Comprehensive and Functional Analysis of Horizontal Gene Transfer Events in Diatoms

Emmelien Vancaester,[1,2,3] Thomas Depuydt,[1,2,3] Cristina Maria Osuna-Cruz,[1,2,3] and Klaas Vandepoele*,[1,2,3]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[2]VIB Center for Plant Systems Biology, Ghent, Belgium

[3]Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium

***Corresponding author:** E-mail: klpoe@psb.vib-ugent.be.

Associate editor: Fabia Ursula Battistuzzi

Abstract

Diatoms are a diverse group of mainly photosynthetic algae, responsible for 20% of worldwide oxygen production, which can rapidly respond to favorable conditions and often outcompete other phytoplankton. We investigated the contribution of horizontal gene transfer (HGT) to its ecological success. A large-scale phylogeny-based prokaryotic HGT detection procedure across nine sequenced diatoms showed that 3–5% of their proteome has a horizontal origin and a large influx occurred at the ancestor of diatoms. More than 90% of HGT genes are expressed, and species-specific HGT genes in *Phaeodactylum tricornutum* undergo strong purifying selection. Genes derived from HGT are implicated in several processes including environmental sensing and expand the metabolic toolbox. Cobalamin (vitamin B12) is an essential cofactor for roughly half of the diatoms and is only produced by bacteria. Five consecutive genes involved in the final synthesis of the cobalamin biosynthetic pathway, which could function as scavenging and repair genes, were detected as HGT. The full suite of these genes was detected in the cold-adapted diatom *Fragilariopsis cylindrus*. This might give diatoms originating from the Southern Ocean, a region typically depleted in cobalamin, a competitive advantage. Overall, we show that HGT is a prevalent mechanism that is actively used in diatoms to expand its adaptive capabilities.

***Key words:* horizontal gene transfer, diatoms, vitamin B12, selection pressure.**

Introduction

Horizontal, also dubbed lateral, gene transfer (HGT) is the transfer of genetic information between reproductively isolated species by a route other than direct exchange from parent to progeny. Although HGT events are widespread and well documented among prokaryotes, they are much rarer in eukaryotes. Nevertheless, recently, several examples of HGT from archaea or bacteria into eukaryotes have been reported. Functional HGT events have been described for almost all unicellular eukaryotic lineages, including fungi (Dean et al. 2018; Gonçalves et al. 2018), extremophilic red algae (Schönknecht et al. 2013), green algae (Krasovec et al. 2018), rumen-associated ciliates (Ricard et al. 2006), oomycetes (Savory et al. 2015), and photosynthetic diatoms (Bowler et al. 2008; Marchetti et al. 2009). Next to events involving the maintenance of pre-existing functions, which occur mainly in endosymbiotic relationships, innovative events have been described, which provide the recipient with new functions or an altered phenotype (Husnik and McCutcheon 2018). Although the uptake of genetic material happens by chance, fixation does not, making HGT predominantly important in the following processes: 1) the alteration of iron uptake and metabolism (Marchetti et al. 2009; Tsaousis et al. 2012; Kominek et al. 2019), 2) adaptation to an anaerobic lifestyle (Stairs et al. 2011, 2018), 3) nucleotide import and synthesis (Alexander et al. 2016; Dean et al. 2018), 4) novel defense mechanisms (Strese et al. 2014; Chou et al. 2015), 5) mechanisms to cope with stressors such as salt (Harding et al. 2017; Foflonker et al. 2018), temperature (Krasovec et al. 2018), and heavy-metal concentrations (Schönknecht et al. 2013), and 6) expansion of metabolic capacities (Ricard et al. 2006; Savory et al. 2015; Gonçalves et al. 2018).

Diatoms (Bacillariophyta) are one of the most abundant and species-rich groups of phytoplankton and release between 20% and 25% of the global amount of oxygen (Field et al. 1998). They can rapidly adapt to local conditions, outcompete other photosynthetic eukaryotes, and dominate oceanic spring blooms, as long as silicon is not limited (Winder and Cloern 2010). Moreover, they are found throughout every aquatic photic zone of this planet, such as oceans, intertidal zones, freshwater bodies, soil and even ice ecosystems (Janech et al. 2006). Molecular clock evidence suggests that diatoms emerged between 225 and 200 million years ago (Nakov et al. 2018), and their origin may be related to the end-Permian mass extinction that occurred around

250 million years ago. In the early Cretaceous, between 150 and 130 million years ago, diatoms split into centric and pennate lineages. Several whole-genome sequences of representatives from polar centrics (*Thalassiosira pseudonana* [Armbrust et al. 2004], *T. oceanica* [Lommer et al. 2012], *Cyclotella cryptica* [Traller et al. 2016]), araphid pennates (*Synedra acus* [Galachyants et al. 2015]), and raphid pennates (*Phaeodactylum tricornutum* [Bowler et al. 2008; Rastogi et al. 2018], *Seminavis robusta* [Osuna-Cruz et al. 2020], *Fistulifera solaris* [Tanaka et al. 2015], *Fragilariopsis cylindrus* [Mock et al. 2017], *Pseudo-nitzschia multistriata* [Basu et al. 2017]) have become available in recent years, which allows the analysis of the evolutionary history within diatoms. It is not fully understood how HGT has contributed to the ecological success of this environmentally important group of organisms. Moreover, diatoms harbor complex red algal plastids taken up after several endosymbiotic events and therefore endosymbiotic gene transfer (EGT) has also contributed to their mosaic genetic setup.

Although HGT detection has been previously performed in diatoms within the context of genome projects (Bowler et al. 2008; Lommer et al. 2012; Traller et al. 2016; Basu et al. 2017; Rastogi et al. 2018), they were based on different methodologies and criteria and are therefore not directly comparable. Although some studies used phylogenetics (Bowler et al. 2008; Basu et al. 2017), others relied purely on sequence homology searches (Lommer et al. 2012; Traller et al. 2016; Rastogi et al. 2018). In this study, we sought to phylogenetically detect HGT events simultaneously across all sequenced diatoms. We delineated genes from horizontal descent using a high-throughput gene family phylogenetics-based approach, which allows dating transfer events. Here, we explore the functional bias of HGT genes in diatoms and for the first time gain insight into their expression dynamics and patterns of selection.

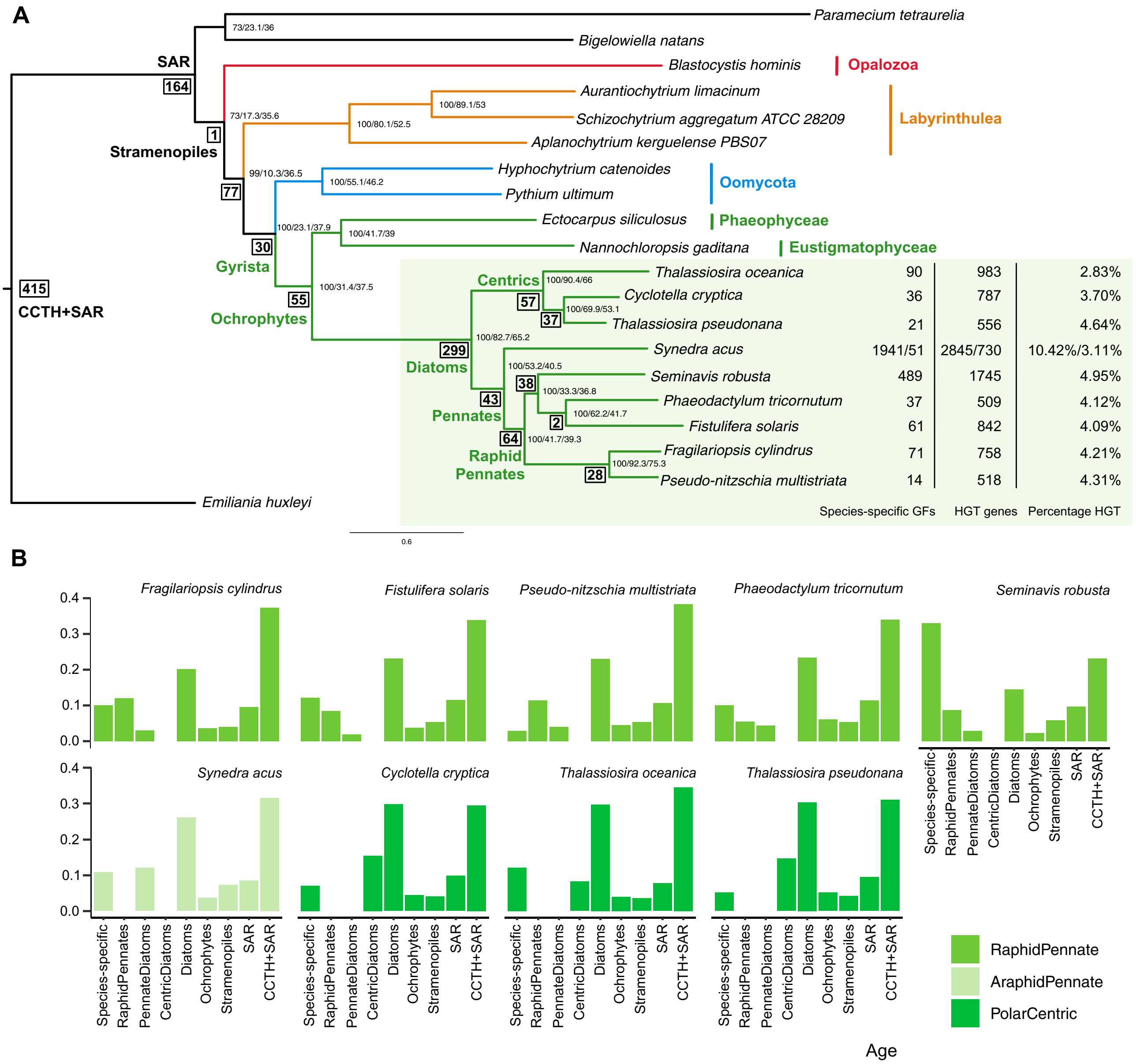
Results

Detection and Phylogenetic Distribution of Diatom HGT Candidates

Twenty unicellular eukaryotic species (supplementary table S1, Supplementary Material online) were selected to deduce the contribution of prokaryotic-derived HGT. All protein-coding genes from these 20 eukaryotic species (fig. 1*a*) were clustered in 145,601 gene families, of which 32% are genes that encode proteins that lack similarity to any other protein in this data set. Moreover, a similarity search was performed against the nonredundant database of NCBI to detect prokaryotic homologs to these proteins. After adding prokaryotic homologs to the gene families, the identification of prokaryotic-to-eukaryotic HGT was achieved by building a phylogenetic tree for 8,476 gene families containing both eukaryotic and prokaryotic proteins. A graphical summary of the methodology can be found in supplementary figure S1, Supplementary Material online. The species topology of these 20 unicellular eukaryotes was also constructed, both based on single-locus trees and a concatenation-based approach of 156 near-single copy gene families (138,948 amino acids) (fig. 1*a*), with the haptophyte *Emiliania huxleyi* as an outgroup. Having the species-level phylogeny available allows for the dating of HGT candidates.

To avoid the misclassification of contaminating DNA present in the genome assembly as genomic regions originating by HGT, several quality analyses were performed. The guidelines proposed by Richards and Monier (2016) were followed to exclude incorrect inference of HGT. Therefore, the gene origin was determined by phylogenetic tree construction followed by inspection of species-specific HGT genes. Also, the percentage GC and the integration of HGT genes across chromosomes were assessed. First, the fraction of species-specific HGT was compared among all diatoms. More than 75% (2,146/2,844) of the predicted HGT genes in *S. acus* were only detected in this genome; whereas in all other species, this fraction was drastically lower (11.58 ± 9.25%) (fig. 1*b*). A donor analysis of these genes revealed that many were derived from *Sphingomonas* sp., which has been described to be associated with *S. acus* in culture (Zakharova et al. 2010). Contigs flagged to be contaminant based on a nucleotide sequence similarity search against all available Sphingomonadales genomes were clearly separable from *S. acus* based on their significantly lower percentage GC (fig. 2*a*) (42.1% vs. 63.3%, $P < 2 \times 10^{-16}$). Therefore, all 695 nuclear contigs having a GC content above 50% were removed, reducing the nuclear *S. acus* genome size by 4–94.38 Mb and retaining 23,719 genes. Interestingly, the HGT detection procedure succeeded in both flagging the contaminant and detecting HGT events in the *S. acus* genome (fig. 2*b*). Despite the fact that in several other diatoms the GC content was significantly different between genes from horizontal and vertical descent, the mean difference never exceeds two percentage points (supplementary fig. S2, Supplementary Material online).

Next, the enrichment of HGTs per contig or chromosome was evaluated to assess whether certain regions are derived from contamination, yielding no clear examples of clustering of HGT genes on specific genomic locations. The distribution of HGTs across the chromosome-level genomes of *P. tricornutum* and *T. pseudonana* is plotted in supplementary figure S3, Supplementary Material online, and shows an unbiased distribution of HGT genes. Moreover, reads derived from the third-generation sequencing were available in four diatoms, to assess the degree of contamination. All reads spanning at least two genes of which one was predicted to have a horizontal origin also encompassed vertical genes in *P. tricornutum* and *T. pseudonana*, whereas this was the case for 99.54% and 99.73% reads in, respectively, *F. cylindrus and S. robusta*. Thus, no prokaryotic contamination in long-read data was detected for four diatoms (supplementary figs. S4, Supplementary Material online). As it has been proposed that the transfer of transposable elements (TEs) could be associated with facilitating gene transfer (Keeling and Palmer 2008), the distance between every gene and its closest TE was calculated in *P. tricornutum*. Species-specific HGT genes were significantly closer to TEs ($P = 1.6 \times 10^{-03}$), whereas the same was also true for vertically descended species-specific genes ($P = 2.7 \times 10^{-14}$). This suggests that novel genes are



	Species-specific GFs	HGT genes	Percentage HGT
Thalassiosira oceanica	90	983	2.83%
Cyclotella cryptica	36	787	3.70%
Thalassiosira pseudonana	21	556	4.64%
Synedra acus	1941/51	2845/730	10.42%/3.11%
Seminavis robusta	489	1745	4.95%
Phaeodactylum tricornutum	37	509	4.12%
Fistulifera solaris	61	842	4.09%
Fragilariopsis cylindrus	71	758	4.21%
Pseudo-nitzschia multistriata	14	518	4.31%

FIG 1. Overview of HGT events across diatoms. (*a*) Species phylogeny determined by IQ-Tree with values at the internal nodes that denote bootstrap, gene- and site-concordance factors respectively. Branches are colored according to their phylogenetic classification. Bold and framed values at internal nodes reflect the number of predicted HGT events. For diatoms, the number of species-specific gene families (GFs) of HGT origin, the total number of HGT genes and its fraction of the proteome to be originating from HGT is tabularized. For *Synedra acus*, the number of HGT genes both prior and after removal of contamination is mentioned. (*b*) Distribution of the age classes of HGTs across the nine investigated diatom species. SAR is the clade comprising Stramenopiles, Alveolates, and Rhizaria, whereas CCTH consists of Cryptophyta, Centrohelida, Telonemia, and Haptophyta and CCTH + SAR denotes the ancestor of both groups.

more likely to integrate and become fixed close to repetitive regions.

Except for a fraction of genes in the *S. acus* data set, we could not identify genomic properties indicating that the identified HGT genes are caused by contamination. In total, 7,415 diatom genes were defined as having HGT origin, covering 1,963 gene families. This reflects 509 to 1,745 genes per species, making 3%–5% of the diatom gene repertoire predicted to be HGT (fig. 1*a*). This is similar to previous phylogenetic-based estimations of HGT content in diatoms, which ranged from 587 genes (4.8%) in *P. tricornutum* (Bowler et al. 2008) to 438 in *P. multistriata* (Basu et al. 2017) (3.6%) and is slightly higher than what was reported in the anaerobic gut parasite *Blastocystis hominis* (Eme et al. 2017) (2.5%), where next to bacterial HGT also other transfers were described. The lower frequency in this stramenopile could be due to its constrained and reduced genome size as a result of its parasitic lifestyle. On average, an HGT gene family consisted of 3.76 diatom genes and 2.55 diatom species. In total, only 106 HGT families were present in all nine diatoms. For 69 gene families, the HGT copies were significantly expanded in at least one species, of which notably 26 and 21 gene families

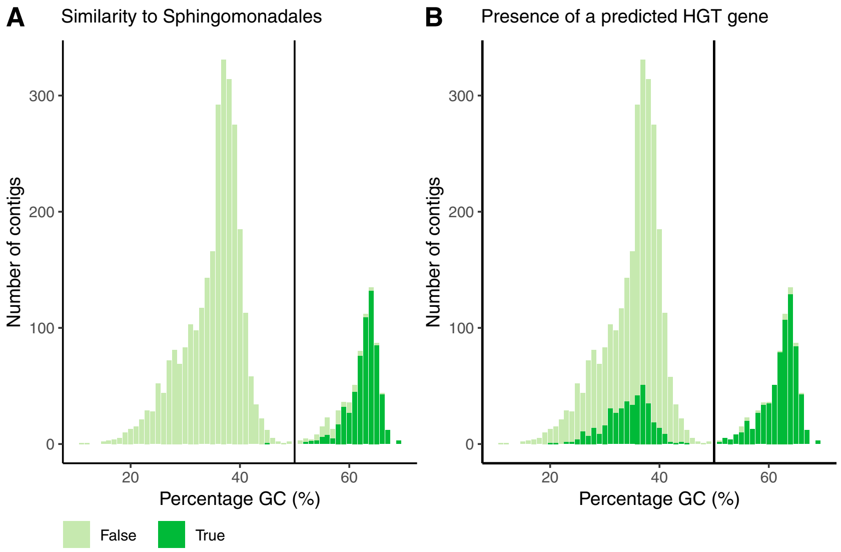


Fig. 2. Contamination of Sphingomonadales in genome of *Synedra acus*. (*a*) Percentage GC of *S. acus* contigs versus nucleotide sequence similarity to Sphingomonadales based on at least 70% identity and 25% alignment coverage. (*b*) Percentage GC of *S. acus* contigs versus presence of an HGT candidate within each contig.

were expanded respectively in *S. robusta* and *S. acus* (supplementary table S2, Supplementary Material online). Indeed, gene family expansion by duplication has been observed before following HGT integration in eukaryotes (Dean et al. 2018; Murphy et al. 2019) and this could be a strategy to diversify the original acquired function.

The age of all gene families of vertical descent was determined based on the last common ancestor (LCA) of the observed species. Similarly, the most likely time point of integration for every HGT was determined using the species composition of the acceptor branch in the phylogenetic tree (fig. 1*a*). The large number of HGT gene families that can be attributed to the ancestor of diatoms is striking, ranging from 15% in *S. robusta* to 30% in *T. pseudonana* (fig. 1*b*). Another study (Rastogi et al. 2018) also detected a continuous flux of genes from prokaryotes during the evolutionary history of *P. tricornutum*. However, they claimed that most influx occurred at the ancestor of the photosynthetic Stramenopiles (Ochrophytes), whereas our results indicate that this happened more recently in the ancestor of the diatom clade.

Finally, several structural gene features were evaluated according to their mode of inheritance. The coding gene length of vertically descended species-specific genes in all diatom species was significantly shorter compared with all other genes ($P < 2 \times 10^{-16}$) and significantly shorter to the species-specific HGT genes in all diatoms, except for *T. pseudonana* (supplementary fig. S5, Supplementary Material online). In yeast, it has also been observed that *de novo* genes were on average shorter than conserved and horizontally transferred genes(Vakirlis et al. 2018). Species-specific HGT genes, on the other hand, were significantly shorter to all other genes in *C. cryptica* ($P = 2.1 \times 10^{-02}$) and *S. robusta* ($P = 1.4 \times 10^{-05}$). Given that introns are a typical eukaryotic gene feature, HGT genes are expected to have a shorter total intron length, especially for recent acquisitions as HGT genes adapt to their recipient genome. The intron length of HGT genes was significantly shorter in several pennate diatoms (*F. solaris*: 1.8×10^{-03}, *P. tricornutum*: 3.4×10^{-02}, *S. robusta*: 1.1×10^{-07} and *S. acus*: 4.6×10^{-03}) and for several diatoms the young species-specific HGT genes had shorter introns than the rest of the gene repertoire (*F. cylindrus*: 1.1×10^{-03}, *F. solaris*: 9.8×10^{-03}, *P. tricornutum*: 1.1×10^{-02}, *S. robusta*: 2×10^{-09} and *C. cryptica*: 4.9×10^{-02}). These results indicate that introns become an emerging property of HGT genes after integration.

The Functional Landscape of Diatom HGT Genes

To gain insight in the functional repertoire of HGT genes, a gene ontology (GO) and functional domain (Interpro) enrichment was performed. Out of the 7,415 diatom HGT genes, 6,024 (81%) were annotated with an Interpro domain and 3,893 (52%) with a GO term. The only GO term that was enriched for HGT genes in all nine diatom species is pseudouridine synthesis (GO: 0001522), whereas enriched protein domains covered pseudouridine synthase (IPR006145), S-adenosyl-L-methionine–dependent methyltransferase (IPR029063) and nitroreductase (IPR029479). An overview of the enriched functional categories across different ages can be found in supplementary figures S6 and S7, Supplementary Material online. A more in-depth exploration of several functional categories is given in supplementary note 1, Supplementary Material online, whereas an overview of all discussed functions and their corresponding gene families can be found in supplementary table S3, Supplementary Material online.

Cobalamin Uptake

Cobalamin (vitamin B12) is a complex molecule composed of a central cobalt-containing corrin ring, a lower ligand of 5,6-dimethylbenzimidizole (DMB) and an upper axial ligand that can be a hydroxy-, cyano-, methyl, or adenosyl group (fig. 3*b*). Vitamin B12 acts as a coenzyme in three enzymes in

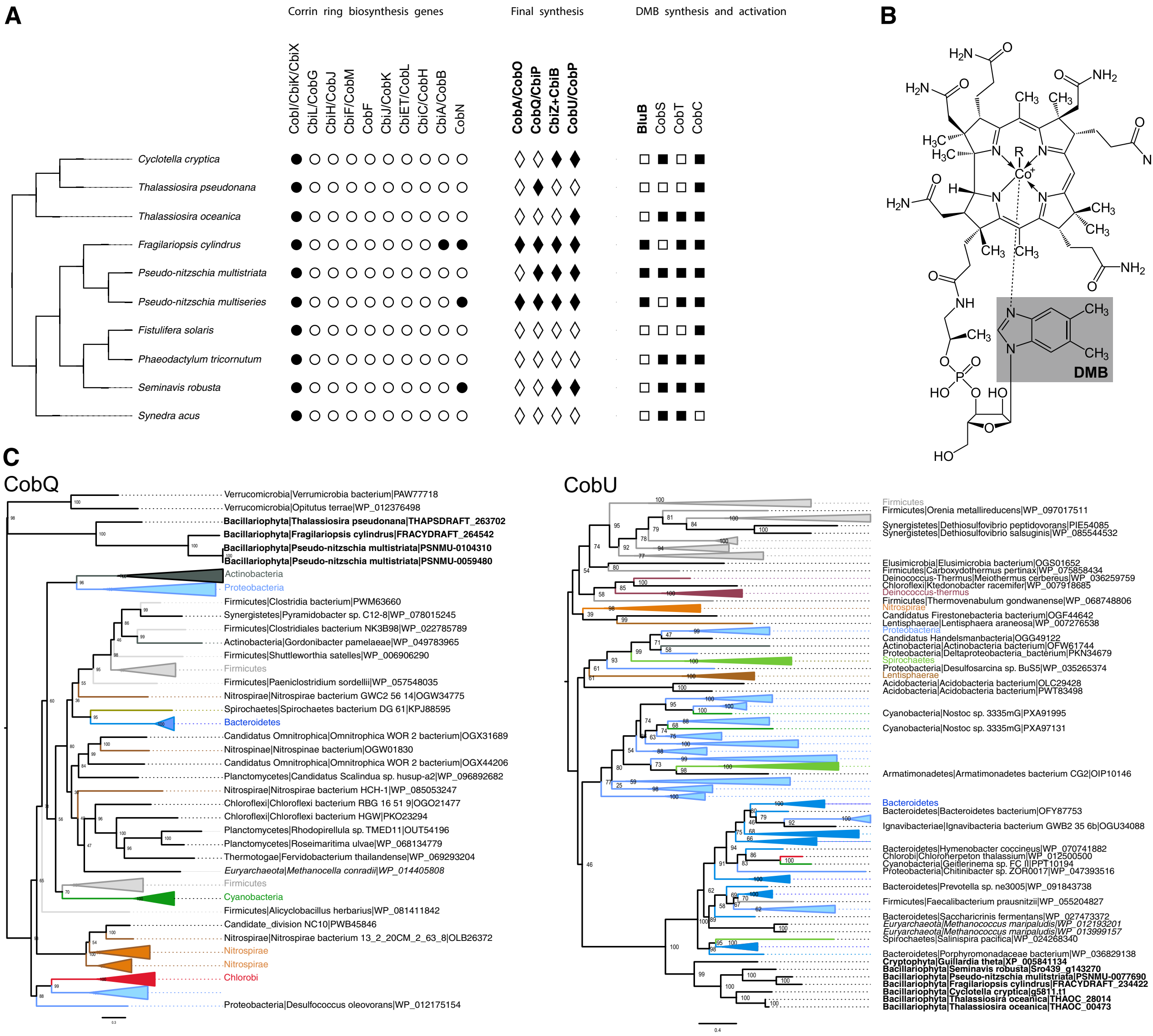

A
Corrin ring biosynthesis genes
Final synthesis
DMB synthesis and activation
CobI/CbiK/CbiX
CbiL/CobG
CbiH/CobJ
CbiF/CobM
CobF
CbiJ/CobK
CbiET/CobL
CbiC/CobH
CbiA/CobB
CobN
CobA/CobO
CobQ/CbiP
CbiZ+CbiB
CobU/CobP
BluB
CobS
CobT
CobC
Cyclotella cryptica
Thalassiosira pseudonana
Thalassiosira oceanica
Fragilariopsis cylindrus
Pseudo-nitzschia multistriata
Pseudo-nitzschia multiseries
Fistulifera solaris
Phaeodactylum tricornutum
Seminavis robusta
Synedra acus
B
H_2N
NH_2
CH_3
H_3C
R
Co^+
N
H
NH
O
P
HO
DMB
C
CobQ
Verrucomicrobia|Verrumicrobia bacterium|PAW77718
Verrucomicrobia|Opitutus terrae|WP_012376498
Bacillariophyta|Thalassiosira pseudonana|THAPSDRAFT_263702
Bacillariophyta|Fragilariopsis cylindrus|FRACYDRAFT_264542
Bacillariophyta|Pseudo-nitzschia multistriata|PSNMU-0104310
Bacillariophyta|Pseudo-nitzschia multistriata|PSNMU-0059480
Actinobacteria
Proteobacteria
Firmicutes|Clostridia bacterium|PWM63660
Synergistetes|Pyramidobacter sp. C12-8|WP_078015245
Firmicutes|Clostridiales bacterium NK3B98|WP_022785789
Actinobacteria|Gordonibacter pamelaeae|WP_049783965
Firmicutes|Shuttleworthia satelles|WP_006906290
Firmicutes
Firmicutes|Paeniclostridium sordellii|WP_057548035
Nitrospirae|Nitrospirae bacterium GWC2 56 14|OGW34775
Spirochaetes|Spirochaetes bacterium DG 61|KPJ88595
Bacteroidetes
Candidatus Omnitrophica|Omnitrophica WOR 2 bacterium|OGX31689
Nitrospinae|Nitrospinae bacterium|OGW01830
Candidatus Omnitrophica|Omnitrophica WOR 2 bacterium|OGX44206
Planctomycetes|Candidatus Scalindua sp. husup-a2|WP_096892682
Nitrospirae|Nitrospirae bacterium HCH-1|WP_085053247
Chloroflexi|Chloroflexi bacterium RBG 16 51 9|OGO21477
Chloroflexi|Chloroflexi bacterium HGW|PKO23294
Planctomycetes|Rhodopirellula sp. TMED11|OUT54196
Planctomycetes|Roseimaritima ulvae|WP_068134779
Thermotogae|Fervidobacterium thailandense|WP_069293204
Euryarchaeota|Methanocella conradii|WP_014405808
Firmicutes
Cyanobacteria
Firmicutes|Alicyclobacillus herbarius|WP_081411842
Candidate_division NC10|PWB45846
Nitrospirae|Nitrospirae bacterium 13_2_20CM_2_63_8|OLB26372
Nitrospirae
Nitrospirae
Chlorobi
Proteobacteria|Desulfococcus oleovorans|WP_012175154
CobU
Firmicutes
Firmicutes|Orenia metallireducens|WP_097017511
Synergistetes|Dethiosulfovibrio peptidovorans|PIE54085
Synergistetes|Dethiosulfovibrio salsuginis|WP_085544532
Elusimicrobia|Elusimicrobia bacterium|OGS01652
Firmicutes|Carboxydothermus pertinax|WP_075858434
Deinococcus-Thermus|Meiothermus cerbereus|WP_036259759
Chloroflexi|Ktedonobacter racemifer|WP_007918685
Deinococcus-thermus
Firmicutes|Thermovenabulum gondwanense|WP_068748806
Nitrospirae
Candidatus Firestonebacteria bacterium|OGF44642
Lentisphaerae|Lentisphaera araneosa|WP_007276538
Proteobacteria
Candidatus Handelsmanbacteria|OGG49122
Actinobacteria|Actinobacteria bacterium|OFW61744
Proteobacteria|Deltaproteobacteria_bacterium|PKN34679
Spirochaetes
Proteobacteria|Desulfosarcina sp. BuS5|WP_035265374
Lentisphaerae
Acidobacteria|Acidobacteria bacterium|OLC29428
Acidobacteria|Acidobacteria bacterium|PWT83498
Cyanobacteria|Nostoc sp. 3335mG|PXA91995
Cyanobacteria|Nostoc sp. 3335mG|PXA97131
Armatimonadetes|Armatimonadetes bacterium CG2|OIP10146
Bacteroidetes
Bacteroidetes|Bacteroidetes bacterium|OFY87753
Ignavibacteriae|Ignavibacteria bacterium GWB2 35 6b|OGU34088
Bacteroidetes|Hymenobacter coccineus|WP_070741882
Chlorobi|Chloroherpeton thalassium|WP_012500500
Cyanobacteria|Geitlerinema sp. FC II|PPT10194
Proteobacteria|Chitinibacter sp. ZOR0017|WP_047393516
Bacteroidetes|Prevotella sp. ne3005|WP_091843738
Firmicutes|Faecalibacterium prausnitzii|WP_055204827
Bacteroidetes|Saccharicrinis fermentans|WP_027473372
Euryarchaeota|Methanococcus maripaludis|WP_012193201
Euryarchaeota|Methanococcus maripaludis|WP_013999157
Spirochaetes|Salinispira pacifica|WP_024268340
Bacteroidetes|Porphyromonadaceae bacterium|WP_036829138
Cryptophyta|Guillardia theta|XP_005841134
Bacillariophyta|Seminavis robusta|Sro439_g143270
Bacillariophyta|Pseudo-nitzschia mulitstriata|PSNMU-0077690
Bacillariophyta|Fragilariopsis cylindrus|FRACYDRAFT_234422
Bacillariophyta|Cyclotella cryptica|g5811.t1
Bacillariophyta|Thalassiosira oceanica|THAOC_28014
Bacillariophyta|Thalassiosira oceanica|THAOC_00473
0.4

Fig. 3. Cobalamin pathway in diatoms. (*a*) Overview of the cobalamin biosynthesis pathway and its presence in diatoms. Genes in bold are of horizontal descent and the presence of a gene is displayed by a filled circle, diamond, or square depending on its position in the pathway. (*b*) Chemical structure of cobalamin, where the lower ligand DMB is emphasized in gray. (*c*) Phylogenetic trees of two genes of horizontal descent in diatoms *CobQ* and *CobU*.

eukaryotes: methylmalonyl-CoA mutase, type II ribonucleotide reductase, and methionine synthase (MetH). More than half of the algal species surveyed (171/326) (Croft et al. 2005) are auxotrophic for vitamin B12, including 37 out of 58 diatoms. Nonetheless, *de novo* synthesis has only been described to occur in prokaryotes. Therefore, cobalamin availability alters the composition of marine phytoplankton communities (Bertrand et al. 2015). The exchange of cobalamin in return for organic compounds is believed to underpin the close mutualistic interactions between heterotrophic bacteria and auxotrophic algae (Heal et al. 2017). Some algae maintain a cobalamin-independent methionine synthase (MetE), although it is anticipated that this enzyme has a 50–100-fold lower turnover rate compared with MetH (Bertrand et al. 2013). A correlation was detected between the scattered phylogenetic pattern of absence of an *MetE* and auxotrophy for this vitamin (Helliwell et al. 2011). It has been suggested that this loss has a biogeographical basis as there is a tendency for diatoms occurring in the Southern Ocean to retain *MetE* more often (Ellis et al. 2017). Moreover, it has been recently shown that cyanobacteria produce the chemical variant pseudo-cobalamin, where adenine substitutes DMB as the lower ligand, which is less bioavailable to eukaryotic algae (Helliwell et al. 2016). However, some species, including *P. tricornutum* and *E. huxleyi* (Helliwell 2017), can remodel this to cobalamin using CobT, CobS, and CobC via the nucleotide loop assembly (Helliwell et al. 2016; Heal et al. 2017). Here *BluB*, necessary for DMB production (Campbell et al. 2006), was detected to have originated by HGT from alphaproteobacteria in *F. cylindrus*, *P. multistriata*, and *P. multiseries* (fig. 3*a*). More than 90% of the cobalamin-producing alpha- and gammaproteobacteria encode *BluB* (Heal et al. 2017).

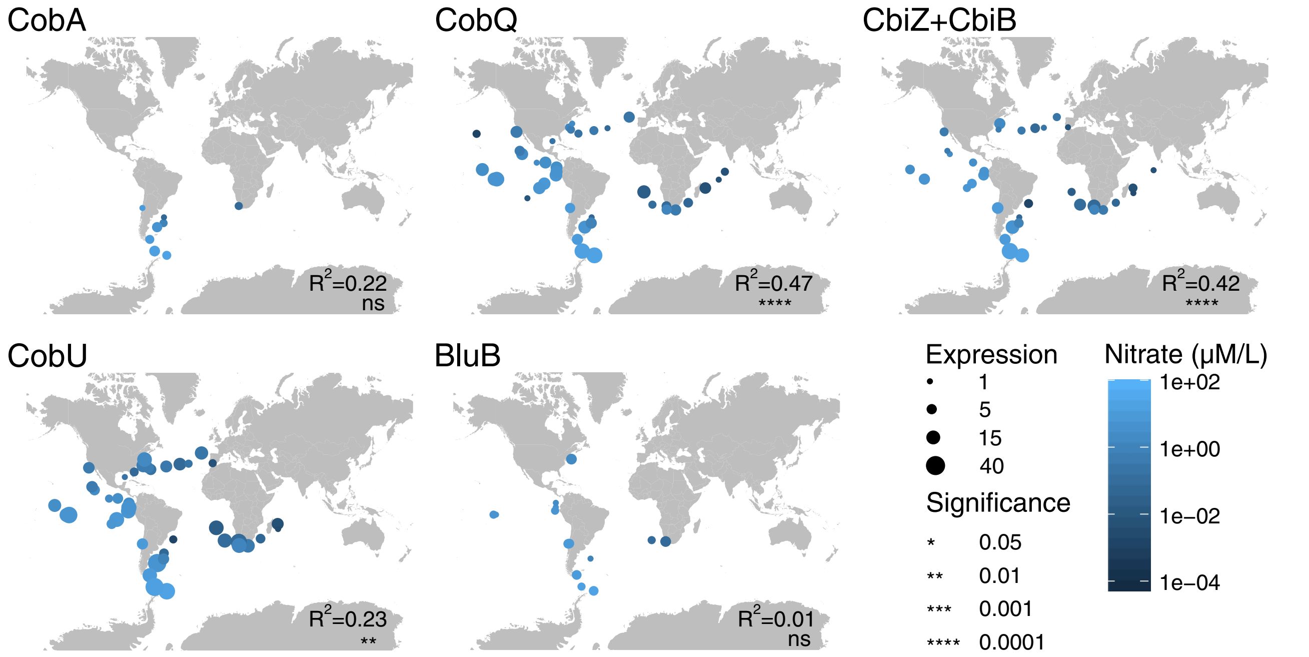


FIG. 4. Expression levels of diatom cobalamin genes in the ocean. Expressed number of diatom sequences of several HGT genes involved in the cobalamin pathway across stations sampled worldwide during the TARA Oceans project in the surface layer, colored according to their nitrate concentration.

Moreover, five HGT genes were detected in the final synthesis of the cobalamin biosynthesis pathway, which can also function as scavenging and repair genes: *CobN*, *CobA*/*CobO*, *CobQ*/*CbiP*, *CobD*/*CbiB*, and *CobU*/*CobP* (fig.3*a* and *c*). These genes were previously also detected in diatom metatranscriptomes and *P. granii* (Cohen et al. 2017, 2018), where *CobN*, *CobS*, and *CobU* were more highly expressed under iron replete conditions. Interestingly, for all diatoms, except for *C. cryptica*, the *CbiB* gene also contains the *CbiZ* domain, which is involved in the removal of the lower ligand (Helliwell 2017). Only *F. cylindrus* and *P. multiseries* contain the full suite of these detected HGT genes in their cobalamin pathway, whereas *P. tricornutum*, *F. solaris*, and *S. acus* possess none (fig. 3*a*). Interestingly, although *P. multiseries* is auxotrophic for cobalamin, *F. cylindrus* is not. Thus, despite the presence of *MetE*, *F. cylindrus* expanded its repertoire of cobalamin synthesis genes and prefers to maximally optimize its uptake to perform methionine synthesis by the more efficient *MetH*.

By querying the metatranscriptomic TARA Oceans data, it was clear that *CobU*, *CbiZ*+*CbiB*, and *CobQ* are significantly correlated with nitrate concentration and day length (fig. 4) (supplementary figs. S8 and S9, Supplementary Material online), whereas *CobU* and *CbiZ*+*CbiB* are anticorrelated with temperature (supplementary fig. S10, Supplementary Material online) and *CobU* and *CobQ* are anticorrelated with iron (supplementary figs. S11, Supplementary Material online). HGT genes in the cobalamin pathway are particularly abundant in the Southern and Pacific Ocean (fig. 4). The lower production rate of bacteria in low temperature and the photodegradation of cobalamins, which could be of particular importance during polar summers, might explain the cobalamin limitation and the specific expression of vitamin B12–related genes in these regions of the ocean. Thus, the acquisition of the full suite of detected HGT genes in *F. cylindrus* might confer a competitive advantage over other primary producers in its polar habitat, which is often depleted in cobalamin.

Environmental Adaptation to Light Sensing and Cold Protection

Diatoms employ photosensory proteins to gain information about their environment and respond to changing light conditions. Proteorhodopsins (PR) perceive light to drive ATP generation and are especially important when photosynthesis is comprised during iron-limiting conditions. This study confirms the bacterial origin of the PR genes in *F. cylindrus* and *Pseudo-nitzschia granii* (Marchetti et al. 2015), next to brown algae, dinophytes and haptophytes. Furthermore, the red/far-red light–sensing phytochrome *DPH1* (Fortunato et al. 2016) was detected as HGT in *P. tricornutum* (1 copy), *S. robusta* (4), *S. acus* (7), *C. cryptica* (1), and *T. pseudonana* (1) and formed together with brown algae, an independent branch from green algal and fungal DPHs, similar as in previous reports (Montsant et al. 2007; Fortunato et al. 2016) and was predicted to have originated from HGT.

Polar diatoms such as *F. cylindrus* undergo periods of prolonged darkness, low temperature, and high salinity. Their ability to thrive in these conditions could be partially attributed to cryoprotectants that interfere with the growth of ice (Bayer-Giraldi et al. 2018). Ice-binding proteins were found to be laterally transferred from a basidiomycete lineage to *Fragilariopsis curta* and *F. cylindrus* (Sorhannus 2011). Also the phylogenetic tree inferred in this study detected relatedness between fungal and diatom antifreeze proteins but was not classified as HGT as this pipeline only detects prokaryotic-to-eukaryotic HGT. However, a second gene family of *F. cylindrus* proteins containing the ice-binding protein domain (IPR021884) was found to be transferred from *Cryobacterium*.

Carbon and Nitrogen Metabolism

Diatoms can rapidly recover from prolonged nitrogen limitation due to presence of the urea cycle that allows for carbon fixation into nitrogenous compounds (Allen et al. 2011). Two genes in the metabolic branches derived from this pathway, carbamate kinase and ornithine cyclodeaminase, were found to be laterally transferred, both here as in previous studies (Bowler et al. 2008; Allen et al. 2011). The latter enzyme is responsible for the conversion of ornithine to proline, which is the main osmolyte during salt stress in diatoms. Another way of nitrogen storage and translocation is the catabolism of purines to urate that can be further degraded to allantoin. It was found that plants and diatoms independently evolved a fusion protein (Urah-Urad domain; allantoin synthase) to perform the second and third steps in this urate degradation pathway (Oh et al. 2018). Exactly as in the study by Oh et al. (2018), this gene was detected to be laterally transferred from alphaproteobacteria, where this fusion event occurred, to the ancestor of haptophytes and stramenopiles.

Moreover, several genes in carbohydrate metabolism were found to be laterally transferred. The acetyl-CoA conversion to acetate occurs in a two-step process where phosphate acetyltransferase (PTA) adds a phosphate group to form acetylphosphate, that is in turn is catalyzed to acetate by acetate kinase (ACK) (Fabris et al. 2012). The PTA gene family was found to have bacterial origins and emerged in the ancestor of haptophytes and stramenopiles. In all diatoms, except for *F. cylindrus* and *P. multistriata*, multiple copies were found of this gene. Also acetate kinase was detected as an HGT gene in the pennate diatoms: *P. tricornutum*, *S. robusta*, and *S. acus*. Furthermore, this enzyme was predicted to be involved in the bifid shunt (Fabris et al. 2012). Here, the key enzyme XPK cleaves xylulose-5-phosphate to acetyl-phosphate and glyceraldehyde-3-phosphate, followed by conversion of acetyl-phosphate to acetate by ACK. Also *XPK* was laterally transferred in the pennate diatoms, single-copy in *P. tricornutum* and significantly expanded to five copies in *S. acus*. *XPK* and *ACK* are syntenic in *P. tricornutum*, what was already suggested to point to a bacterial origin as this spatial organization is also detected in Proteobacteria and Cyanobacteria (Fabris et al. 2012). Interestingly, *S. acus* has also conserved the physical association of *XPK* and *ACK* and maintained a bidirectional promoter, although an inversion of the gene order occurred (supplementary fig. S12, Supplementary Material online). Furthermore, phosphofructokinase and the cytosolic fructose-bisphosphate aldolase *Fba4* in the glycolysis (Allen et al. 2012), phosphopentose epimerase (Whitaker et al. 2009) in the pentose phosphate pathway, and a putative D-lactate dehydrogenase are enzymes that were predicted to be transferred from bacteria present in diatoms. Finally, also bacterial xylanases, glucanases, and glucosidases expanded the carbohydrate metabolic repertoire in diatoms.

The biosynthetic aspartate-derived pathway to synthesize the four amino acids, lysine, threonine, methionine, and isoleucine, was completed due to HGT (Sun and Huang 2011). Aspartate semialdehyde dehydrogenase (asd) performs the second step in this pathway and is derived from Proteobacteria. The end product L-aspartate 4-semialdehyde can be used by either dihydrodipicolinate synthase (dapA) toward lysine biosynthesis or homoserine dehydrogenase (thrA) toward threonine and methionine. Both genes were laterally transferred from bacteria. The metabolic pathways of other amino acids were also affected; the last step in tryptophan synthesis is achieved by tryptophan synthase. Although in diatoms the alpha and beta subunit of this enzyme are merged, in *P. tricornutum*, an extra copy of the beta subunit is present (Jiroutová et al. 2007) (Phatr3_J52286) that was deemed bacterial. Also alanine racemase, arginine biosynthesis *ArgJ*, leucyl-tRNA synthetase *leuRS2*, glycyl-tRNA synthetase *glyRS2*, and tyrosine-tRNA ligase *tyrRS2* were laterally transferred.

Selection Pattern for HGT Genes in *P. tricornutum*

Genomic sequence information from 10 *Phaeodactylum* accessions, belonging to four clades sampled across the world (Rastogi et al. 2020), was used to do determine the maintenance and selection patterns across the detected HGT genes. The retention of species-specific HGT genes across different strains confirmed their horizontally derived origin and did not point to contamination (for more details, see supplementary Note 2, Supplementary Material online). Moreover, analyzing gene selection patterns gives an indication on the strength of functional conservation. Variant calling resulted in a data set of 585,715 high-confidence bi-allelic SNPs. The total number of SNPs per strain across the genome was low and ranged from 0.96% to 1.37% (supplementary table S4, Supplementary Material online). To detect selective pressure, $\pi N/\pi S$ was calculated. This metric compares the fraction of synonymous and nonsynonymous mutations within a coding open-reading frame across strains. A gene experiencing neutral evolution has a $\pi N/\pi S$ value of 1, whereas a value smaller than 1 signifies negative purifying selection. The smaller the ratio of nonsynonymous and synonymous nucleotide diversity, the stronger is the level of purifying selection acting on the gene. The average synonymous nucleotide diversity (πS) across all accessions is 0.009, whereas the nonsynonymous nucleotide diversity (πN) is 0.003, thus the genome-wide average $\pi N/\pi S$ ratio is 0.3. This value is similar to what was described by Rastogi et al. (2020) and means most genes undergo strong purifying selection. The average $\pi N/\pi S$ for genes of vertical descent is 0.302, whereas for HGT genes, it is significantly lower at 0.268 ($P = 5.9 \times 10^{-4}$). Whereas repeat-associated HGT genes have higher $\pi N/\pi S$ values, indicating less purifying selection, the mean number and length of introns is higher than in HGT genes not associated with repeats (supplementary table S5, Supplementary Material online). This pattern suggests that the presence of repeats has a positive effect on the sequence evolution of HGT genes. When comparing the

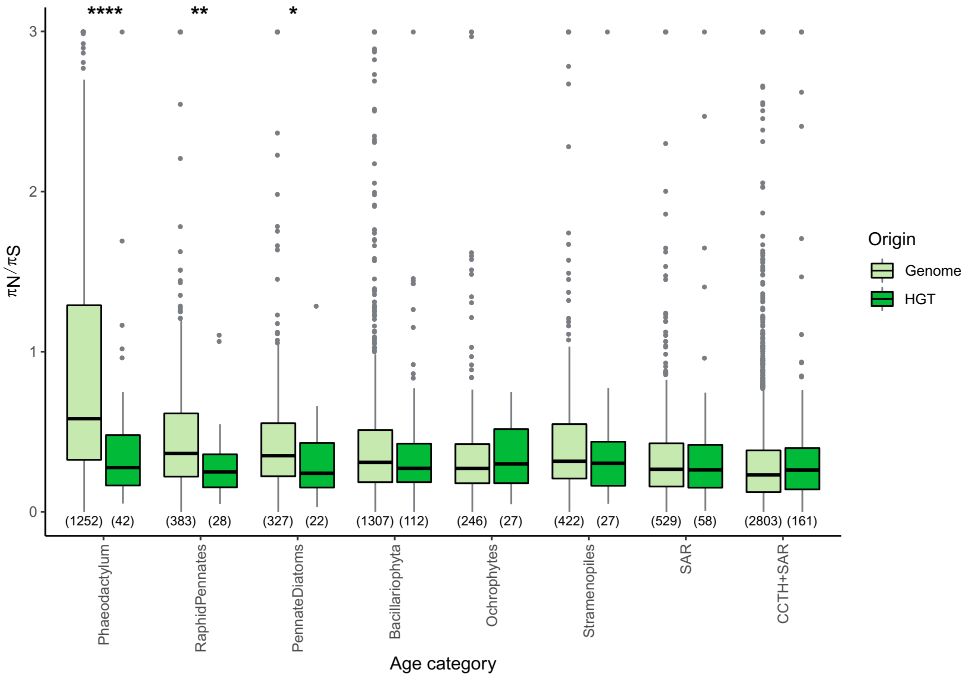


Fig. 5. Selective pressure over time in *P. tricornutum*. Distribution of selective pressure, measured by $\pi N/\pi S$, across age classes sorted from young to old and per origin in *P. tricornutum*. SAR is the clade comprising Stramenopiles, Alveolates and Rhizaria, whereas CCTH consists of Cryptophyta, Centrohelida, Telonemia, and Haptophyta and CCTH + SAR denotes the ancestor of both groups. Number of genes is indicated in between parenthesis. The asterisks denote a statistical difference per type within the same age category and have the following confidence range for *P* values; $*\leq 0.05$, $**\leq 0.01$, $***\leq 0.001$, $****\leq 0.0001$.

$\pi N/\pi S$ ratios for HGT and vertical genes across age classes (fig. 5), it is apparent that this difference is due to the youngest gene categories, being the *Phaeodactylum*, raphid pennate diatoms and pennate diatoms-specific genes, where vertical genes are less constrained than HGT genes in those age categories. To the best of our knowledge, this is the first time the selection pressure of prokaryotic HGT genes is assessed in unicellular eukaryotes and compared with vertically descended genes while taking age into account. Although it has already been observed that *de novo* genes display patterns of rapid evolution and the strength of purifying selection increases with age (Vakirlis et al. 2018), it is remarkable to observe that HGT genes deviate from this pattern. Unlike recent innovations from vertical descent, young HGT genes are quickly integrated in the biological network exemplified by their high levels of purifying selection.

Expression and Coexpression Network Analysis of HGT Genes

The availability of RNA sequencing experiments in several diatoms allows for the construction of genome-wide gene expression atlases quantifying gene expression levels across a wide range of conditions. These compendia consisted of 13–76 conditions per species, all having biological replicates per condition (supplementary table S6, Supplementary Material online). The vast majority of HGT genes are expressed: in *P. tricornutum*, all HGT genes are expressed, in *T. pseudonana* 558 out of 580 (96%), in *S. robusta* 1,597 out of 1,741 (92%), and in *F. cylindrus* 741 out of 762 (97%). Given that most HGT genes are kept under purifying selection in *P. tricornutum* and are transcribed in diatoms, this is indicative that they are functional and can play a vital role in expanding the functional repertoire. Indeed, 64% of the predicted HGT genes in *P. tricornutum* were translated into proteins in a proteogenomic analysis (Yang et al. 2018). This is similar to 63% of all proteins in the genome that were detected to be translated.

Next, the expression specificity was calculated per gene, where a low value signifies broad expression in many (or even all) conditions and a value close to one indicates expression in one or a few. Species-specific genes have a higher mean condition-specific expression, both for vertical- and horizontal-derived genes in *P. tricornutum* ($P < 2 \times 10^{-16}$, 1.1×10^{-06}), *S. robusta* ($<2 \times 10^{-16}$, 1.6×10^{-13}), *F. cylindrus* ($<2 \times 10^{-16}$, 2.2×10^{-03}), and *T. pseudonana* ($<2 \times 10^{-16}$, 3.7×10^{-02}). A declining trend of condition specificity was observed over time. Whenever there was a significant difference in condition specificity between HGT and vertical genes within the same age category, HGT genes consistently displayed on average a more specific expression pattern (fig. 6). A high tissue specificity for species-specific genes which decreases over time has also been observed in mouse (Lehner and Fraser 2004; Freilich et al. 2005). Interestingly, the selection pressure in *P. tricornutum* across all age classes and per mode of inheritance is not strongly correlated with expression specificity ($R^2 = 0.00121$ for vertical genes, $R^2 = 0.000729$ for horizontal genes) (supplementary fig. S13, Supplementary Material online), showing that genes having a highly specific expression pattern are not necessarily under less purifying selection, defying the trend that was previously observed in mammals (Zhang and Li 2004).

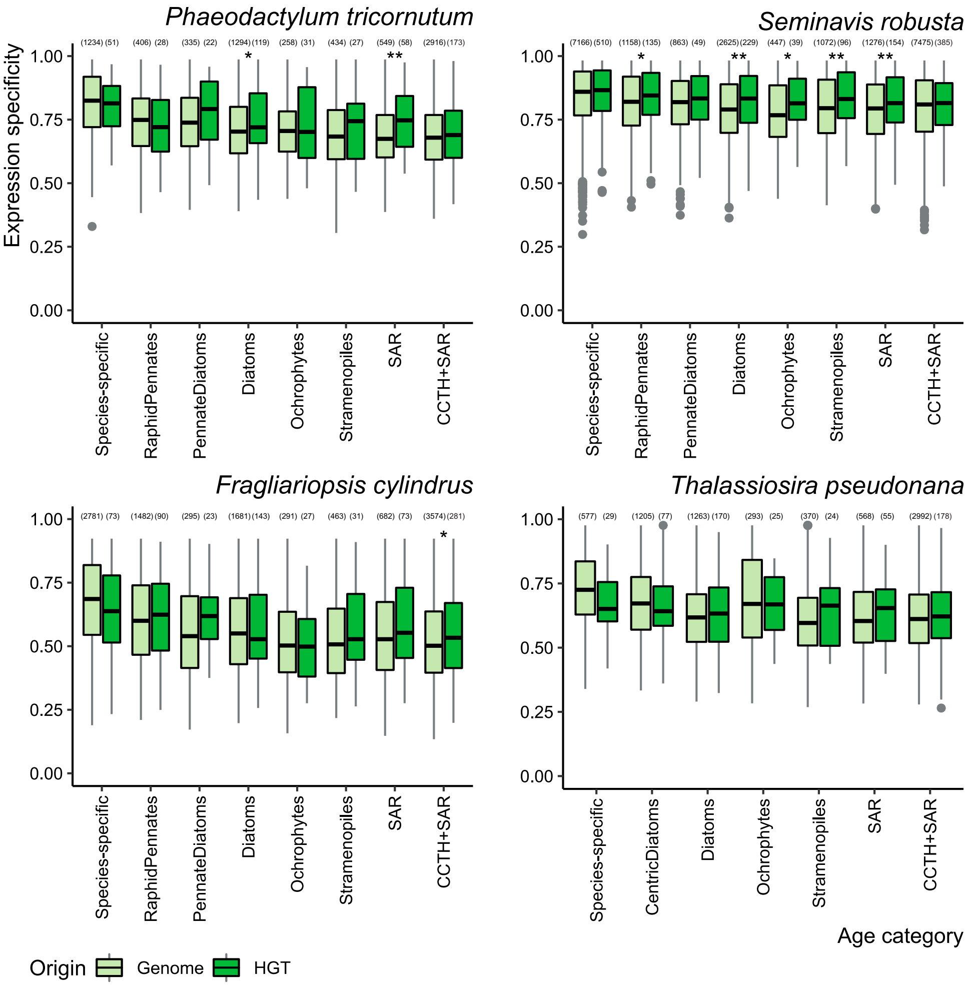


Fig. 6. Expression specificity over time in diatoms. Distribution of expression specificity across age classes sorted from young to old and per origin in four diatoms. SAR is the clade comprising Stramenopiles, Alveolates, and Rhizaria, whereas CCTHs consist of Cryptophyta, Centrohelida, Telonemia, and Haptophyta and CCTH + SAR denotes the ancestor of both groups. The number of genes is indicated in between parenthesis. The asterisks denote a statistical difference per type within the same age category and have the following confidence range for *P* values; *≤0.05, **≤0.01, ***≤0.001, ****≤0.0001.

Based on a global coexpression *P. tricornutum* network constructed using an expression atlas comprising 211 samples, for every gene in *P. tricornutum* the coexpression neighborhood, based on its highest reciprocal ranks, was defined as a module. Subsequently, these modules and the known gene functions for genes part of this module were used, through guilt-by-association analysis, to gain functional insights in the detected HGT genes. For 19 HGT genes, the coexpression modules confirmed enrichment for at least one known function. Fructose-bisphosphate aldolase *Fba4* is enriched in its coexpression module for genes involved in a carbohydrate metabolic process and aspartate semialdehyde dehydrogenase (*asd*) has enrichment for amino acid biosynthesis. New functions were attributed to 320 out of 509 HGT genes based on significant GO enrichment of the coexpression modules. For example, two HGT proteins involved in amino acid synthesis—tryptophan synthase β chain ($P = 2.2 \times 10^{-16}$) and *ArgJ* ($P = 1 \times 10^{-15}$)—are predicted to be coregulated with photosynthetic genes, although both proteins do not contain the chloroplast-targeting peptide. The coexpression neighborhood of phosphofructokinase is significantly enriched ($P = 3.7 \times 10^{-06}$) to be involved in the Krebs cycle, whereas this protein is part of the glycolysis and directly upstream of the citric acid cycle. Also Phatr3_J40382, which contains a pyruvate kinase-like domain, is enriched for this GO term, and this corroborates its metabolic function. Methionine sulfoxide reductase *MsrB* (Phatr3_J13757) was predicted to have a similar expression pattern to genes partaking in iron–sulfur cluster assembly ($P = 1.1 \times 10^{-03}$) and metal ion transport ($P = 1.4 \times 10^{-02}$). In yeast, these proteins were already shown to have a protective role for FeS clusters during oxidative stress (Sideri et al. 2009). The far-red light phytochrome *DPH1* is enriched for genes involved in transcriptional regulation ($P = 5.5 \times 10^{-03}$), which could point to its primary role in the light-sensing cascade. These results demonstrate that coexpression network analysis offers a pragmatic means to predict the biological processes HGT genes are involved in.

Discussion

Through the application of phylogeny-based HGT detection, we identified 1,963 gene families with a prokaryotic origin in diatoms. Although HGT detection has been previously performed in diatoms, this is the first large-scale phylogenetic approach of HGT detection across all available sequenced diatoms. Although some previous studies were based on phylogenetics (Bowler et al. 2008; Basu et al. 2017), most relied purely on sequence homology searches (Lommer et al. 2012; Traller et al. 2016; Rastogi et al. 2018), while it has been shown that the degree of gene similarity does not necessary necessarily reflect phylogenetic relationships (Koski and Golding 2001; Philippe et al. 2011). Although HGT had been previously predicted in *P. multistriata*, *C. cryptica*, *T. oceanica*, and twice in *P. tricornutum*, only a fraction of HGT genes were confirmed across these studies, going from 6% to 45% (supplementary fig. S14, Supplementary Material online), which is more than expected by chance. This could be due to the usage of different methods, criteria and underlying databases. For example, horizontal genes were defined in *T. oceanica* if they did not show high similarity with any other stramenopile and thus contain *T. oceanica*–specific genes from both prokaryotic- and eukaryotic-to eukaryotic origin. Whereas the different guidelines to identify HGT, proposed by Richards and Monier (2016), were followed in this study, different methodological aspects of phylogenomics make the interpretation of complex tree topologies challenging. Both incomplete lineage sorting and uncertainty in the tree topology can lead to erroneous inference (Ravenhall et al. 2015), potentially inflating the number of identified HGT events. Nevertheless, when only selecting HGT genes inferred here that were confirmed by at least one other study, the trend of strong purifying selection and declining expression specificity over time remained unchanged (supplementary figs. S15, Supplementary Material online), confirming the reliability of our findings.

Exploring genome-wide nucleotide diversity information revealed that laterally transferred genes are showing stronger levels of purifying selection compared with new genes originating from noncoding regions, so-called *de novo* genes. Moreover, most species-specific HGT genes are present across all strains in *P. tricornutum*, suggesting they were acquired prior the divergence of these strains. An alternative scenario would be that the HGT event occurred in one strain and was subsequently spread through admixture. Independent of which scenario is most likely, we do observe that a large majority of these HGT genes are maintained in all *P. tricornutum* strains. Indeed, in grasses, it was recently shown that several plant-to-plant HGT fragments were rapidly integrated and spread across the population, after which erosion occurred on neutrally selected genes within those fragments (Olofsson et al. 2019).

Complementary to studying selection patterns of HGT genes, expression analysis revealed that 92%–100% of the transferred genes are transcriptionally active. Furthermore, in *P. tricornutum*, 64% of the identified HGT showed evidence for translation, corroborating their functional importance. Given that for 19 HGT genes the functions learned from their coexpression module confirmed the biological processes they are potentially involved in, this approach can shed light on the functional relevance of other HGT genes in diatoms. Among the HGT events, we detected the transfer of five concurrent genes of the vitamin B12 biosynthetic pathway. A cobalamin addition experiment in a high-nutrient low chlorophyll (HNLC) region in the Gulf of Alaska significantly altered the species composition, going from diatom-dominated plankton to an increased fraction of ciliates and dinoflagellates (Koch et al. 2011). This could be explained by the presence of these HGT genes and the corresponding enhanced uptake mechanism of vitamin B12 and its analogues, which give diatoms a competitive advantage during limiting conditions. In conclusion, our results support a high genetic plasticity and ability for local adaptation in diatoms due to HGT.

Materials and Methods

Gene Family Construction

The publicly available genomes of 17 stramenopiles, 1 alveolate, 1 rhizarian, and 1 haptophyte (listed in supplementary table S1, Supplementary Material online) were downloaded, and their nuclear proteomes, totaling to 398,001 protein-coding genes, were searched for similarity in an all-against-all fashion with BLASTp (version 2.6+) using an e-value cut-off of 10^{-5} and retaining maximum 4,000 hits. Next, clustering of these protein-coding genes was performed using OrthoFinder (version 2.1.2) (Emms and Kelly 2015).

Species-Level Phylogeny

To delineate the species phylogeny for all SAR members, which is the clade comprising Stramenopiles, Alveolates and Rhizaria, using *Emiliania huxleyi* as an outgroup, OrthoFinder gene families where all species have a copy number of either 1 or 2 genes were selected, and one gene sequence was randomly picked in case of duplication. MUSCLE (Edgar 2004) was used to build a sequence alignment per locus, followed by concatenation. Afterward IQ-Tree (version 1.7.0b7) (Nguyen et al. 2015) was used to build a concatenated tree using an edge-linked-proportional partition model and 1,000 bootstraps. Additionally, IQ-Tree was used to estimate every single-locus tree. In both analyses, the best protein of model of evolution per gene family was selected using ModelFinder. Finally, the gene- and site-concordance factors of the inferred species-level phylogeny were calculated (Minh et al. 2020).

HGT Detection

The NCBI nonredundant protein database (download date June 8, 2018) was complemented with the proteomes of 20 species (supplementary table S1, Supplementary Material online). Diamond (version 0.9.18.119) (Buchfink et al. 2015) searches were performed in sensitive mode against this database for all proteins of these 20 species, retaining maximum 1,000 hits per query. Hits were reduced to maximum five sequences for each order, and 15 sequences per phylum. Genes families with at least one copy in a diatom and at least one-third of the diatom members having a prokaryotic hit were analyzed. The hits of all diatom members were combined and clustered using CD-HIT (version 4.6.1) (Fu et al. 2012) based on a 95% identity cut-off. Next, the sequences were aligned with MAFFT (version 7.187) (Katoh et al. 2002) in automatic mode. Maximum likelihood trees were produced using IQTree (version 1.6.5) (Nguyen et al. 2015) including a test for the best fitting protein model (-mset JTT, LG, WAG, Blosum62, VT, Dayhoff) (Kalyaanamoorthy et al. 2017). The FreeRate model was used to account for rate heterogeneity across sites (-mrate R), empirical base frequencies were calculated (-mfreq F), and 1,000 rounds of ultraparametric bootstrapping (-bb 1000) (UFBoot2) (Hoang et al. 2018) were run.

Phylogenetic trees were reordered based on midpoint rooting, unless the whole eukaryotic fraction formed a cluster and then this cluster was used as a subtree for rooting. Every node having a bootstrap support $\geq$ 90, and consisting of a prokaryotic and eukaryotic subtree, was considered to have originated from HGT and the LCA of both subtrees was inferred. However, to avoid HGT inference due to spurious database hits, HGT calls were only kept if prokaryotic sequences made up more than 15% of the sequences belonging to that node when its eukaryotic subtree was larger than 20 sequences. Moreover, to avoid classifying EGT incorrectly as HGT, only prokaryotic-to-eukaryotic events were analyzed and events older than SAR + Haptophytes, also dubbed SAR + CCTH, were discarded. When several subsequent nodes in a gene tree complied to be HGT, the overall LCA of the prokaryotic subset was considered the donor of this event.

Contamination Detection

To assess the degree of contamination from *Sphingomonas sp.* in *S. acus*, 914 genomes of the order Sphingomonadales were retrieved from NCBI and a nucleotide blast against the *S. acus* genome was performed. Contigs having at least 70% identity and 25% alignment coverage were deemed to have Sphingomonadales origin.

Next, the remaining *S. acus* contigs as well as all scaffolds from the other diatoms that contained predicted genes were subjected to a BLAST against the nucleotide nonredundant database (download date October 1, 2019). Again, contigs having at least 70% identity and 25% alignment coverage having similarity to nonstramenopiles as best hits were classified as contamination. For *F. cylindrus*, *F. solaris*, *P. multiseries*, *P. tricornutum*, *S. robusta*, and *T. pseudonana* no contaminating contigs were detected. For *S. acus* and *T. oceanica*, 1 and 7 contigs, respectively, were detected. However, none contained predicted HGT genes. In the *C. cryptica* genome, 17 contigs were found, mostly from *Escherichia coli*, which contained in total 8 misclassified HGT genes that were removed from the analyses.

Additionally, detection of contaminating sequences was analyzed in the context of long reads. First, long-read data were mapped to the genome. Nanopore MinION reads for *P. tricornutum* (SRX4617960) and *T. pseudonana* (SRX4617979) were mapped using GraphMap (Sović et al. 2016), whereas PacBio libraries for *S. robusta* (PRJEB36614) and *F. cylindrus* (PRJEB15040) were mapped using BLASR (Chaisson and Tesler 2012) (-m 4 -bestn 2 -maxAnchorsPerPosition 100 -advanceExactMatches 10 -affineAlign -affineOpen 100 -affineExtend 0 -insertion 5 -deletion 5 -extend -maxExtendDropoff). Next, the overlapping genes on the mapped long reads were delineated using bedtools (Quinlan and Hall 2010) and the fraction of predicted HGT genes on the total genes spanned by a read was calculated.

Gene Family Expansion

Expanded families were delineated by calculating the Z-score profile of the gene copy number per HGT family across all diatoms excluding the allodiploid *F. solaris*. Families where the variance is larger than two and the Z-score for a particular species is larger than three were deemed expanded in that species.

Structural Properties of HGT Genes

Structural genomic annotation features for sequenced diatoms were retrieved and GC content, coding sequence length, number of introns per gene, and intron length were compared between horizontally and vertically transferred genes among several age categories. Also the distance for every gene in *P. tricornutum* to the closest TE as defined by (Rastogi et al. 2018) and centromeric and telomeric region elucidated by (Diner et al. 2017) was calculated and compared among the different origins. Statistical significance was calculated by the Wilcoxon rank sum test. For the diatoms *T. pseudonana* and *P. tricornutum*, whose genomes are resolved on chromosome-scale level, the distribution of HGT genes was plotted using R.

Functional Interpretation of HGT Genes

The proteomes of all species were functionally annotated using Interproscan (version 60) (Jones et al. 2014) in order to obtain functional domain annotations and GO terms. KEGG orthology identifiers (Ogata et al. 1999) were attained using EggNOG-mapper(Huerta-Cepas et al. 2017). For all diatoms, the chloroplast targeting signal was predicted using ASAFind (version 1.1.7) (Gruber et al. 2015). Only GO terms within the subtree "biological process" were taken into account. These terms were expanded to also contain all ancestral functional information. GO and Interpro domain enrichment was performed on the HGT genes per species using hypergeometric testing, and multiple hypothesis testing was constrained using Benjamini–Hochberg correction ($q < 0.05$). Functional enrichments found in at least two species were visualized using the ComplexHeatmap package (Gu et al. 2016) (R version 3.4) and clustered using the complete linkage method.

Tandem duplicates were defined as genes belonging to the same gene family and located within 15 genes of each other and identified using i-ADHoRe v3.0 (Proost et al. 2012) (alignment method: gg2, gap size 15, tandem gap 15, cluster gap 15, $q = 0.85$, probability cut-off 0.01, anchor_points 3, level_2_only FALSE, FDR as method for multiple hypothesis correction).

Metatranscriptome Analysis

For several selected HGT gene families involved in cobalamin synthesis, a HMM profile was created using hmmer3 v3.1b2

(Eddy 2011) and these were uploaded to the Ocean Gene Atlas webserver (Villar et al. 2018) to query the eukaryotic MATOU gene data set (Carradec et al. 2018) (hmmer, evalue cut-off 10^{-10}) linked to the metatranscriptomic TARA Oceans data. Only sequences taxonomically assigned as diatoms were further analyzed. The abundance was estimated as the number of sequence per station and depth.

Population Genetics

Data from ten resequencing strains were downloaded from the public repository SRA (https://www.ncbi.nlm.nih.gov/sra) (SRR6476693–SRR6476702), and these reads were mapped to the *P. tricornutum* genome using BWA-mem (version 0.7.17) (Li and Durbin 2009). The read alignments per strain were filtered to only include unique mappings without chimeric alignments using samtools (version 1.6) (Li et al. 2009). SGSGeneLoss (version 0.1) (Golicz et al. 2015) was run in a relaxed mode to determine to presence/absence pattern of all genes across the ten strains (minCov = 1, lostCutoff = 0.05, thus requiring only one read and 5% gene coverage to be perceived as present). The resulting phylogenetic pattern of HGT genes was visualized using the ComplexHeatmap package (Gu et al. 2016) (R version 3.4) and clustered using the complete linkage method.

SNP calling was performed per strain using HaplotypeCaller, after which SNPs were integrated using GenotypeGVCFs. Both methods are available within the GATK framework (version 3.7) (McKenna et al. 2010). SNPs were filtered using the GATK recommended hard filters ($QD < 2.0$; $FS > 60.0$; $MQ < 40.0$; $MQRankSum \leq 12.5$; $ReadPosRankSum \leq 8.0$) (Van der Auwera et al. 2013) and only bi-allelic SNPs were retained.

To estimate the degree of negative-purifying pressure across the proteome, only coding positions having a read depth ≥ 10 across all strains were considered, calculated using SAMtools mpileup (Li 2011). In total, 89% of all genic positions could be analyzed and 272,235 SNPs were observed in these regions. We used SnpEff (version 4.3t) (Cingolani et al. 2012) to predict the individual effect per SNP and $\pi N/\pi S$ was calculated taking only the callable positions for complete codons into account and correcting for the allele frequency of the mutation in the population. Statistical significance of difference in selective pressure across mode of inheritance and age classes was calculated by the Wilcoxon rank sum test.

Expression and Coexpression Analysis

An expression atlas for every diatom species which has RNA-Seq expression data available was generated. First, relevant experiments were searched using Curse (Vaneechoutte and Vandepoele 2019), which also allows the user to identify and curate replicates. The experiments listed in (supplementary table S6, Supplementary Material online) were used to generate the expression compendia. Next, the atlas was generated using Prose (Vaneechoutte and Vandepoele 2019), which uses the SRA toolkit to download the raw data locally, FastQC (Andrews 2010) to perform quality control and adapter detection, Trimmomatic (Bolger et al. 2014) for automatic read trimming, and finally kallisto (Bray et al. 2016) for expression quantification in transcripts per million (TPM). Genes were deemed expressed when having a TPM value of at least 3. The condition specificity, also known as tau (Kryuchkova-Mostacci and Robinson-Rechavi 2016), of every gene was calculated as follows, where *x* is the TPM value per condition, max is the maximal expression of a gene and n is the number of conditions in the expression compendium:

$$\mathrm{tau} = \frac{\sum_{1}^{n}(1 - (x/\mathrm{max}))}{n}.$$

Condition-specific genes were defined as having a tau value of ≥ 0.9.

The generated expression atlas for *P. tricornutum* was also used to define coexpression clusters. The Pearson correlation was calculated in a pairwise manner between all genes and the highest reciprocal rank (HRR) (Mutwil et al. 2010; Liesecke et al. 2018) was determined at 23, by maximizing the recovery of known GO annotations, while restraining the number of novel predictions. A cluster was defined for every gene based on this cut-off and GO enrichment using hypergeometric testing was run per cluster. Multiple hypothesis testing was constrained using Benjamini–Hochberg correction ($q < 0.05$).

Data Availability

All gene families, phylogenetic trees of horizontal descent, and the dating of the HGT events within these gene families are available on Zenodo (https://zenodo.org/record/3889669).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful to Dr Laura Eme for her careful reading of the manuscript and her constructive comments. E.V. wants to acknowledge the funding obtained by the BOF project GOA01G01715.

Author Contributions

E.V. wrote the article and performed species topology delineation, horizontal gene transfer (HGT) detection analysis, functional interpretation, metagenomic, and population genomic analysis. T.D. aided in HGT delineation and performed coexpression analysis. C.M.O-C. aided in population genomic analysis and performed expression analysis generation of *S. robusta*. K.V. supervised the project. All authors read, edited, and approved the article.

References

Alexander WG, Wisecaver JH, Rokas A, Hittinger CT. 2016. Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. *Proc Natl Acad Sci U S A.* 113(15):4116–4121.

Allen A, Dupont C, Oborník M, Horák A, Adriano N-N, John M, Zheng H, Johnson D, Hu H, Fernie A, et al. 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473(7346):203–207.

Allen AE, Moustafa A, Montsant A, Eckert A, Kroth PG, Bowler C. 2012. Evolution and functional diversification of fructose bisphosphate aldolase genes in photosynthetic marine diatoms. *Mol Biol Evol*. 29(1):367–379.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Armbrust E, Berges J, Bowler C, Green B, Martinez D, Putnam N, Zhou S, Allen A, Apt K, Bechner M, et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306(5693):79–86.

Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best Practices Pipeline. *Current Protocols in Bioinformatics*. 43(1):11.10.1–11.10.33.

Basu S, Patil S, Mapleson D, Russo M, Vitale L, Fevola C, Maumus F, Casotti R, Mock T, Caccamo M, et al. 2017. Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytol*. 215(1):140–156.

Bayer-Giraldi M, Sazaki G, Nagashima K, Kipfstuhl S, Vorontsov DA, Furukawa Y. 2018. Growth suppression of ice crystal basal face in the presence of a moderate ice-binding protein does not confer hyperactivity. *Proc Natl Acad Sci U S A*. 115(29):7479–7484.

Bertrand EM, McCrow JP, Moustafa A, Zheng H, McQuaid JB, Delmont TO, Post AF, Sipler RE, Spackeen JL, Xu K, et al. 2015. Phytoplankton–bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. *Proc Natl Acad Sci U S A*. 112(32):9938–9943.

Bertrand EM, Moran DM, McIlvin MR, Hoffman JM, Allen AE, Saito MA. 2013. Methionine synthase interreplacement in diatom cultures and communities: implications for the persistence of B12 use by eukaryotic phytoplankton. *Limnol Oceanogr*. 58(4):1431–1450.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

Bowler C, Allen A, Badger J, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar R, et al. 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 456(7219):239–244.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 34(5):525–527.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 12(1):59–60.

Campbell GRO, Taga ME, Mistry K, Lloret J, Anderson PJ, Roth JR, Walker GC. 2006. *Sinorhizobium meliloti* bluB is necessary for production of 5,6-dimethylbenzimidazole, the lower ligand of B12. *Proc Natl Acad Sci U S A*. 103(12):4634–4639.

Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-Mendez G, Rocha F, Tirichine L, Labadie K, Tara Oceans Coordinators, et al. 2018. A global ocean atlas of eukaryotic genes. *Nat Commun*. 9(1):373.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*. 13(1):238.

Chou S, Daugherty M, Peterson S, Biboy J, Yang Y, Jutras B, Lillian F-L, Ferrin M, Harding B, Christine J-W, et al. 2015. Transferred interbacterial antagonism genes augment eukaryotic innate immune function. *Nature* 518(7537):98–101.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6(2):80–92.

Cohen NR, A. Ellis K, Burns WG, Lampe RH, Schuback N, Johnson Z, Sañudo-Wilhelmy S, Marchetti A. 2017. Iron and vitamin interactions in marine diatom isolates and natural assemblages of the Northeast Pacific Ocean: iron and vitamin interactions in diatoms. *Limnol Oceanogr*. 62(5):2076–2096.

Cohen NR, Mann E, Stemple B, Moreno CM, Rauschenberg S, Jacquot JE, Sunda WG, Twining BS, Marchetti A. 2018. Iron storage capacities and associated ferritin gene expression among marine diatoms: iron storage and ferritin expression in diatoms. *Limnol Oceanogr*. 63(4):1677–1691.

Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. 2005. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* 438(7064):90–93.

Dean P, Sendra K, Williams T, Watson A, Major P, Nakjang S, Kozhevnikova E, Goldberg A, Kunji E, Hirt R, et al. 2018. Transporter gene acquisition and innovation in the evolution of Microsporidia intracellular parasites. *Nat Commun*. 9(1):1709.

Diner RE, Noddings CM, Lian NC, Kang AK, McQuaid JB, Jablanovic J, Espinoza JL, Nguyen NA, Anzelmatti MA, Jansson J, et al. 2017. Diatom centromeres suggest a mechanism for nuclear DNA acquisition. *Proc Natl Acad Sci U S A*. 114(29):E6015–E6024.

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol*. 7(10):e1002195.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32(5):1792–1797.

Ellis KA, Cohen NR, Moreno C, Marchetti A. 2017. Cobalamin-independent methionine synthase distribution and influence on vitamin B12 growth requirements in marine diatoms. *Protist* 168(1):32–47.

Eme L, Gentekaki E, Curtis B, Archibald J, Roger A. 2017. Lateral gene transfer in the adaptation of the anaerobic parasite blastocystis to the gut. *Curr Biol*. 27(6):807–820.

Emms D, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 16(1):157.

Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart G. 2012. The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner–Doudoroff glycolytic pathway. *Plant J*. 70(6):1004–1014.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski PG. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281(5374):237–240.

Foflonker F, Mollegard D, Ong M, Yoon HS, Bhattacharya D. 2018. Genomic analysis of *Picochlorum* species reveals how microalgae may adapt to variable environments. *Mol Biol Evol*. 35(11):2702–2711.

Fortunato AE, Jaubert M, Enomoto G, Bouly J-P, Raniello R, Thaler M, Malviya S, Bernardes JS, Rappaport F, Gentili B, et al. 2016. Diatom phytochromes reveal the existence of far-red-light-based sensing in the ocean. *Plant Cell* 28(3):616–628.

Freilich S, Massingham T, Bhattacharyya S, Ponsting H, Lyons PA, Freeman TC, Thornton JM. 2005. Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. *Genome Biol*. 6(7):R56.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150–3152.

Galachyants Y, Zakharova Y, Petrova D, Morozov A, Sidorov I, Marchenkov A, Logacheva M, Markelov M, Khabudaev K, Likhoshway Y, et al. 2015. Sequencing of the complete genome of an araphid pennate diatom *Synedra acus* subsp. *radians* from Lake Baikal. *Dokl Biochem Biophys*. 461(1):84–88.

Golicz AA, Martinez PA, Zander M, Patel DA, Van De Wouw AP, Visendi P, Fitzgerald TL, Edwards D, Batley J. 2015. Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Funct Integr Genomics*. 15(2):189–196.

Gonçalves C, Wisecaver JH, Kominek J, Oom M, Leandro M, Shen X-X, Opulente DA, Zhou X, Peris D, Kurtzman CP, et al. 2018. Evidence for loss and reacquisition of alcoholic fermentation in a fructophilic yeast lineage. *eLife* 7(e33034):.

Gruber A, Rocap G, Kroth P, Armbrust E, Mock T. 2015. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant J*. 81(3):519–528.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32(18):2847–2849.

Harding T, Roger AJ, Simpson AG. 2017. Adaptations to high salt in a halophilic protist: differential expression and gene acquisitions through duplications and gene transfers. *Front Microbiol*. 8:944.

Heal KR, Qin W, Ribalet F, Bertagnolli AD, Coyote-Maestas W, Hmelo LR, Moffett JW, Devol AH, Armbrust EV, Stahl DA, et al. 2017. Two distinct pools of B $_{12}$ analogs reveal community interdependencies in the ocean. *Proc Natl Acad Sci U S A*. 114(2):364–369.

Helliwell KE. 2017. The roles of B vitamins in phytoplankton nutrition: new perspectives and prospects. *New Phytol*. 216(1):62–68.

Helliwell KE, Lawrence AD, Holzer A, Kudahl UJ, Sasso S, Kräutler B, Scanlan DJ, Warren MJ, Smith AG. 2016. Cyanobacteria and eukaryotic algae use different chemical variants of vitamin B12. *Curr Biol*. 26(8):999–1008.

Helliwell KE, Wheeler GL, Leptos KC, Goldstein RE, Smith AG. 2011. Insights into the evolution of vitamin B12 auxotrophy from sequenced algal genomes. *Mol Biol Evol*. 28(10):2921–2933.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 35(2):518–522.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol*. 34(8):2115–2122.

Husnik F, McCutcheon J. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol*. 16(2):67–79.

Janech MG, Krell A, Mock T, Kang J-S, Raymond JA. 2006. Ice-binding proteins from sea ice diatoms (Bacillariophyceae). *J Phycol*. 42(2):410–416.

Jiroutová K, Horák A, Bowler C, Oborník M. 2007. Tryptophan biosynthesis in stramenopiles: eukaryotic winners in the diatom complex chloroplast. *J Mol Evol*. 65(5):496–511.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14(6):587–589.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30(14):3059–3066.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*. 9(8):605–618.

Koch F, Marcoval MA, Panzeca C, Bruland KW, Sañudo-Wilhelmy SA, Gobler CJ. 2011. The effect of vitamin B $_{12}$ on phytoplankton growth and community structure in the Gulf of Alaska. *Limnol Oceanogr*. 56(3):1023–1034.

Kominek J, Doering DT, Opulente DA, Shen X-X, Zhou X, Jeremy D, Hulfachor AB, Groenewald M, Mcgee MA, Karlen SD, et al. 2019. Eukaryotic acquisition of a bacterial operon. *Cell*. 176(6):1356–1366.e10.

Koski LB, Golding GB. 2001. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*. 52(6):540–542.

Krasovec M, Vancaester E, Rombauts S, Bucchini F, Yau S, Hemon C, Lebredonchel H, Grimsley N, Moreau H, Sophie S-B, et al. 2018. Genome analyses of the microalga *Picochlorum* provide insights into the evolution of thermotolerance in the green lineage. *Genome Biol Evol*. 10(9):2347–2365.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2016. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform*. 18(2):205–214.

Lehner B, Fraser AG. 2004. Protein domains enriched in mammalian tissue-specific or widely expressed genes. *Trends Genet*. 20(10):468–472.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25(14):1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Liesecke F, Daudu D, Dugé de Bernonville R, Besseau S, Clastre M, Courdavault V, de Craene J-O, Crèche J, Giglioli-Guivarc'h N, Glévarec G, et al. 2018. Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Sci Rep*. 8(1):10885.

Lommer M, Specht M, Roy A-S, Kraemer L, Andreson R, Gutowska MA, Wolf J, Bergner SV, Schilhabel MB, Klostermeier UC, et al. 2012. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol*. 13(7):R66.

Marchetti A, Catlett D, Hopkinson B, Ellis K, Cassar N. 2015. Marine diatom proteorhodopsins and their potential role in coping with low iron availability. *ISME J*. 9(12):2745–2748.

Marchetti A, Parker M, Moccia L, Lin E, Arrieta A, Ribalet F, Murphy M, Maldonado M, Armbrust E. 2009. Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature*. 457(7228):467–470.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297–1303.

Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets. *Mol Biol Evol*. doi:10.1093/molbev/msaa106.

Mock T, Otillar R, Strauss J, Mark M, Paajanen P, Schmutz J, Salamov A, Sanges R, Toseland A, Ward B, et al. 2017. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* 541(7638):536–540.

Montsant A, Allen AE, Coesel S, Martino AD, Falciatore A, Mangogna M, Siaut M, Heijde M, Jabbari K, Maheswari U, et al. 2007. Identification and comparative genomic analysis of signaling and regulatory components in the diatom *Thalassiosira pseudonana*. *J Phycol*. 43(3):585–604.

Murphy CL, Youssef NH, Hanafy RA, Couger MB, Stajich JE, Wang Y, Baker K, Dagar SS, Griffith GW, Farag IF, et al. 2019. Horizontal gene transfer as an indispensable driver for evolution of *Neocallimastigomycota* into a distinct gut-dwelling fungal lineage. *Appl Environ Microbiol*. 85(15):e00988–19.

Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S. 2010. Assembly of an Interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. *Plant Physiol*. 152(1):29–43.

Nakov T, Beaulieu J, Alverson A. 2018. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (diatoms, Bacillariophyta). *New Phytol*. 219(1):462–473.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 32(1):268–274.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 27(1):29–34.

Oh J, Liuzzi A, Ronda L, Marchetti M, Corsini R, Folli C, Bettati S, Rhee S, Percudani R. 2018. Diatom allantoin synthase provides structural insights into natural fusion protein therapeutics. *ACS Chem Biol*. 13(8):2237–2246.

Olofsson JK, Dunning LT, Lundgren MR, Barton HJ, Thompson J, Cuff N, Ariyarathne M, Yakandawala D, Sotelo G, Zeng K, et al. 2019. Population-specific selection on standing variation generated by lateral gene transfers in a grass. *Curr Biol*. 29(22):3921–3927.e5.

Osuna-Cruz CM, Bilcke G, Vancaester E, De Decker S, Bones AM, Winge P, Poulsen N, Bulankova P, Verhelst B, Audoor S, et al. 2020. The *Seminavis robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. *Nat Commun*. 11(1):3320

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 9(3):e1000602.

Proost S, Fostier J, Witte D, Dhoedt B, Demeester P, de Peer Y, Vandepoele K. 2012. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res*. 40(2):e11–e11.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.

Rastogi A, Maheswari U, Dorrell R, Vieira F, Maumus F, Kustka A, James M, Allen A, Kersey P, Bowler C, et al. 2018. Integrative analysis of large scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome and evolutionary origin of diatoms. *Sci Rep*. 8(1):4834.

Rastogi A, Vieira FRJ, Deton-Cabanillas A-F, Veluchamy A, Cantrel C, Wang G, Vanormelingen P, Bowler C, Piganeau G, Hu H, et al. 2020. A genomics approach reveals the global genetic polymorphism, structure, and functional diversity of ten accessions of the marine model diatom *Phaeodactylum tricornutum*. *Isme J*. 14(2):347–363.

Ravenhall M, Škunca N, Lassalle F, Dessimoz C. 2015. Inferring horizontal gene transfer. *Plos Comput Biol*. 11(5):e1004095.

Ricard G, McEwan N, Dutilh B, Jouany J-P, Macheboeuf D, Mitsumori M, McIntosh F, Michalowski T, Nagamine T, Nelson N, et al. 2006. Horizontal gene transfer from Bacteria to rumen Ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics*. 7(1):22.

Richards T, Monier A. 2016. A tale of two tardigrades. *Proc Natl Acad Sci U S A*. 113(18):4892–4894.

Savory F, Leonard G, Richards T. 2015. The role of horizontal gene transfer in the evolution of the oomycetes. *Plos Pathog*. 11(5):e1004805.

Schönknecht G, Chen W-H, Ternes C, Barbier G, Shrestha R, Stanke M, Bräutigam A, Baker B, Banfield J, Garavito R, et al. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* 339(6124):1207–1210.

Sideri TC, Willetts SA, Avery SV. 2009. Methionine sulphoxide reductases protect iron-sulphur clusters from oxidative inactivation in yeast. *Microbiology*. 155(2):612–623.

Sorhannus U. 2011. Evolution of antifreeze protein genes in the diatom genus *Fragilariopsis:* evidence for horizontal gene transfer, gene duplication and episodic diversifying selection. *Evol Bioinform Online*. 7:EBO.S8321

Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. 2016. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun*. 7(1):11307.

Stairs C, Eme L, Muñoz-Gómez S, Cohen A, Dellaire G, Shepherd J, Fawcett J, Roger A. 2018. Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. *Elife*. 7. doi: 10.7554/eLife.34292.

Stairs CW, Roger AJ, Hampl V. 2011. Eukaryotic pyruvate formate lyase and its activating enzyme were acquired laterally from a firmicute. *Mol Biol Evol*. 28(7):2087–2099.

Strese Å, Backlund A, Alsmark C. 2014. A recently transferred cluster of bacterial genes in Trichomonas vaginalis - lateral gene transfer and the fate of acquired genes. *BMC Evol Biol*. 14(1):119.

Sun G, Huang J. 2011. Horizontally acquired DAP pathway as a unit of self-regulation: gene transfer and metabolic network. *J Evol Biol*. 24(3):587–595

Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Maréchal E, Bowler C, Muto M, Sunaga Y, Tanaka M, et al. 2015. Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *Plant Cell*. 27(1):162–176.

Traller JC, Cokus SJ, Lopez DA, Gaidarenko O, Smith SR, John PM, Gallaher SD, Podell S, Thompson M, Cook O, et al. 2016. Genome and methylome of the oleaginous diatom *Cyclotella cryptica* reveal genetic flexibility toward a high lipid phenotype. *Biotechnol Biofuels*. 9(1):258.

Tsaousis AD, Ollagnier de Choudens S, Gentekaki E, Long S, Gaston D, Stechmann A, Vinella D, Py B, Fontecave M, Barras F, et al. 2012. Evolution of Fe/S cluster biogenesis in the anaerobic parasite Blastocystis. *Proc Natl Acad Sci*. 109(26):10426–10431.

Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2018. A molecular portrait of de novo genes in yeasts. *Mol Biol Evol*. 35(3):631–645.

Vaneechoutte D, Vandepoele K. 2019. Curse: building expression atlases and co-expression networks from public RNA-Seq data. *Bioinformatics* 35(16):2880–2881.

Villar E, Vannier T, Vernette C, Lescot M, Cuenca M, Alexandre A, Bachelerie P, Rosnet T, Pelletier E, Sunagawa S, et al. 2018. The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Res*. 46(W1):W289–W295.

Whitaker JW, McConkey GA, Westhead DR. 2009. The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. *Genome Biol*. 10(4):R36.

Winder M, Cloern JE. 2010. The annual cycles of phytoplankton biomass. *Phil Trans R Soc B*. 365(1555):3215–3226.

Yang M, Lin X, Liu X, Zhang J, Ge F. 2018. Genome annotation of a model diatom *Phaeodactylum tricornutum* using an integrated proteogenomic pipeline. *Mol Plant*. 11(10):1292–1307.

Zakharova Y, Adel'shin R, Parfenova V, Bedoshvili Y, Likhoshway Y. 2010. Taxonomic characterization of the microorganisms associated with the cultivable diatom *Synedra acus* from Lake Baikal. *Microbiology+*. 79(5):679–687.

Zhang L, Li W-H. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol*. 21(2):236–239.