. 2010). Six of the 39 (~15%) heat shock factors in a diatom-only clade, Group 2, are positively selected (fig. 1 of Rayko et al. 2010). The Group 2 clade of HSFs highlights the importance of gene duplication in evolution and the potential for mutations in one of the duplicates to eventually provide a selective advantage (Zhang et al. 1998; Briscoe et al. 2010).

In _T. pseudonana_, 77% (112) of the proteins over-represented within any functional category are engaged in protein–protein interactions. The specific GO term over-represented in _T. pseudonana_ is not among those over-represented among positively selected genes in other organisms, but genes encoding proteins with related functions including signal transducers, and receptors for various proteins are over-represented in animals (Bustamante et al. 2005; Voolstra et al. 2011).

Lineage-specific, or orphan, genes are found throughout the evolutionary tree of life (Tautz and Domazet-Lošo 2011). These rapidly evolving genes experience relaxed levels of purifying selection; the increased numbers of nonsynonymous mutations may lead to novel protein functions and adaptive advantages (Domazet-Lošo and Tautz 2003; Voolstra et al. 2011). In addition, the creation of orphan genes provides a long-term source of genetic variability upon which selection may act. We define orphan genes of _T. pseudonana_ as those genes lacking any sequence homology to the genes of other organisms at an e-value cutoff of $10^{-5}$. There are proportionally more orphan genes in the positively selected set (24%) than in the genome as a whole (15%). These genes, frequently encoding proteins of unknown function, are important to conferring an adaptive advantage to individuals of
**Expression of Positively Selected Genes**

Levels of gene expression in yeast and vertebrates are inversely correlated to protein divergence and are hypothesized to exercise indirect control over mutation rates, such that highly expressed genes are the most conserved (Pál et al. 2001; Subramanian and Kumar 2004). Positively selected genes are expressed at lower levels than genes subject to neutral or purifying selection and tend to be expressed in restricted conditions or specific tissues in multicellular organisms (Kosiol et al. 2008). Positively selected genes in *T. pseudonana* CCMP1335 had significantly lower expression than more conserved genes across six different experimental treatments (fig. 2). Similar to other organisms, the adaptive genes in *T. pseudonana* are likely functioning in specialized capacities and enhance survival during suboptimal growth conditions.

Sexual reproduction genes are a classic example of genes with restricted expression and rapid rates of evolution in diverse organisms (Clark et al. 2006; Oliver et al. 2010). *Thalassiosira pseudonana* belongs to the multipolar diatoms that produce flagellated sperm, as do the centric diatoms. Genes encoding flagella-associated and putative sperm-activating proteins are under positive selection in *T. pseudonana* and are among the genes with the lowest levels of expression among six experimental growth conditions (fig. 3 and Mock et al. 2008). This pattern of expression for genes involved in sexual reproduction is consistent with observations that there were no cells differentiating into gametangia in these experiments (Mock et al. 2008). Sexual reproduction is not yet documented in *T. pseudonana*, but it is an obligate and episodic phase of the life history of many diatoms (Chepurnov et al. 2004). The signature of positive selection suggests that these genes are functional; otherwise, they would be expected to degrade into nonfunctional pseudogenes.

Coexpression of multiple genes suggests that their protein products are functioning in the same metabolic pathway or that they are affecting the same phenotype. The environmental conditions in which the genes are expressed provide additional information when the function of their encoded proteins is unknown. One of the most obvious patterns of coexpression in *T. pseudonana* CCMP1335 was associated with five positively selected genes that were up-regulated when the cells were limited for silicon and iron (fig. 4, Mock et al. 2008). The first gene in this group is a MYB transcription factor that also is highly expressed under other conditions, similar to expression patterns of MYB genes in plants where they respond to environmental stress and regulate growth and development (Yanhui et al. 2006). Two of the five genes encode proteins that appear to interact with chitin: protein 12594 possesses a chitin-binding domain and a signal peptide suggesting that it is secreted and protein 21085 does not have a signal peptide, but it does have a transmembrane domain and it is a member of a superfamily of proteins that bind lectins, suggesting that it may also interact with chitin. The fourth member of this cluster encodes a protein (21587) with a transmembrane domain but possesses no other functional annotation. The fifth gene contains a V5/Tpx-1 domain, commonly found in extracellular sensory proteins that recognize signals and proteins from other organisms that are frequently pathogenic (Cantacessi et al. 2009). Gene expression of this cluster is associated with a distinctive morphology induced by both silicon and iron limitation—an elongated, bent-cell phenotype in which chitin is deposited in the girdle region of the wall (Durkin et al. 2009). Healthy cells have tight connections between silica cell wall components; therefore, secretion of chitin-binding proteins is hypothesized to shore up the silica cell wall when those connections are weakened and vulnerable to pathogenic attack (Davis et al. 2005; Durkin et al. 2009). The pattern of expression and putative localization of these five proteins provides a possible link between selective agents of nutrient limitation, functional genes, and a cell wall phenotype.

**Selection of Specified Strains with Respect to Environment**

We used branch-site models of positive selection to explore the hypothesized selective forces of geographic isolation, ocean environment, and temperature on subsets of the strains collected from sites associated with these environmental variables (table 3). Three hydrogeographic barriers separate the Adriatic Sea from the Atlantic Ocean, and these boundaries differentiate population structure in planktonic organisms living in the Mediterranean Sea (Patarnello et al. 2007; Yebra et al. 2011). Although the *T. pseudonana* Adriatic strain is potentially geographically isolated, there do not appear to be any differentiating environmental pressures. Instead, the genes of the Adriatic strain are evolving under purifying and neutral selection.

The open ocean strain from the Pacific Gyre was collected from an environment very different than the other six estuarine strains, including the one from the Adriatic. The gyre has stable light, salinity, and low nutrients compared with coastal regions and estuaries, which are characterized by dynamic fluctuations of those same parameters (Karl 1999). Therefore, we hypothesized that the open ocean strain would have a strong genetic signal of differential adaptation to its unique environment. Surprisingly, only two genes were identified as positively selected in the open ocean strain. Both encode predicted proteins with unknown functions.

Temperature can be a strong selective agent (Husby et al. 2011), and we tested two hypotheses associated with seasonal variation in SSTs. The four strains collected from locations...
where summer–winter SSTs varied by 6°C, or less, shared 121 positively selected genes. The presence of three heat shock factors, one heat shock protein, and over-represented protein-binding proteins within this set of positively selected genes suggests adaptation to stress conditions. Fewer positively selected genes are shared by the three strains collected from regions where there is a strong seasonal fluctuation in temperature, but one gene is notable because it encodes a UV radiation resistance protein. This protein may not be directly affected by temperature, but it would be affected by available sunlight altering the temperature of water. These data provide circumstantial evidence that seasonal temperature variation may act as a selective agent; however, it is more likely that absolute temperature, the timing of seasonal changes, and interactions between available light and temperature are more effective in selecting fit phenotypes (e.g., Namroud et al. 2008).

In the two previous sections, we explore associations between genotype and selective force. It is important to note that these strains have been in culture for decades (table 3). Although there is no evidence that culturing promotes positive selection, it is unlikely that any strain retains its original genotype; therefore, care must be taken in interpreting the results. Our results are robust in two ways: first the most tractable genes for further study tend to be grouped in the expression profile or are shared by a subset of strains from the branch-site models, and second, the majority of positively selected genes detected by the branch-site models are in the original set of 809 genes.

Together, our results present an opportunity to test hypotheses that integrate positively selected genes in T. pseudonana with their associated phenotypes and selective forces through manipulative laboratory experiments and in the field by taking a population genetics approach to determine allele distributions for populations living in different regions.

Supplementary Material
Supplementary figure S1 and tables S1–S4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
The authors thank Renee George, Megan Kogut, Bob Morris, and Larry Ruzzo for insightful discussion and Robin Kodner for assistance with the T. pseudonana consensus tree. Special thanks are given to those who shared data sets. The T. pseudonana data set was constructed and curated by Rhonda Morales, Ellen Lin, Vaughn Iverson, Dave Schruth, and Chris Berthiaume. The Pseudo-nitzschia multiseries and P. australis data sets are courtesy of Micaela S. Parker. Uwe John and Wiebe Kooistra provided the Pseudo-nitzschia multistriata data set, and M. Valeria Ruggiero provided the P. multistriata 18S sequence. This work was supported by the Gordon and Betty Moore Foundation Marine Microbiology Investigator Award to E.V.A., and by the National Institutes of Health HD057974 and the National Science Foundation’s Division of Environmental Biology 0918106 to W.J.S.

References

432


