Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes

Richard G Dorrell, Adrien Villain, Benoît Perez-Lamarque, Guillemette Audren de Kerdrel, Giselle Mccallum, Andrew K Watson, Ouardia Ait-Mohamed, Adriana Alberti, Erwann Corre, Kyle R Frischkorn, et al.

Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes

Richard G. Dorrell[a,1], Adrien Villain[b,2,3], Benoît Perez-Lamarque[a,c,2], Guillemette Audren de Kerdrel[a,4], Giselle McCallum[d], Andrew K. Watson[c], Ouardia Ait-Mohamed[a,5], Adriana Alberti[e], Erwann Corre[f], Kyle R. Frischkorn[a,6], Juan J. Pierella Karlusich[a], Eric Pelletier[e,g], Hélène Morlon[a], Chris Bowler[a,g], and Guillaume Blanc[b,1]

[a]Institut de Biologie de l'École Normale Supérieure, Ecole Normale Supérieure, CNRS, INSERM, Université Paris Sciences et Lettres, 75005 Paris, France; [b]Aix Marseille University, Universite de Toulon, CNRS, Institut de Recherche pour le Développement (IRD), Mediterranean Institute of Oceanography (MIO) UM 110, 13288 Marseille, France; [c]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, École Pratique des Hautes Études (EPHE), Université des Antilles (UA), 75005 Paris, France; [d]Department of Biology, Concordia University, H3G 1M8 QC, Montreal, H3G 1M8 QC, Canada; [e]Metabolic Genomics, Genoscope, Institut de Biologie François Jacob, Commissariat à l'Energie Atomique, CNRS, Université Evry, Université Paris-Saclay, 91000 Evry, France; [f]FR 2424 CNRS, Analysis and Bioinformatics for Marine Science, Station Biologique de Roscoff, Université Pierre et Marie Curie Paris 06, 75005 Paris, France; and [g]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans Global Ocean Systems Ecology and Evolution, 75016 Paris, France

Edited by John M. Archibald, Dalhousie University, Halifax, Canada, and accepted by Editorial Board Member W. F. Doolittle November 30, 2020 (received for review May 24, 2020)

Horizontal gene transfer (HGT) is an important source of novelty in eukaryotic genomes. This is particularly true for the ochrophytes, a diverse and important group of algae. Previous studies have shown that ochrophytes possess a mosaic of genes derived from bacteria and eukaryotic algae, acquired through chloroplast endosymbiosis and from HGTs, although understanding of the time points and mechanisms underpinning these transfers has been restricted by the depth of taxonomic sampling possible. We harness an expanded set of ochrophyte sequence libraries, alongside automated and manual phylogenetic annotation, in silico modeling, and experimental techniques, to assess the frequency and functions of HGT across this lineage. Through manual annotation of thousands of single-gene trees, we identify continuous bacterial HGT as the predominant source of recently arrived genes in the model diatom *Phaeodactylum tricornutum*. Using a large-scale automated dataset, a multigene ochrophyte reference tree, and mathematical reconciliation of gene trees, we note a probable elevation of bacterial HGTs at foundational points in diatom evolution, following their divergence from other ochrophytes. Finally, we demonstrate that throughout ochrophyte evolutionary history, bacterial HGTs have been enriched in genes encoding secreted proteins. Our study provides insights into the sources and frequency of HGTs, and functional contributions that HGT has made to algal evolution.

RNAseq | MMETSP | ornithine-urea cycle | phylogenomics | stramenopile

Since their origin, the eukaryotes have split into a multitude of different lineages (1, 2). These comprise several "crown supergroups," some of which (e.g., opisthokonts, including animals and fungi) are exclusively nonphotosynthetic, while others (e.g., archaeplastids, including plants, and stramenopiles) contain photosynthetic members (1, 2). Within these photosynthetic eukaryotes, the initial endosymbiotic acquisition of a chloroplast occurred (3, 4) in a common ancestor of the archaeplastids (green algae and plants, red algae, and glaucophytes), which has subsequently been transmitted into other photosynthetic groups (e.g., cryptomonads, haptophytes, ochrophytes within the stramenopiles, and dinoflagellates) via secondary and tertiary endosymbioses (5, 6). Alongside endosymbiotic transfers, nonendosymbiotic or horizontal gene transfers (HGTs) from bacteria and eukaryotes have shaped the biology of individual eukaryotic groups (7, 8).

The ochrophytes are a diverse group of algae, spanning from single-celled pico-phytoplankton of no more than a few microns diameter (e.g., chrysophytes, pelagophytes), to giant macroalgae of over 100 m in length (e.g., kelps, within the phaeophytes) (5, 9), and originating over 500 million years ago (3, 4). Ochrophytes constitute well over half of the photosynthetic eukaryote barcodes retrieved in the *Tara* Oceans census of marine planktonic life (10), and have important roles as producers and consumers in freshwater habitats (11, 12). Among the ochrophytes, diatoms are particularly important, removing as much carbon dioxide from the atmosphere per year as tropical forests (13, 14).

Genomic inspection of ochrophytes, led by pioneering studies on the diatom *Phaeodactylum tricornutum* (14, 15), have revealed large numbers of endosymbiotically and horizontally acquired genes. These include the probable acquisition of at least 400 genes of red algal origin (14, 16, 17) and between 100 and >1,500

Significance

The ochrophytes are an ancient and important group of eukaryotic algae, including diatoms, the most important photosynthesisers in the modern ocean, and a wide range of other species. Throughout their history, ochrophytes have exchanged genes with bacteria and eukaryotes through horizontal gene transfer (HGT), diversifying their cell biology. Here, we profile thousands of phylogenetic trees, showing that HGTs from bacteria contribute particularly to the recent evolution of diatoms, occurring more frequently than HGTs from eukaryotes and, potentially, more frequently in diatoms than other ochrophytes. Using experimental and computational techniques, we show that bacterial HGTs predominantly impact the diatom secreted proteome, which may contribute to the striking evolutionary success of diatoms.

Author contributions: R.G.D., A.V., B.P.-L., H.M., C.B., and G.B. designed research; R.G.D., A.V., B.P.-L., G.A.d.K., G.M., A.K.W., K.R.F., and G.B. performed research; O.A.-M., A.A., E.C., J.J.P.K., and E.P. contributed new reagents/analytic tools; R.G.D., B.P.-L., A.K.W., and G.B. analyzed data; and R.G.D., B.P.-L., K.R.F., H.M., C.B., and G.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. J.M.A. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: dorrell@bio.ens.psl.eu or guillaume.blanc@mio.osupytheas.fr.

[2]A.V., and B.P.-L. contributed equally to this work.

[3]Present address: BIOASTER, 69007 Lyon, France.

[4]Present address: Centre d'Études Prospectives et d'Informations Internationales, 75007 Paris, France.

[5]Present address: Immunity and Cancer Department, Institut Curie, Paris Sciences and Letters Research University, INSERM U932, 75005 Paris, France.

[6]Present address: Nature Research, New York, NY 10004.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2009974118/-/DCSupplemental.

Published January 8, 2021.

genes from green algae (5, 16, 18). Previous studies have indicated that both the red and green algal genes in the *P. tricornutum* genome are probably tied to the origins of the ochrophyte chloroplast (5, 6). Moreover, the *P. tricornutum* genome contains a large number of genes of bacterial origin [e.g., 584 genes in the version 1 genome annotation (14)], which may also have arrived with the ochrophyte chloroplast or have been acquired through alternative HGTs (5, 6). Many of these bacterially derived genes may facilitate the environmental success of diatoms. For example, genes encoding an ornithine-urea cycle (19, 20), and siderophore transport proteins of inferred bacterial origin (21) may, respectively, enable diatoms to effectively manage intracellular nitrogen reserves and acquire iron from the environment.

Previous studies have identified different numbers of HGTs in ochrophytes, reflecting the techniques used for identifying gene transfers [e.g., phylogenetics and automated tree sorting (8, 16), and manual phylogenetic inspection of specific subsets of genes (5, 18)], along with the size and taxonomic completeness of the reference libraries used (5, 13). For example, many of the "red" and "green" algal gene transfers previously identified in ochrophyte genomes might be derived from other algae with secondary chloroplasts (5, 6). Furthermore, the lack of genomic references available for ochrophytes has limited our understanding of the timing of many of the HGTs associated with ochrophytes, even within recent studies (8). For example, it remains to be determined what HGTs differentiate diatoms from other, less ecologically prominent ochrophytes.

In the last decade, alongside improved genomic sampling (17), over 150 different transcriptome libraries have been sequenced from ochrophytes (22, 23), dramatically expanding our understanding of the functional biology of this lineage. Herein we harness a densely sampled, transcriptome-enriched reference sequence dataset, manual and automated phylogenetic annotations, and experimental characterization to assess the extent, timing, and functions of HGT across the ochrophytes. We manually assess all prokaryotic and eukaryotic HGTs during ochrophyte evolution that are detectable in the genome of *P. tricornutum*, and uncover a continuous flux of genes from bacteria as the predominant contributor of recent HGTs in diatom evolution. Through a larger-scale automated analysis of all genes likely to be derived from HGTs from bacteria across the ochrophytes, we identify large numbers of bacterial HGTs underpinning the evolution of diatoms, following their divergence from other ochrophytes. Using mathematical reconciliation techniques, we identify evidence for a greater frequency of bacterial HGTs in diatoms than other ochrophytes, which may contribute to the striking evolutionary success of diatoms in the contemporary ocean. Finally, using in silico prediction and GFP localization, we demonstrate that HGTs from bacteria, throughout the evolutionary history of ochrophytes, have been enriched in proteins with secretory localizations.

Results

Phylogenomic Profiling Reveals the Composite Eukaryotic and Bacterial History of the *P. tricornutum* Genome. We first wished to determine the most frequent evolutionary identities of the organisms that have donated genes into ochrophytes—and have received genes from ochrophytes—via HGT. To do this, we developed a custom pipeline to generate a global estimate of all HGTs associated with the genome of the model ochrophyte *P. tricornutum*, based on the manual inspection of single-gene trees with a densely sampled dataset of prokaryotic and eukaryotic genomes and transcriptomes (5, 24) (Dataset S1). We divided this dataset into 12 different lineages comprising 151 different taxonomic subcategories (shown in Dataset S1) based on phylogenetic affinity, and previously estimated contributions of different taxa to ochrophyte genome evolution (5, 14). This pipeline was used to trace all genes in the version 3 annotation (17) of the *P. tricornutum* genome that have undergone horizontal transfer since the ochrophyte common ancestor. Our dataset improves on previous analyses by considering HGTs between ochrophytes and other algae with chloroplasts of secondary origin (e.g., haptophytes and dinoflagellates) (5, 6), and by considering the direction of transfer events, to explore the possibility that ochrophytes are donors as well as recipients of HGTs into other branches of the tree of life (17, 21).

We identified 1,979 genes (16.3% of the total *P. tricornutum* genome) (Dataset S1) with potential histories of horizontal transfer during ochrophyte evolution. These included 1,347 genes (11.0% total genome content) where the single gene tree implicated an HGT into ochrophytes at some point in their evolutionary history, and 1,771 HGTs, partitioned over 1,184 genes (9.7% total genome content) from ochrophytes into another branch of the tree of life (Fig. 1*A* and Dataset S1). Eighteen exemplar trees are provided in *SI Appendix*, Fig. S1.

Reflecting the complex patterns of HGT likely to occur in ochrophytes, 547 (4.4% total genome content) gene trees revealed an HGT into the ochrophytes, followed by one or more HGTs out from the ochrophytes into other lineages, while 397 (3.3% total genome content) revealed transfers from ochrophytes into at least two different groups across the tree of life (Dataset S1). In 76 cases, the sister-group to the ochrophytes contained multiple nonstramenopile lineages, consistent with an HGT origin, but with no clear ancestral identity.

Smaller numbers of HGTs were supported by >50% (727 HGTs in; 798 HGTs out), and >80% bootstrap support (415 HGTs in; 464 HGTs out), and phylogeneties with concordant sets of nonochrophyte BLAST top-hits (5, 17) (152 HGTs in; 240 HGTs out). Broadly, these preserved the general dynamics and features discussed below (*SI Appendix*, Fig. S2 and Dataset S1).

Complex Patterns of Horizontal Transfer between *P. tricornutum* and other Eukaryotic Algae. The majority of genes detectable in the *P. tricornutum* genome that could be inferred to be transferred into the ochrophytes (1,007 of 1,347 gene transfers in, 75%) are derived from eukaryotic algae with both primary and secondary chloroplasts (cryptomonads, haptophytes, dinoflagellates, red or green algae), with similar numbers of genes attributed to each (Fig. 1*A* and Dataset S1). Transfers from eukaryotic algae comprised the majority of HGTs recovered with >50% (569 of 727 HGTs in, 78%) and >80% bootstrap support (326 of 415 HGTs in, 79%), and a plurality of HGTs into ochrophytes recovered by both phylogenetic and BLAST analysis (72 of 152 HGTs in, 47%) (*SI Appendix*, Fig. S2). Exemplar trees for genes transferred from eukaryotic algae into ochrophytes are provided in *SI Appendix*, Fig. S1 *A–H*.

The majority of transfers out from the ochrophytes likewise involved eukaryotic algae with secondary chloroplasts (1,747 of 1,771 total HGTs out, 99%), with the greatest number identified in haptophytes (817 of 1,771 inferred transfers out, 46%) (Fig. 1*A* and Dataset S1). HGTs into haptophytes were recovered as the single largest category in trees filtered on 50% (448 of 898 HGTs out, 50%) and 80% bootstrap support (252 of 464 HGTs out, 54%), and in transfers recovered concordantly by phylogenetic and BLAST top-hit analysis (138 of 240 HGTs out, 58%) (*SI Appendix*, Fig. S2). Exemplar tree topologies for genes transferred from ochrophytes into other eukaryotic algae are provided in *SI Appendix*, Fig. S1 *A–K*.

The greatest number of transfers from ochrophytes into haptophytes (383 of 817 HGTs into haptophytes, 47%) were assigned to a common ancestor of Khakista (diatoms, bolidophytes) and Hypogyristea (pelagophytes and dictyochophytes) (Fig. 1*B* and Dataset S1). This result was likewise recovered in trees filtered on 50% (249 of 448 HGTs into haptophytes, 56%) and 80% bootstrap support (145 of 252 HGTs into haptophytes, 58%); and in transfers recovered concordantly by phylogenetic and BLAST top-hit analysis (82 of 138 HGTs into haptophytes, 59%)

(*SI Appendix*, Fig. S2). Analysis of the composition of the clades involved in each transfer (*SI Appendix*, Fig. S3 *A*, *i*) furthermore showed a specific enrichment in pelagophytes and dictyochophytes in the ochrophyte sister-group to the haptophyte clade (making up >40% weighted sister-group presence). Sequences from pelagophytes and dictyochophytes typically reciprocally retrieved haptophyte sequences in internal BLAST top-hit searches of each alignment (χ^2 enrichment $P < 1E\text{-}05$) (*SI Appendix*, Figs. S4 and S5 and Dataset S2). In contrast, this signal was evenly distributed across all haptophyte groups, including the early-diverging pavlovophytes (*SI Appendix*, Figs. S3 *A*, *ii*, S4, and S5 and Dataset S2). This association is specifically concordant with large amounts of HGT from a pelagophyte/dictyochophyte common ancestor into a haptophyte common ancestor, as has been previously proposed (e.g., via an ancient chloroplast endosymbiosis) (5, 21).

Consideration of both clade compositions (*SI Appendix*, Fig. S3*B*) and internal BLAST searches of each alignments (*SI Appendix*, Figs. S4 and S5) also revealed an association between pelagophytes and dictyochophytes with all dinoflagellates sampled, which may suggest a direct or indirect (e.g., via haptophyte) ochrophyte origin of the peridinin dinoflagellate chloroplast (21, 25). This signal was not conserved in the small numbers (26) of gene trees that implied a horizontal transfer from the ochrophytes into other plastid-bearing alveolates [e.g., apicomplexans, and chromerids (25)], which showed a weak enrichment in signal toward synchromophytes, synurophytes, and chrysophytes (*SI Appendix*, Fig. S3*C*). This may be the result of separate HGTs from the ochrophytes into the dinoflagellates and chromerids/apicomplexans, and potentially even separate origins of the dinoflagellate and apicomplexan chloroplasts (25).

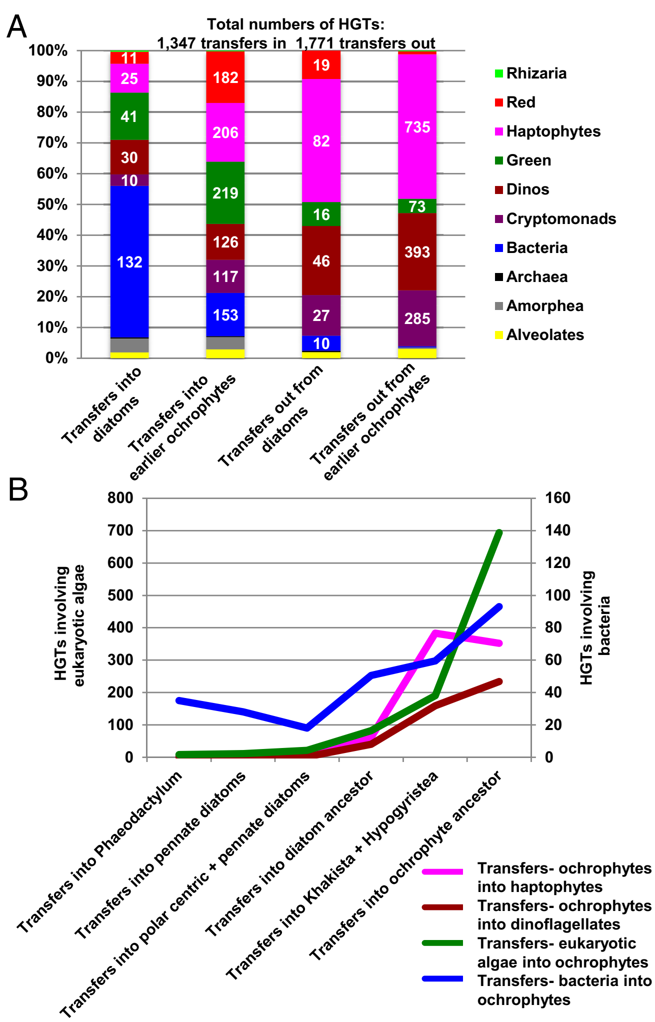

Fig. 1. Dynamics and identity of horizontally transferred genes in the *P. tricornutum* genome. (*A*) Evolutionary origins of 1,347 genes in the version 3 annotation of the *P. tricornutum* genome annotated by manual inspection of phylogenies as having been transferred into ochrophytes from nonstramenopile sources, and 1,771 gene transfers identified out of ochrophytes into nonstramenopile groups. Frequencies are shown for the most abundant contributors to each HGT category; full values are tabulated in Dataset S1. (*B*) Number of HGTs involving different groups of eukaryotic algae (shown on the left vertical axis) and bacteria (right axis) and ochrophytes, manually resolved to (left four) four time points in diatom evolution, and (right two) two earlier points in ochrophyte evolution.

Only a small proportion of the transfers involving eukaryotic algae detectable in the *P. tricornutum* genome occurred after the divergence of diatoms from other ochrophytes (123 of 1,007 eukaryotic algal HGTs in; 12%; 194 of 1,747 eukaryotic algal HGTs out, 11%) (Fig. 1*B*). This result was also recovered when considering only HGTs with >50% bootstrap support (77 of 569 eukaryotic algal HGTs in, 14%; 69 of 896 eukaryotic algal HGTs out, 7%); >80% bootstrap support (51 of 326 eukaryotic algal HGTs in, 16%; 28 of 462 eukaryotic algal HGTs out, 6%); and transfers supported by BLAST top-hit analysis (11 of 72 eukaryotic algal HGTs in, 15%; 31 of 238 eukaryotic algal HGTs out, 13%) (*SI Appendix*, Fig. S2).

Continuous Bacterial HGT Dominates Recent Arrivals in the *P. tricornutum* Genome. We identified a substantial number of genes (285, 2.3% total genome content) that have been transferred into the *P. tricornutum* genome, since the ochrophyte common ancestor, from bacterial partners (Fig. 1*A* and Dataset S1). The transfer events from bacteria are largely asymmetric, with only 18 plausible cases of genes across the *P. tricornutum* genome (0.1% total genome content) transferred into bacterial groups from ochrophyte donors (Fig. 1*A* and Dataset S1). Smaller numbers of bacterial HGTs into ochrophytes could be detected with 50% bootstrap support (146 HGTs in), 80% bootstrap support (81 HGTs in), and concordant BLAST top-hit and phylogenetic annotations (76 HGTs in); effectively no HGTs out from ochrophytes to bacteria received more than minimal bootstrap support (*SI Appendix*, Fig. S2).

The gene transfers inferred from bacteria into ochrophytes were attributed nearly equally to having occurred within diatoms (132 HGTs in) and earlier time points (153 HGTs) (Fig. 1 and Dataset S1). Exemplar trees showing HGTs from bacteria into early ancestors of the diatom lineage are provided in *SI Appendix*, Fig. S1 *M*–*O*, while trees documenting recent bacterial HGTs in the evolutionary history of *P. tricornutum* are shown in *SI Appendix*, Fig. S1 *P*–*R*. Bacteria were responsible for nearly half of the gene transfers within diatoms detectable in *P. tricornutum* (49%, from 267 HGTs into diatoms) (Fig. 1*A* and Dataset S1), and in trees with >50% bootstrap support (74 of 152 HGTs into diatoms, 48%); >80% bootstrap support (38 of 90 HGTs into diatoms, 42%), and a majority of transfers with concordant BLAST top-hit and phylogenetic annotations (59 of 71 HGTs into diatoms, 83%) (*SI Appendix*, Fig. S2).

The bacterial HGTs were inferred to be derived from a wide range of taxa, considering BLAST top-hit assignation (Dataset S1). Notably, Phatr3_Jdraft542.p1 shows highest BLAST similarity with sequences from the nitrogen-fixing diatom endobiont *Richelia intracellularis* (27); and Phatr3_J31257.p1 shows greatest similarity to the dinoflagellate epibiont *Labrenzia aggregata* (28) (Dataset S1), suggesting possible origins from bacterial symbionts.

Thousands of Bacterial Sequences Detectable in Ochrophyte Genomes and Transcriptomes. Given the overwhelmingly bacterial signature of HGTs in diatoms that can be detected in the *P. tricornutum* genome, we wished to quantify how prevalent bacterial HGT was

EVOLUTION

elsewhere in the ochrophyte tree. We identified genes that had phylogenetic histories compatible with prokaryotic HGTs based on 162 ochrophyte genomes and transcriptomes, including 109 diatom sequence libraries, of which 7 transcriptomes were generated for this study (Dataset S1), and a custom BLAST and phylogenetic pipeline (*Materials and Methods* and *SI Appendix*, Fig. S6). A total of 41,620 (1.6%) ochrophyte sequences yielded exclusively noneukaryotic BLAST matches (outside of other ochrophytes), hence were considered prokaryote-derived sequences de facto, and a further 46,444 (1.7% total) candidates were inferred to have prokaryotic origins through automated phylogenetic analysis (*Materials and Methods* and *SI Appendix*, Fig. S6). We present consensus MrBayes and RAxML topologies for six exemplar clusters in *SI Appendix*, Fig. S7.

The majority (97%) of the detected sequences were from bacterial (as opposed to archaeal, or viral) lineages (*SI Appendix*, Fig. S8). The donor bacterial groups identified were diverse, although were dominated by Proteobacteria, Bacteroidetes, Terrabacteria, and Planctomycete-Verrumicrobiaceae-Chlamydialean taxa, consistent with biases observed in previous diatom HGT analyses (8, 14, 17). No overrepresentation of genes originating from known endosymbiont bacteria was detected: only 38 ochrophyte gene families had their closest sister group composed of Rickettsiales, Holosporales, Legionellales, or Chlamydiae lineages (*SI Appendix*, Fig. S8). A limited number of genes (397) of viral origin were also predicted, of which the greatest (139, of predominant Phycodnaviridae origin) were restricted to the brown alga *Ectocarpus*, consistent with previous reports of integrated viral segments in this genome (29, 30) along with 493 genes across ochrophytes of apparent archaeal origin.

Next, we searched within our data for bacterial sequences that we could clearly identify as HGTs, rather than bacterial contaminants in individual ochrophyte libraries (25). A small number of ochrophyte libraries [principally the genome of the diatom *Synedra acus* (31), and the diatom transcriptomes for *Thalassiosira miniscula*, *Chaetoceros* sp. *UNC2102*, *Chaetoceros* sp. *neogracile*, and *Chaetoceros* sp. *curvisetus*, and the transcriptome of the dictyochophyte *Pseudopedinella elastica* (22)] contained visibly abnormal proportions of bacterial sequences confined to these species (*SI Appendix*, Fig. S9). To verify whether single-species HGT events in our dataset might be influenced by contamination, we compared GC content, as well as the distribution of unique 5mer patterns in bacterial sequences uniquely found in ochrophyte species, to species-specific models of nucleotide composition generated from a training dataset of 902,839 "vertically inherited" genes inferred from this library (*SI Appendix*, Fig. S10). The bacterial sequences identified in only one ochrophyte library frequently had different normalized GC contents and 5mer composition distributions than vertically inherited genes (*SI Appendix*, Fig. S7). In contrast, bacterial sequences found in more than one ochrophyte library had normalized GC contents and 5mer compositions that overlapped with those of vertically inherited genes (*SI Appendix*, Fig. S10), suggesting that they correspond to bacterial HGTs that have drifted toward the sequence composition of their recipient ochrophyte genome.

To test the effectiveness of the automated tree assignations, we compared the set of *P. tricornutum* genes inferred to be compatible with bacterial HGT through automated analysis with those detected in the previous manual assignation of *P. tricornutum* HGTs (*SI Appendix*, Fig. S11 and Dataset S3). A total of 37 genes identified in the automated analysis were independently confirmed manually to have undergone bacterial HGT; only 2 of the genes in the automated analysis were found to have vertical origins in manual assignations, and 5 were inferred to have originated through horizontal transfers from nonbacterial sources (Dataset S3). We identified concordant topologies for genes that either retrieved and for genes that did not retrieve eukaryotic homologs in the automatically reconciled gene tree, and for gene trees that were reconciled with >70% and <70% SH-aLRT branch support (*SI Appendix*, Fig. S11). We note, moreover, that the number of HGTs per node predicted with low SH-aLRT branch supports (i.e., <70%) is strongly correlated (R = 97.6) with the number of HGTs predicted with higher branch supports (i.e., >70%), demonstrating that the two HGT categories contained the same signal (*SI Appendix*, Fig. S12). We chose accordingly to work with the bacterial HGTs found in more than one ochrophyte library for downstream analysis, regardless of branch support or presence of eukaryotic homologs.

As an explicit detection of genuine bacterial HGTs, we used a previously defined BLAST methodology (32) to search for genes that contain composite bacterial and ochrophyte domains, which arise through the recombination of bacterial and host sequences (*SI Appendix*, Fig. S13 and Dataset S5). We identified 12 genes containing composite bacterial and eukaryotic domains that are shared between more than one ochrophyte library, hence are unlikely to be sequence assembly artifacts (*SI Appendix*, Fig. S13). These include, for example, a rhamnose transporter of proteobacterial origin, which has fused to a G protein-coupled receptor domain of ochrophyte origin in the polar centric diatoms *Minutocellus polymorphus*, *Extubocellulus spinifer*, and *Skeletonema dohrnii* (*SI Appendix*, Fig. S14).

Bacterial HGTs Occur at Greater Frequency in Diatoms than Other Ochrophytes. Considering the large contribution of bacteria for HGTs into diatoms (Fig. 1*B*), we wished to determine whether diatoms have undergone greater numbers of bacterial HGTs than other ochrophytes. We mapped the distribution of 2,796 sequence clusters compatible with HGTs from bacteria into ochrophytes and shared across at least 2 ochrophyte sequences to a 63-gene tree covering 213 species, including all 161 ochrophyte libraries referenced in this study (Fig. 2 and Dataset S4). This topology confirms phylogenetic relationships proposed by previously published phylogenies: placing *Attheya* as the sister-group to all pennate diatoms and *Striatella* as the earliest-diverging pennate diatom genus (33, 34), and identifying two distinct early-branching ochrophyte clades, the first consisting of pinguiophytes, synchromophytes, and chrysophytes, and the second consisting of eustigmatophytes, phaeophytes, xanthophytes, and raphidophytes (35) (Fig. 2). Topologies were verified by the consideration of the single-gene trees used to generate the final concatenation; by long-branch and fast-site removal approaches, and by the serial removal of individual taxa from the concatenated alignment (*SI Appendix*, Figs. S15–S17).

We noted that many HGTs could be traced to foundational nodes in diatom evolution. The largest number (462) were identified on the node for the common ancestor of polar centric and pennate diatoms (Fig. 2). Over 100 individual transfers were also resolved to the common ancestors of all diatoms, of Khakista and Hypogyristea, and of pennate and polar centric diatoms, and radial centric diatoms within the Coscinodiscophytes (Fig. 2). In contrast, none of the nondiatom ochrophyte nodes considered were marked by >50 HGTs (Fig. 2). Similar results were detected when only considering HGTs reconciled with >70% bootstrap support (Fig. 2, *SI Appendix*, Fig. S18, and Dataset S3).

We then wished to determine whether HGTs were truly more frequent in diatoms than in other ochrophytes. We noted limited differences in the missing fraction of BUSCO libraries (36) across ochrophytes, with most libraries presenting >60% completeness (*SI Appendix*, Fig. S19*A*), suggesting that library completeness does not account for the enrichment of HGTs detected in diatoms. Next, we used amalgamated likelihood estimation (ALE) (37, 38), a probabilistic gene reconciliation approach that can be used to infer HGT origins, to select, among the 675 HGT proteins present in at least 10 ochrophyte species, those that were inferred to be predominantly vertically inherited (and are therefore not contaminations or multiple independent acquisitions) (Fig. 3*A*). This ALE-verified dataset contained 435 HGT protein clusters that originated up to 10 times more frequently within diatoms compared to other ochrophyte lineages (t test, $t_{df = 327} = 3.4$, $P = 0.0008$) (Fig. 3*B*, *SI Appendix*, Fig. S19*B*, and Dataset S4). We confirmed using randomizations that this diatom enrichment in bacterial HGTs was not due to random chance (Fig. 3*C*). We also found that many of the bacterial HGT proteins have been subsequently lost in diatoms (*SI Appendix*, Fig. S19*C*).

Finally, we applied subsampling controls in which we compared the number of ALE-verified HGTs attributable to diatom-inclusive groups or nondiatom ochrophytes when randomly sampling 10 selected species, over 100 replicates. We noted elevated frequencies in diatom versus nondiatom groups, considering total HGTs identified (*SI Appendix*, Fig. S20) and total HGTs assigned to the common ancestor of each group (*SI Appendix*, Fig. S21). Thus, even when sampling of diatoms and other groups are similar, our data indicate elevated frequencies of bacterial HGT into diatoms compared to other ochrophytes.

Functional Insights into the Roles of Bacterial HGTs in Ochrophytes. Next, we considered what functional annotations might be associated with bacterial HGTs into ochrophytes. Using the *P. tricornutum* genome as a well-annotated model, we profiled biological functions, structural homology (17), and predicted localizations (39–42) of all genes of predicted bacterial origin extending to the ochrophyte ancestor (Dataset S6), considering genes identified in manual annotations of *P. tricornutum* trees (Dataset S1) and *P. tricornutum* genes identified in automated HGT analyses of all ochrophyte genomes and transcriptomes (Dataset S3).

The bacterial HGTs identified in *P. tricornutum* perform diverse functions. These included three vacuolar iron transport proteins (Phatr3_J37632.p1, Phatr3_J43313.p1, and Phatr3_J43314.p1) (43), most likely transferred into the diatom common ancestor from an actinobacterial donor (*SI Appendix*, Fig. S1*M*). The inclusion of these proteins reflects the likely bacterial origin of iron-scavenging pathways unique to diatoms (21). Our dataset also included cytoplasmic enzymes involved in putative siderophore (Phatr3_J46626.p1, isochorismatase), and menaquinone (Phatr3_J47199.p1/Phatr3_J48624.p1, demethylmenaquinone methyltransferase, Phatr3_J44670.p1, 1 4-dihydroxy-2-naphthoate polyprenyltransferase) biosynthesis, reflecting the probable role of HGT for diversifying eukaryotic secondary metabolism (44, 45).

Other pathways previously annotated as being bacterially derived in *P. tricornutum* were sparsely recovered in our dataset. For example, in the case of the diatom ornithine-urea cycle, which was originally proposed to contain multiple genes derived from bacterial HGT (20, 46), comparative analysis of our data (Dataset S6) and other published analyses (17, 19, 47) reveals a mosaic origin. These include enzymes likely to have arisen from bacterial HGT (Phatr3_J35643.p1, mitochondrial ornithine cyclodeaminase; Phatr3_J24238.p1, mitochondrial carbamate kinase; and Phatr3_J3969.p1, chloroplast acetylglutamate kinase), alongside others derived from within-eukaryotic HGT (e.g., Phatr3_J29702.p1, mitochondrial urease, acquired from dinoflagellates; and Phatr3_J50577.p1/ J54222.p1, chloroplast acetylornithine transaminase, acquired from haptophytes or green algae; *SI Appendix*, Fig. S1*H*), or with presumed vertical origin. These data underlie the importance of interaction between host- and HGT-derived proteins for the elaboration of diatom metabolism (5, 48).

Bacterial HGTs Frequently Encode Proteins with Secretory Localizations. We considered what localization patterns may be associated with ochrophyte proteins of bacterial HGT origin. Across our entire bacterial HGT dataset, we observed a strong enrichment in proteins with signal peptides, but nonchloroplastic localizations [assessed using signalP/ASAFind (41, 49) and HECTAR (40)], and secretory consensus predictions (either extracellular or plasma membrane) with WolfPSort (42). We identified these enrichments among all genes inferred to originate from bacterial HGTs throughout ochrophyte evolution (Fig. 4*A* and Dataset S3) and among bacterial HGTs detectable in manual annotations of the *P. tricornutum* genome (*SI Appendix*, Fig. S22).

We tested the localization of proteins encoded by 10 genes of inferred bacterial origin detected in the *P. tricornutum* genome, using full-length GFP C-terminal overexpression constructs (Dataset S6). Four of the proteins studied had clear secretory localizations. Phatr3_J33371.p1, a sialate O-acetylesterase manually annotated to have been acquired by a pennate diatom ancestor from bacteria; and Phatr3_J42871.p1, a protein of unknown function conserved across pennate and polar centric diatoms, as well as radial centric diatoms within the coscinodiscophytes, showed clear localization around the endoplasmic reticulum, verified using ER-Tracker staining (Fig. 4*B* and *SI Appendix*, Fig. S22). Furthermore, Phatr3_J50959.p1, a homolog of the bacterial CoA-binding protein YccU conserved across all ochrophytes (*SI Appendix*, Fig. S7*C*), as well as Phatr3_J8596.p1, a protein of unknown function shared across all ochrophytes except for chrysophytes, synurophytes, and pinguiophytes localized to the plasma membrane (Fig. 4*B*).

Not all of the bacterial HGTs assessed either through in silico or GFP analysis localized to the endomembrane system, with other HGTs experimentally localized to the pyrenoid, mitochondria, and cytoplasm (*SI Appendix*, Fig. S23). We noted a weak enrichment in chloroplast-targeting predictions associated with bacterial genes in ochrophyte genomes. Within the genes manually identified from *P. tricornutum*, this was only found to be significant ($P < 0.01$) with ASAFind but not HECTAR, and was much weaker than equivalent biases (χ^2 $P < 10^{-10}$) for genes inferred to have been horizontally transferred with eukaryotic algae (Dataset S1). Within the ochrophyte-wide bacterial HGT dataset, this enrichment was only detected in bacterial HGTs acquired at early time points in ochrophyte (Fig. 4*A*). This likely reflects that some of the earliest bacterial transfers in ochrophyte evolution occurred as part of a chloroplast endosymbiosis event (5), whereas more recent bacterial HGTs have nonchloroplast origins.

Time-Dependent Dynamics of Bacterial HGT Expression Patterns. We also considered whether bacterially derived genes in ochrophyte genomes had specific epigenetic characteristics or expression patterns, focusing on orthologs from *P. tricornutum*, for which substantial epigenetic (17, 50, 51) and quantitative gene-expression data are available (52). Considering structure and location, we noted that bacterial HGTs contain fewer introns per average (0.58 per gene) than other genes across the *P. tricornutum* genome (0.71 per gene, one-way ANOVA $P = 0.03$); although 107 of 298 (36%) contained at least one intron (*SI Appendix*, Fig. S24*A*). We found no obvious biases in the frequencies of intron retention or exon splicing (17) documented in bacterial HGTs compared to other genes (Dataset S1).

We identified little preference in the distribution of HGT genes across *P. tricornutum* chromosomes, barring a possible ($P < 0.05$, χ^2) enrichment on chromosomes 9 and 32 (Dataset S1). We noted that 14 of 298 (4.8%) of the bacterial HGTs were adjacent, and in the same transcriptional orientation, to other bacterial HGTs, whereas only 6 (2.1%) were adjacent to other bacterial HGTs but in opposing transcriptional orientation (*SI Appendix*, Fig. S24*B* and Dataset S1). However, structural inspection of the bacterial HGTs that are adjacent to one another suggest that they predominantly possess similar functions to one another, suggesting that they are likely the result of recent paralogy events, as opposed to being a remnant of horizontally acquired bacterial operons (*SI Appendix*, Fig. S24*B*).

Considering epigenetic marks associated with *P. tricornutum* genes, genes we could manually assign as having undergone HGTs with eukaryotic algae during ochrophyte evolution were enriched in activating histone marks (H3K9Ac, H3K4Me2) and depleted in repressive histone marks (H3K9Me2, H3K9Me3, H3K27Me3; χ^2 test, $P < 0.01$), and had higher mean reads per million kilobases

EVOLUTION

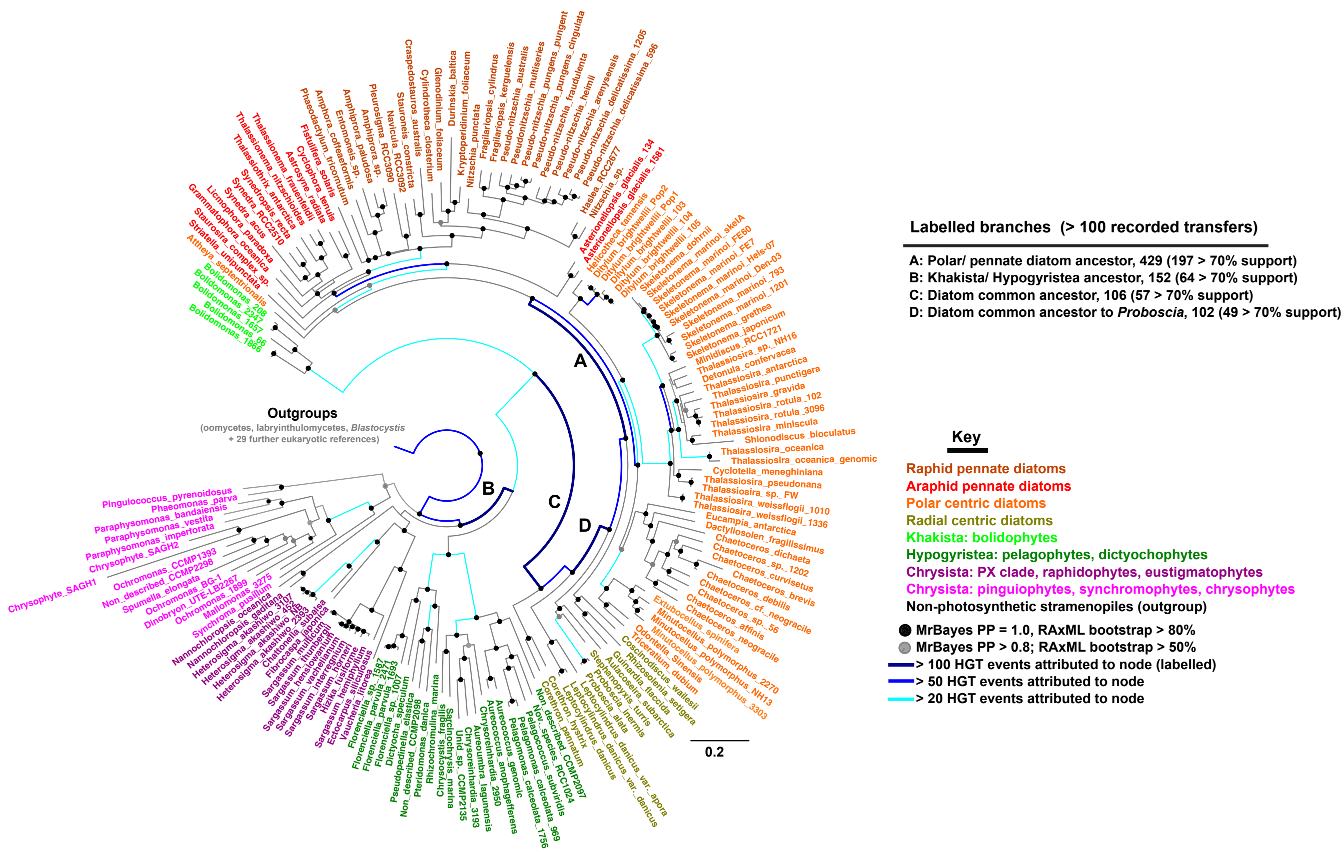

Fig. 2. Distribution of 2,796 bacterial HGTs automatically identified across ochrophyte genomes. This figure shows the consensus MrBayes (GTR, WAG) and RAxML (GTR, JTT, WAG) topology of a 213 taxa by 26,399-aa concatenated alignment, consisting of 63 genes with high occupancy in 161 ochrophyte genomes and transcriptomes, rooted on a set of 52 nonochrophyte outgroups (collapsed here to show only aplastidic stramenopile groups). Leaf names are shaded by evolutionary origin. Support values for each node are shown with shaded circles. The total number of bacterial HGTs attributed to each node by MRCA analysis is depicted by branch color and thickness; four branches with >100 HGTs are individually labeled. An analogous tree topology, showing only the frequencies of HGTs reconciled with >70% bootstrap support, is provided in *SI Appendix*, Fig. S18.

mapped (RPKM) values (456.7) in transcriptomes generated under replete conditions than vertically inherited genes (340.6, $P < 0.01$, one-way ANOVA) (*SI Appendix*, Fig. S22). In contrast, the *P. tricornutum* genes with bacterial HGT origins did not significantly deviate, either in histone marks or mean RPKM value (304.3) from other genes in the genome (*SI Appendix*, Fig. S22), suggesting an absence of particular transcriptional activation compared to vertically derived genes.

Finally, we exploited previously published metastudies of microarray (53) and RNA-sequencing (RNA-seq) data (52) from *P. tricornutum* lines subject to light (53), phosphate (26), nitrogen (54), and iron stress (55) to investigate the coexpression patterns of bacterial HGTs. Bacterially derived genes from the *P. tricornutum* genome show distinct patterns of coregulation, falling within different modules of coregulated genes to genes horizontally transferred with eukaryotic algae (*SI Appendix*, Fig. S22). Moreover, genes inferred to have been acquired within the pennate diatoms exhibited stronger transcriptional coregulation with one another than did bacterially derived genes inferred to have been acquired earlier in ochrophyte evolution (*SI Appendix*, Fig. S25). This may suggest that recently acquired HGTs show distinctive expression patterns compared to HGTs with more ancient acquisitions.

Discussion

In this study, we dissected two independently curated surveys of HGT in diatoms: a targeted survey of all HGTs, involving both prokaryotic and eukaryotic partners into and out of the ochrophytes throughout their evolutionary history that are detectable in the genome of the model diatom *P. tricornutum* (Fig. 1 and *SI Appendix*, Figs. S1–S5), and an automated survey of all genes with phylogenetic histories compatible with bacterial HGT across the diversity of ochrophyte genomes and transcriptomes (Fig. 2 and *SI Appendix*, Figs. S6–S12). These data reveal concordant features in the diversity (*SI Appendix*, Figs. S8, S13, and S14), dynamics (Fig. 3 and *SI Appendix*, Figs. S15–S21), and functions (Fig. 4 and *SI Appendix*, Figs. S22–S25) of gene transfer from bacteria into ochrophytes. Both the taxonomically larger transcriptome-enriched dataset and the combination of automatically and manually resolved phylogenies allowed the detection of a wide range of different HGTs (e.g., horizontal transfers involving other algae with secondary chloroplasts) (Fig. 1), and an improved resolution of the probable time points when each transfer occurred (Figs. 2 and 3). Our data underscore how HGTs from a range of different sources have impacted on the evolution of eukaryotic micro-organisms, a concept that has been the subject of extensive debate (56–58). Further phylogenetic reconstructions, using denser taxonomic sampling strategies may detect even greater numbers of HGTs in *P. tricornutum*.

We note the specific importance of bacterial HGTs in the recent evolution of diatoms. Considering the diversity of HGTs detectable in *P. tricornutum*, we show that HGT from bacteria has occurred continuously throughout ochrophyte evolution, including the recent evolutionary history of diatoms, in contrast to HGT with eukaryotic partners, which is predominantly restricted to earlier time points (Fig. 1). The bacterial HGT is largely unidirectional (Fig. 1), with few identifiable gene transfers from ochrophytes into bacteria, which may reflect transcriptional or

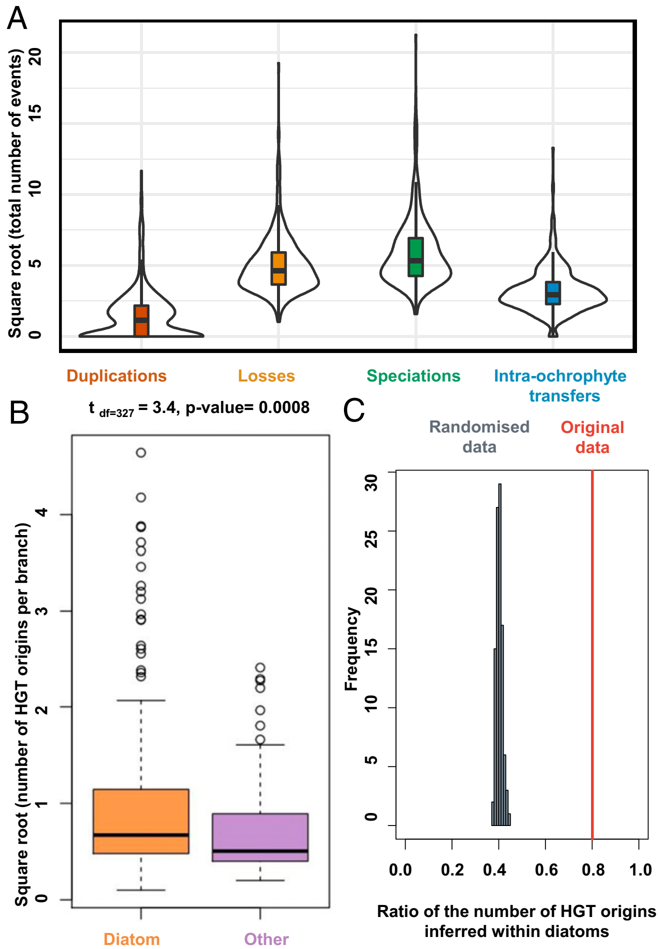

Fig. 3. ALE inference of HGT origins in ochrophytes. (*A*) Distribution of the total number of duplications, losses, speciations, and intraochrophyte transfers averaged over the set of 100 ALE reconciliations for each of the 435 HGT protein clusters inferred as single bacterial HGT origins. (*B*) Distribution of the number of HGT origins per branch inferred by ALE for branches within diatoms (orange) or outside diatoms averaged over the 100 reconciliations. The difference (*t* test) between the distributions is reported on the top of the plot. (*C*) Ratio of the number of HGT origins inferred within diatoms in the original data (red) compared to the same ratio observed when the data are randomized (gray, null expectations obtained by shuffling the species names).

post-translational features that impede their functional expression in bacteria (59, 60). Considering the frequencies of bacterial HGTs identifiable across multispecies ochrophyte data, we furthermore found large numbers of bacterial HGTs at foundational points in the diatom backbone (Fig. 2), which occur at higher frequencies in diatoms than in other ochrophyte groups (Fig. 3 and *SI Appendix*, Figs. S4 and S18–S21). Ultimately, our data point to a continuous bombardment of ochrophyte genomes with bacterial DNA, which may have occurred more substantially in diatoms than other ochrophytes. We note that large numbers of bacterial HGTs into the diatom backbone have been recently detected in a parallel study, using a dataset composed of genomic sequences only (58), underscoring that this result is unlikely to be artifactual.

The discrepancy between the number of bacterial HGTs attributed to early points in diatom evolution (e.g., >450 potential HGTs occurring in a common ancestor of polar centric and pennate diatoms) (Fig. 2) and the number of bacterial genes detectable in extant diatom genomes (e.g., 295 retrieved in *P. tricornutum*) (Fig. 1) is intriguing. While some of these discrepancies may be due to different reference data that is automated versus manual annotation, it is possible that there is a high rate of HGT loss within diatoms. This might be consistent with bacterial HGTs supporting recurrent adaptations to the environment, in which case genes might be frequently lost as a result of streamlining pressure (8, 61). Accounting for missing gene models, we project significantly higher frequencies of HGT losses in diatoms than in nondiatom ochrophytes (Fig. 3*A* and *SI Appendix*, Fig. S19*C*).

We additionally notice specific biases in the functions attributed to bacterial HGTs. While early bacterial HGTs in ochrophyte genomes are enriched in chloroplast-targeting predictions, suggesting origins associated with the ochrophyte chloroplast, more recently arrived bacterial HGTs are enriched in secretory proteins, and show distinct patterns of expression (Fig. 4 and *SI Appendix*, Figs. S22 and S25). A secretory bias (7) has been previously inferred for horizontally acquired genes in the basal stramenopile *Blastocystis*, and diminished expression levels of horizontally acquired genes (as also inferable for bacterial HGTs compared to algal HGTs in our dataset) has previously been shown in many microbial eukaryotes (8, 62), suggesting that these may be general functional features of HGT.

It remains to be determined why large numbers of bacterial HGTs have occurred into diatoms and, indeed, how bacterial HGT has facilitated the success of diatoms in the contemporary ocean. We note that diatoms in the wild typically coexist with bacterial exo- and endosymbionts, with which they are particularly dependent for both nutrient uptake (21, 63, 64), and which may even attach to diatom frustules (65). These interactions may provide a potential route for the horizontal exchanges of genes that facilitate symbiotic interactions. We detect a small number of genes with closest evolutionary affinity to known bacterial symbionts of algae in our data, albeit without obvious quantitative enrichment (Dataset S1). Further studies, with greater taxonomic sampling of ochrophyte lineages known to engage in phagotrophy [e.g., chrysophytes (11, 66)], will be invaluable for confirming whether there is a specifically higher frequency of bacterial HGT into diatoms compared to other ochrophytes.

More broadly, it remains to be determined how HGT has allowed diatoms to explore different niches. This may be accomplished via the functional characterization of individual genes—for example, using transformable model species such as *P. tricornutum* (14)—while the inspection of environmental sequence data, for example from *Tara* Oceans, will be key to understanding how bacterial HGT underpins the versatility of diatoms in the modern ocean (67, 68). Functional characterization of horizontally transferred genes may provide insights into the ecological success of diatoms, and the intertwined genetic landscapes of eukaryotic algae throughout their evolutionary history.

Materials and Methods

Transcriptome Data. Transcriptome libraries were generated for seven diatom species: These were *Guinardia flaccida* RCC3088, *Haslea* sp. RCC2677, *Minidiscus* sp. RCC1721, *Navicula* sp. RCC3092, *Pleurosigma* sp. RCC3090, *Shionodiscus bioculatus* RCC1991, and *Synedra* sp. RCC2510 (Datasets S1 and S3). These species were selected based on their high abundances in *Tara* Oceans metabarcode data (10, 69). Cultures were grown in f/2 medium enriched with silicate, at a 19 °C constant temperature, in a 12-h 30-μE m^{-2} s^{-1}:12-h dark light cycle.

Total cellular RNA was harvested from exponential phase cultures using TRIzol reagent (Invivogen) and precipitated in RNase-free isopropanol, as previously described (5). One microgram RNA, as quantified with a nanodrop, was treated with one U RNase-free DNase (Promega) following the manufacturer's instructions, and repurified with isopropanol. RNA integrity was confirmed using agarose gel electrophoresis.

RNA-seq was performed at Genoscope. RNA-seq libraries were prepared from 1 μg total RNA using the TruSeq Stranded mRNA kit (Illumina), which

A

		% ASAFind-endomembrane	% HECTAR-signal peptide	% WolfPSort-secreted	% WolfPSort-Plasma membrane	% ASAFind-chloroplast	% HECTAR-chloroplast
Totals	All ochrophyte genes	5.7	11.8	10.9	13.4	6.0	2.4
	All HGTs	11.2	14.6	16.8	17.1	7.1	2.2
Ochrophyte	All ochrophytes	8.1	13.8	14.6	15.9	9.1	3.1
	Khakista/ Hypogyristea + PRE clade	9.8	12.5	17.9	9.7	16.1	10.3
	Khakista/ Hypogyristea ancestor	9.6	14.5	14.2	20.8	10.1	3.4
	Khakista	11.1	16.9	16.8	18.1	6.8	2.2
Diatom backbone	Diatom ancestor	15.8	21.0	18.0	15.4	8.7	0.8
	Diatoms to Leptocylindrus	13.3	17.2	12.2	29.2	8.6	2.1
	Diatoms to Probosciales	11.5	14.8	17.0	12.9	6.7	1.9
	Diatoms to Coscinodiscales	12.7	14.5	17.5	12.8	7.9	3.6
	Polar/ pennate diatom ancestor	10.3	13.3	13.5	16.5	7.0	2.4
Within diatoms	Polar centric diatoms	13.8	17.6	13.3	19.0	6.0	0.7
	Ditylum ancestor	10.4	16.7	14.4	9.6	7.1	0.0
	Skeletonema/ Thalassiosirales	18.1	19.2	23.3	20.9	5.5	3.2
	Raphid pennates + Fragilariopsids	10.9	15.0	14.1	16.3	4.9	0.7
	Raphid pennate diatoms	20.2	23.3	22.8	11.4	6.9	1.6

Enriched. $P < 10^{-10}$

Enriched. $P < 0.01$

Depleted. $P < 0.01$

B

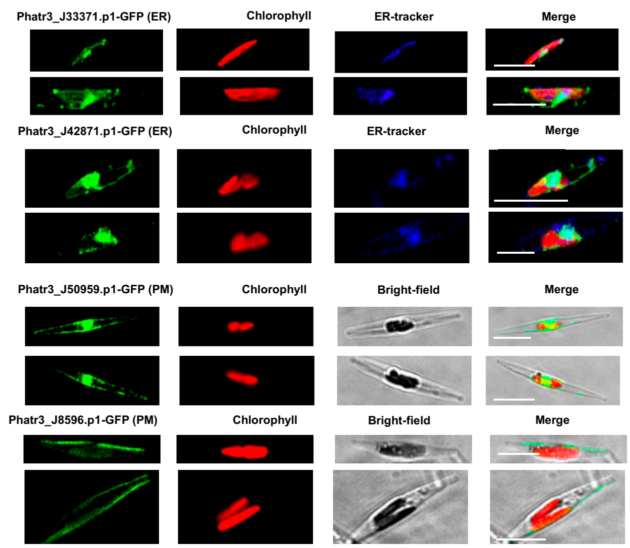

Fig. 4. Continuous secretory enrichment of bacterial HGTs in ochrophytes. (*A*) Heatmap showing the percentage of all ochrophyte genes in the combined 162 species genome and transcriptome library, all genes in the 2,786 identified bacterial HGT clusters, and all HGT genes that originate at different time points in ochrophyte evolution that possess different targeting predictions. Cells are shaded by the significance of enrichment, inferred by (two-tailed) χ^2 tests against the total number of genes in species contained within the corresponding node. (*B*) GFP-fluorescence constructs of four exemplar genes in the *P. tricornutum* genome with endomembrane or secretory localizations and bacterial origins: *Phatr3_J33371.p1* and *Phatr3_J42871.p1*, with inferred endoplasmic reticulum localizations (verified by ER-Tracker); and *Phatr3_J50959.p1* and *Phatr3_J85969.p1*, proteins with partial plasma membrane or cell wall localizations. (Scale bars, 10 μm.)

allows mRNA strand orientation (sequence reads occur in the same orientation as antisense RNA). Poly(A) + RNA was selected with oligo(dT) beads, chemically fragmented and converted into single-stranded cDNA using random hexamer priming. After second-strand synthesis, double-stranded cDNA was 3′-adenylated and ligated to Illumina adapters. Ligation products were PCR-amplified following the manufacturer's recommendations. Ready-to-sequence Illumina libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (Kapa Biosystems), and library profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies). Each library was sequenced using 101-bp paired-end reads chemistry on a HiSeq 2000 Illumina sequencer and, on average, 25 billion paired-end raw reads were obtained for each sample.

Trimming was performed using an in-house software based on the FastX package (hannonlab.cshl.edu/fastx_toolkit/index.html). Raw data were filtered to remove clusters with excessive intensity in bases other than the base called. Adapters and primers were removed, along with all nucleotides after the first nucleotide with quality value lower than 20. Reads shorter than 30 nucleotides after trimming, and reads and their mates that mapped onto run quality-control sequences (PhiX genome) were serially removed. Finally, ribosomal RNA-like reads were excluded using SortMeRNA (70). Assembly was performed by Genoscope using trinity-v2.4.0 (71) under default parameters. Sequencing data are archived in the European Nuclear Archive (ENA) https://www.ebi.ac.uk/ena/data/view/PRJEB34158 under the accession numbers ERR3497271 to ERR3497276.

Peptide sequences for de novo transcriptomes sequenced for this study were generated using TransDecoder under default conditions. Peptide sequences for transcriptome data for which translations have not been made publicly available [e.g., independent chrysophyte transcriptomes (24, 66)] were generated using getorf using -table 0, -minsize 100, -find 1 as default settings.

Identification of all Genes of Horizontal Origin in the *P. tricornutum* Genome. Possible HGTs in the *P. tricornutum* genome since the radiation of the ochrophytes were manually identified using a pipeline adapted from previous studies (5, 17, 24). Each peptide sequence in the version 3 annotation of the *P. tricornutum* genome (17), available on ENSEMBL (protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index), was searched using BLASTp against a composite library consisting of a complete version of uniprot (accessed June 2018) (72), decontaminated versions of the MMETSP (5, 22) and 1-kp transcriptome projects (23, 73), the seven diatom transcriptomes listed above, alongside other independent genome and transcriptome data (24, 27). This library was divided into 151 nonredundant categories, based on current phylogenomic understanding of the divisions of the tree of life, grouped into nine eukaryotic lineages (amorphea, alveolates, cryptomonads, dinoflagellates, green algae, haptophytes, red algae, rhizaria, stramenopiles), two prokaryotic lineages (archaea, bacteria), and viruses. A list of the categories considered, and the libraries from which they are constructed, is provided in Dataset S1.

The FASTA sequence of the single top-hits for each query gene against each reference subcategory, with a threshold e-value of 10^{-05}, were extracted and compiled to form a local reference library of up to 151 closest homologs of the query sequence across the tree of life. The query sequence was then searched again by BLASTp against this local reference library, and the best-scoring hits were ranked. Genes that retrieved a best BLASTp hit, excluding ochrophyte orthologs, to a nonphotosynthetic stramenopile category (either oomycetes, labyrinthulomycetes, *Blastocystis*, or bicosoecids) (Dataset S1), were inferred to have vertical origin. Exactly 8,000 genes, which retrieved a best-scoring hit, excluding all ochrophyte orthologs, to a nonstramenopile lineage were retained for subsequent phylogenetic inspection.

Manual Verification of *P. tricornutum* HGTs. Alignments were generated for each retained gene using the local reference library previously assembled, MAFFT v7.271 (74) using the --auto option. These alignments were subsequently refined using the in-built alignment program in GeneIOUS v10.0.9 (75) under highly stringent settings (gap open penalty 12, gap extension penalty 3), and were manually curated to remove poorly aligned sequences (defined as sequences with <50% pairwise coverage with the *P. tricornutum* query). A small number of genes (21) were split into separate N- and C-terminal domain alignments, based on the alignment of nonoverlapping sets of reference sequences to different regions of the *P. tricornutum* query. Alignments for which fewer than 2 nonochrophyte sequences could be aligned to the *P. tricornutum* sequence were rejected, retaining 3,339 curated alignments. Alignments were trimmed with trimAl using the -gt 0.5 option (76), and tree topologies were generated using RAxML v8 with the PROTGAMMAJTT substitution matrix (77) and 100 bootstrap replicates.

Consensus tree topologies were rooted on either: 1) The nearest nonphotosynthetic stramenopile outgroup to the *P. tricornutum* query, 2) if no such homolog were detected the node containing the greatest phylogenetic diversity of nonochrophyte lineages (following methodology established in ref. 5), or 3) if neither criteria was met, any node that recapitulated the basal divergence of chrysistan ochrophytes (raphidophytes, PX clade, eustigmatophytes, pinguiophytes, or SC clade) to ochrophytes from Khakista and Hypogyristea (pelagophytes, dictyochophytes, bolidophytes, and diatoms). The monophyletic or paraphyletic group of ochrophyte sequences containing the *P. tricornutum* query was then manually read to identify possible horizontal gene transfers. First, the tree was read to determine the position of aplastidic stramenopiles (e.g., are oomycetes sister to ochrophytes?); followed by the identification of paraphyletic relationships (e.g., do ochrophytes resolve within a clade containing multiple, successive sister-groups from the same lineage?); and finally, and only in the absence of either condition above, by considering which single lineage shows greatest phylogenetic proximity to the ochrophytes.

The direction of each horizontal transfer event was defined as either being "in," if the lineage formed a sister-group to ochrophytes; "out," if the lineage was contained within a paraphyletic group of ochrophytes; or "unknown," if neither relationship could be exclusively attributed. In the case of unknown transfers, the BLAST top-hits of the query was consulted to classify the most likely direction of the transfer, dependent on whether the sequences involved were more distantly related (transfer in) or more closely related (transfer out) to the *P. tricornutum* query than the sequences of the ochrophyte ingroup. Sister-groups containing two or more lineages were either assigned a most probable evolutionary origin, based on the most parsimonious common ancestor of the sister-group, or if this were not possible (e.g., a sister-group consisting of two separate, monophyletic lineages), each constituent lineage within the sister-group was annotated with a weighted probability (totalling 1) that it was the originator of the ochrophyte sequences.

Each transfer event was assigned a probable time point, based on the identity of the ochrophyte sequences separating the donor or recipient lineage from *P. tricornutum*. These were: *Phaeodactylum* (transfers involving *P. tricornutum* only); pennate diatoms (involving *P. tricornutum* and at least one other pennate diatom); polar centric/pennate diatoms (*P. tricornutum* and at least one polar centric diatom); basal diatoms (*P. tricornutum* and at least one radial centric diatom); Khakista and Hypogyristea (*P. tricornutum* and at least one of bolidophytes, pelagophytes, or dictyochophytes); and the ochrophyte ancestor (*P. tricornutum*, and at least one of the PX clade, raphidophytes, eustigmatophytes, pinguiophytes, or the SC clade).

Specific associations between individual nonochrophyte and ochrophyte subcategories were identified by internal BLASTp searches between all pairs of species retained in each curated alignment. The best-matching searches of each nonochrophyte sequence against all ochrophyte sequences, and of each ochrophyte sequence against each nonochrophyte sequence, calculated using the -max_target_seqs 1 option and a threshold e-value of 10^{-05} in each case.

An exemplar series of tree topologies are provided in *SI Appendix*, Fig. S1, and full assignations are tabulated in Dataset S1. All BLAST outputs, alignments, trees, and internal BLAST search outputs are publically accessible from the osf.io library https://osf.io/t52wk/, in the folder "All HGTs into *P. tricornutum* genome." Bootstrap replicate files are provided in this folder for trees resolved with the JTT matrix (as annotated above), along with tree reconciliations with LG and WAG matrices, for user exploration. Modified versions of Dataset S1, retaining only HGTs supported by >50% and >80% bootstrap supported, are provided within the same osf.io library, in the folder "Supporting Datasets."

Automatic Annotation of all Bacterial HGTs in Ochrophytes. Bacterial HGTs were detected using an automated pipeline (*SI Appendix*, Fig. S3) across 162 ochrophyte genomes and transcriptomes (Dataset S3). Transcriptome libraries studied were decontaminated using a previously defined pipeline, comparing the BLAST similarity scores between pairs of transcriptome libraries as described elsewhere (5, 70, 78).

Each ochrophyte protein dataset was clustered at a 90% similarity cutoff with CD-HIT (79) to reduce sequence redundancy. The longest protein sequence of each cluster was retained, leaving a total of 2,677,637 nonredundant ochrophyte proteins. These proteins were searched against a modified version of NR GenBank using BLASTP (e-value < 10^{-05}), excluding all ochrophyte sequences and separately clustering bacterial, archaeal, viral, and the remaining eukaryotic sequences at 90% similarity with CD-HIT, and the best 5 overall hits, best 30 hits among nonochrophyte eukaryotes, and the best 30 hits among noneukaryotes were recorded.

A total of 41,620 ochrophyte proteins were found to have prokaryotic, but no nonochrophyte eukaryotic hits, and were considered HGT candidates de facto. A further 308,833 proteins, which were found to possess at least 1 noneukaryotic hit within the 5 best overall hits were aligned against their 30 best prokaryotic and eukaryotic homologs using MAFFT 7.271 (74) using the --auto option. All columns containing gaps in the ochrophyte sequence were removed, and homologs were removed if their aligned sequence (after the removal of gapped columns) contained fewer than 100 aligned amino acids and their length was shorter than 80% the length of the ochrophyte sequence. Phylogenies were built for 302,717 curated protein alignments, containing >100-aa aligned ochrophyte sequence, using FastTree under default parameters (80). Trees were automatically rooted on the sequence that has the longest phylogenetic distance to the focal ochrophyte sequence.

A total of 44,694 ochrophyte proteins for which the first 2 sister-groups in the phylogenetic tree contained noneukaryotes of the same nature (bacteria, archaea or viruses), along with the 41,620 ochrophyte proteins with no clear eukaryotic homologs, were organized into protein families using ORTHOMCL, returning 8,646 families and 21,795 singleton proteins. Each of the proteins in each ORTHOMCL family, plus the nonredundant top 30 eukaryotic hits and top 30 noneukaryotic hits as defined above were merged, and passed through a second round of the same automated phylogenetic pipeline, using IQTree, which implements ModelFinder (using the setting -m MFP) to determine the best-fit substitution model for each protein (81–83). Cleaning was performed using the longest ochrophyte sequence in the family as reference. Automated parsing of the resulting trees identified 2,653 multispecies HGTs (i.e., sequences originating from the same HGT event found in multiple ochrophytes) and 7,426 single-species HGTs (i.e., sequences originating from an HGT event found in a single ochrophyte). All alignments and trees are publically accessible from the osf.io library https://osf.io/t52wk/, in the folder "Quantification of bacterial HGTs."

Trees for exemplar HGT clusters (*SI Appendix*, Fig. S7 and Dataset S6) were realized with MrBayes 3.2.6 hosted on the CIPRES Science Gateway web portal, using three substitution matrices (GTR, Jones, and WAG) (84, 85). For each dataset, two runs were performed, each comprising four chains for 600,000 Markhov Chain Monte Carlo generations, sampling at every 5,000 points, with the first 50% used as burn-in for the consensus. Analyses were carried out using RAxML v8.1.17 (77), using three substitution matrices (GTR, JTT, and WAG), with the Γ correction and automatic bootstrapping.

Identification of Composite Bacterial/Nonbacterial Genes in HGT Data. Bacterial HGTs identified in the above dataset, with at least one eukaryotic detected homolog, and >70% bootstrap support for the placement of ochrophytes and bacteria, were searched for composite domain origin based on a previously published pipeline (5, 32) using CompositeSearch (86). Each protein sequence from each bacterial HGT group was added to a previously published library of bacterial, archaeal, viral, plasmid, and unicellular eukaryote genomes (32), and used in an all-versus-all BLASTp search with an expect value of 1×10^{-05} and a minimum identity of 30%. Sequences were grouped into families if they overlapped by more than 80% of their lengths. CompositeSearch was used to identify potential composite families and their components, and composite families including sequences from a bacterial HGT group were investigated further. The identity of each domain in a composite gene was assessed by profiling the consensus origin (either bacteria, archaea, or ochrophytes) of the top 10 BLAST hits identified between the ochrophyte protein domain and their components in the library (Dataset S5). All BLAST search outputs are publically accessible from the osf.io library https://osf.io/t52wk/, in the folder "Identification of chimeric sequences."

Construction of a Multigene Reference Tree. The *P. tricornutum* genome (17) was searched by BLASTp against a uniref library (71) from which all lineages that currently or may have historically possessed plastids of secondary or higher endosymbiotic derivation (i.e., ochrophytes, cryptomonads, dinoflagellates, haptophytes, euglenids, and plastid-bearing rhizarians and apicomplexans) had been removed (5), and in each case the ID and expect value of the top BLAST hit were extracted. Similar BLAST searches were performed using the *P. tricornutum* query protein sequences against the 162 ochrophyte genomes and transcriptomes inspected for bacterial HGTs (Dataset S3), plus a further 42 reference eukaryotic genomes (Dataset S4). Manual alignments were constructed for 65 genes, for which the top-hit in >90% of the ochrophyte libraries inspected had an expect value that was smaller than the expect value associated with the uniref top-hit sequence, using MAFFT (74) and GeneIOUS (75), as above.

Guide single-gene trees for each trimmed cluster were inferred using 100 bootstrap replicates of the neighbor-joining tree program built into GeneIOUS (75). Ochrophyte sequences that failed to resolve with other ochrophyte sequences were removed from each cluster, and two clusters which retained fewer than 80% occupancy of ochrophyte species were rejected (Dataset S1). Concatenated multigene sequences were built for each species using the 63 remaining clusters, and trimmed with trimAl using the -gt 0.8 option to remove all residues that corresponded to gaps in >20% of the sequences (76). Complete multispecies alignments are provided in Dataset S4.

Bayesian analyses were performed using MrBayes 3.2.6 using two substitution matrices (GTR and WAG), and RAxML v 8 using three matrices (GTR, JTT, and WAG) hosted on the CIPRES Science Gateway web portal (84, 85), as above. Long-branch removal was performed by calculating the averaged branch length of each taxon in the concatenated tree, and serially removing the 10 longest diatom, nondiatom ochrophyte, or nonochrophyte branches from the alignment (87). Site rate evolution categories were calculated using TIGER with the -b 100 option applied (88), and the eight fastest-evolving categories (corresponding to the fastest evolving 35% sites) were serially removed from the alignment. Long-branch removal, fast-site removal, and clade-deduction trees were inferred using RAxML, with the GTR substitution matrix and Γ correction, on CIPRES, as detailed above.

EVOLUTION

Nexus format outputs of the MrBayes trees, Newick format outputs of each RAxML tree, and support values obtained for the modified data and a comparison between the topologies obtained and previously published ochrophyte multigene trees (9, 33–35, 89–93) are shown in Dataset S4. Raw data, including single-gene and concatenated alignments and trees, within the Open Science Framework (OSF) library, https://osf.io/t52wk/, in the folder "Multigene trees."

ALE Inference of Bacterial HGT Data. ALE analysis (37, 38), which estimates maximum-likelihood rates of gene speciation, duplication, transfer and loss, was performed on the 675 bacterial HGT protein clusters that contained sequences in at least 10 ochrophyte species. The phylogeny of each protein cluster was reconstructed using PhyloBayes (94) with an LG model (95) and a discrete γ-distribution with four categories, while making sure to keep only the sequences attributed to ochrophytes. PhyloBayes was run for 4,000 generations, sampling at every generation after an initial burn-in of 1,000 generations to obtain a distribution of gene trees, which can be directly used by ALE. ALE was informed that each ochrophyte sequence library was incomplete, with the number of BUSCOs retrieved using eukaryote references provided as a prior of library completeness (36).

For each cluster, ALE generated a set of 100 reconciliations, output the average number of speciations, duplications, interspecific transfers, and losses of each HGT, and indicated likely origins on the species tree. To identify genes that were vertically inherited by ochrophytes following the bacterial HGT origin, the inferred number of events was compared to null expectations under independent genes-species evolutions (96), generated by shuffling the species across the ochrophyte tree 100 times. Protein clusters were considered to have been vertically inherited if the observed number of speciations was higher than 95% of the null expectations and if the number of transfers was lower than 95% of the null expectations. A total of 435 protein clusters met this criterion.

HGT losses across ochrophytes were estimated by comparing the number of detected protein clusters in each species to the expected number, considering missing BUSCO sequences. A negative deviation (if the observed number is smaller than the expected number) likely indicates an excess of losses of HGTs. All randomization, subsampling and reconciliation controls are made publicly accessible within the osf library https://osf.io/t52wk/, in the folder "ALE inference of bacterial HGTs."

Functional Characterization of HGTs. Genes inferred to have either been horizontally transferred into or out of ochrophyte genomes through manual phylogenetic annotation of the *P. tricornutum* genome (Dataset S1), and *P. tricornutum* sequences identified in automated annotations of ochrophyte-wide HGTs (Dataset S3), were compared to functional annotations (biological processes, PFAM domains) assembled for the *P. tricornutum* version 3 genome annotation (17); epigenetic marks and RPKM values for gene expression under replete media assembled from previous studies (50, 51). Separate enrichment and expression values were calculated for *P. tricornutum* genes manually identified to be of bacterial origin, and for *P. tricornutum* genes acquired from a eukaryotic algal group (either cryptomonads, haptophytes, dinoflagellates, red algae, green algae, photosynthetic alveolates, or photosynthetic rhizaria).

Targeting predictions were obtained for all HGTs using ASAFind v2.0 in conjunction with Signal P v3.0 (41, 49) MitoFates, with threshold prediction value 0.35 (5, 39); HECTAR (40, 97); and WolfPSort v 1.1, taking the consensus prediction using animal, plant and fungi reference models (42).

Gene coregulation values were calculated for *P. tricornutum* genes in published diatom microarray metastudy data (53), and a novel dataset consisting of combat-normalized RNA-seq data from three previously published studies of *P. tricornutum* gene expression under N, P, and Fe stress (26, 52, 54, 55). Coregulation values were calculated for bacterial HGTs identified by automated analysis, and attributed from MRCA estimations to have been transferred into *P. tricornutum* only, or an earlier ancestor of pennate diatoms, pennate and polar centric diatoms, basally divergent diatoms, Khakista and Hypogyristea, or all ochrophytes.

Targeting predictions, and gene coregulation outputs, are publically accessible within the Open Science Framework (OSF) library, https://osf.io/t52wk/, in the folders "Localization of HGTs" and "Co-regulation of HGTs."

GFP Localization of Exemplar Bacterial HGTs. Ten bacterial HGTs identified in the *P. tricornutum* genome (*SI Appendix*, Fig. S4 and Dataset S6) were experimentally localized using full-length C-terminal GFP expression constructs. The full-length cDNA construct of the gene, amplified from total cellular RNA of *P. tricornutum* strain 1.86 grown under replete enhanced seawater medium (ESAW) without silicate, in a 19 °C light:dark cycle as defined above, was cloned into a pPhat: extrabright GFP vector using HiFi assembly Mix (New England Biolabs), following the manufacturers' instructions. Primers used for the GFP ligation are provided in Dataset S6. The complete GFP fusion vector was transformed into 5×10^7 late exponential phase *P. tricornutum* 1.86 cells, plated on replete 1% agar ESAW plates without silicate, using biolistic transformation, as previously described (98, 99). Transformant colonies were selected on agar ESAW plates supplemented with 100 μg mL^{-1} zeocin (New England Biolabs) and verified by Western blotting using anti-GFP antibodies (Agrisera).

Expressing lines were visualized using a Leica SP5 (ER-tracked lines) or SP8 (brightfield lines) confocal microscope (Imaging Facility, Institut de Biologie de l'Ecole Normale Supérieure) using previously defined spectra (5, 24). Where appropriate, cell lines were stained with either 0.1 μg mL^{-1} MitoTracker Orange, or 0.25 μg mL^{-1} ER-Tracker Blue-White (Thermo-Fisher), suspended in marine PBS (mPBS) for 25 min at room temperature, prior to being washed in unstained mPBS and visualized per the manufacturer's instructions. Unstained lines, and wild-type cells, were used as controls to verify the emission and detection windows used to visualize GFP and stain patterns (*SI Appendix*, Fig. S23*B*).

Supporting Data. Transcriptome libraries sequenced within this study are available in the ENA under the project link https://www.ebi.ac.uk/ena/data/view/PRJEB34158). All supporting information, including alignments and single-gene trees of manually and automatically annotated HGTs, randomization, and subsampling controls for mathematical reconciliation of HGT frequencies, and in silico predictions, transcriptional abundance calculations, and raw microscopy data, are publically accessible through the Open Science Framework (osf) portal: https://osf.io/t52wk/. Each folder within the OSF portal contains a "readme" file detailing its contents; and a global inventory is provided on the portal wiki page.

Data Availability. Osf library data have been deposited in the Diatom HGT paper dataset (10.17605/OSF.IO/T52WK) (100). Data pertaining to the construction and inspection of the stramenopile multigene tree are provided through the University of Cambridge Apollo Repository https://www.repository.cam.ac.uk/handle/1810/261420 (101). Sequencing data are archived in the ENA, https://www.ebi.ac.uk/ena/data/view/PRJEB34158 (accession nos. ERR3497271–ERR3497276).

ACKNOWLEDGMENTS. We thank Fuhai Liu, Leïla Tirichine, and Catherine Cantrel (Institut de Biologie de l'École Normale Supérieure) for assistance with the maintenance of diatom cultures; Guita Niang and Arnaud Meng (Genoscope) for assistance with sequence assembly and data management; Vincent Daubin and Bastien Boussau for helpful discussions about the amalgamated likelihood estimation approach; Marianne Jaubert and Angela Falciatore (Université Pierre et Marie Curie) for assistance with biolistic transformations; and Eric Bapteste (Université Pierre et Marie Curie) for guidance in the identification of chimeric bacterial/nonbacterial proteins, respectively. R.G.D. was supported by a European Molecular Biology Organization Early Career Fellowship (ALTF 1124/2014) and a CNRS Momentum Fellowship (awarded 2019–2021). C.B. acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (Diatomic; Grant 835067), Agence Nationale de Recherches (ANR) Grant ANR-19-CE20-0020 (BrownCut), and the French Government Investissements d'Avenir programme OCEANOMICS (ANR-11- BTBR-0008). C.B. and H.M. acknowledge funding from MEMO LIFE (ANR-10- LABX-54) and Paris Sciences and Letters Research University (ANR-11- IDEX-0001-02). H.M. and B.P.-L. additionally acknowledge a PhD fellowship from the Ecole Normale Supérieure (to B.P.-L.). A.V. and G.B. acknowledge funding from the Excellence Initiative of Aix-Marseille University, A*MIDEX, a French "Investissements d'Avenir" program (ANR-11-IDEX-0001-02) and the MICROBIO-E program.

1. G. Walker, R. G. Dorrell, A. Schlacht, J. B. Dacks, Eukaryotic systematics: A user's guide for cell biologists and parasitologists. *Parasitol* **138**, 1638–1663 (2011).
2. S. M. Adl *et al*., The revised classification of eukaryotes. *J. Eukaryot. Microbiol.* **59**, 429–493 (2012).
3. E. Evanovich, J. Guerreiro, A timescale for the radiation of photosynthetic eukaryote. bioRXiv:10.1101/2020.04.18.047969 (20 April 2020).
4. L. Eme, S. C. Sharpe, M. W. Brown, A. J. Roger, On the age of eukaryotes: Evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* **6**, a016139 (2014).
5. R. G. Dorrell *et al*., Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *eLife* **6**, 23717 (2017).
6. J. W. Stiller *et al*., The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat. Commun.* **5**, 5764 (2014).

7. L. Eme, E. Gentekaki, B. Curtis, J. M. Archibald, A. J. Roger, Lateral gene transfer in the adaptation of the anaerobic parasite *Blastocystis* to the gut. *Curr. Biol.* **27**, 807–820 (2017).
8. X. Fan *et al*., Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions. *Sci. Adv.* **6**, eaba0111 (2020).
9. R. Derelle, P. López-García, H. Timpano, D. Moreira, A phylogenomic framework to study the diversity and evolution of stramenopiles (=Heterokonts). *Mol. Biol. Evol.* **33**, 2890–2898 (2016).
10. C. de Vargas *et al*.; Tara Oceans Coordinators, Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
11. P. Škaloud *et al*., Speciation in protists: Spatial and ecological divergence processes cause rapid species diversification in a freshwater chrysophyte. *Mol. Ecol.* **28**, 1084–1095 (2019).
12. N. Joli *et al*., Need for focus on microbial species following ice melt and changing freshwater regimes in a Janus Arctic Gateway. *Sci. Rep.* **8**, 9405 (2018).
13. R. G. Dorrell, A. G. Smith, Do red and green make brown? Perspectives on plastid acquisitions within chromalveolates. *Eukaryot. Cell* **10**, 856–868 (2011).
14. C. Bowler *et al*., The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244 (2008).
15. E. V. Armbrust *et al*., The genome of the diatom *Thalassiosira pseudonana:* Ecology, evolution, and metabolism. *Science* **306**, 79–86 (2004).
16. A. Moustafa *et al*., Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**, 1724–1726 (2009).
17. A. Rastogi *et al*., Integrative analysis of large scale transcriptome data draws a comprehensive landscape of *Phaeodactylum tricornutum* genome and evolutionary origin of diatoms. *Sci. Rep.* **8**, 4834 (2018).
18. P. Deschamps, D. Moreira, Reevaluating the green contribution to diatom genomes. *Genome Biol. Evol.* **4**, 683–688 (2012).
19. S. R. Smith *et al*., Evolution and regulation of nitrogen flux through compartmentalized metabolic networks in a marine diatom. *Nat. Commun.* **10**, 4552 (2019).
20. A. E. Allen *et al*., Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**, 203–207 (2011).
21. E. Kazamia *et al*., Endocytosis-mediated siderophore uptake as a strategy for Fe acquisition in diatoms. *Sci. Adv.* **4**, eaar4536 (2018).
22. P. J. Keeling *et al*., The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).
23. One Thousand Plant Transcriptomes Initiative, One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
24. R. G. Dorrell *et al*., Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 6914–6923 (2019).
25. T. Ševčíková *et al*., Updating algal evolutionary relationships through plastid genome sequencing: Did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Sci. Rep.* **5**, 10134 (2015).
26. M. H. Cruz de Carvalho, H. X. Sun, C. Bowler, N. H. Chua, Noncoding and coding transcriptome responses of a marine diatom to phosphate fluctuations. *New Phytol.* **210**, 497–510 (2016).
27. D. Beisser *et al*., Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. *PeerJ* **5**, e2832 (2017).
28. F. Gómez, K. Furuya, S. Takeda, Distribution of the cyanobacterium Richelia intracellularis as an epiphyte of the diatom *Chaetoceros compressus* in the western Pacific Ocean. *J. Plankton Res.* **27**, 323–330 (2005).
29. J. M. Cock *et al*., The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617–621 (2010).
30. N. Delaroque, W. Boland, The genome of the brown alga *Ectocarpus siliculosus* contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol. Biol.* **8**, 110 (2008).
31. Y. P. Galachyants *et al*., Sequencing of the complete genome of an araphid pennate diatom *Synedra* acus subsp. radians from Lake Baikal. *Dokl. Biochem. Biophys.* **461**, 84–88 (2015).
32. R. Méheust, E. Zelzion, D. Bhattacharya, P. Lopez, E. Bapteste, Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 3579–3584 (2016).
33. M. B. Parks, N. J. Wickett, A. J. Alverson, Signal, uncertainty, and conflict in phylogenomic data for a diverse lineage of microbial eukaryotes (diatoms, bacillariophyta). *Mol. Biol. Evol.* **35**, 80–93 (2018).
34. M. Yu *et al*., "Evolution of the plastid genomes in diatoms" in *Plastid Genome Evolution*, S. Chaw, R. Jansen, Eds. (Elsevier:, 2018), vol. 85, pp. 129–155.
35. F. Burki *et al*., Untangling the early diversification of eukaryotes: A phylogenomic study of the evolutionary origins of Ccentrohelida, Haptophyta and Cryptista. *Proc. Biol. Sci.* **2016**, 283 (1823).
36. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
37. G. J. Szöllősi, A. A. Davín, E. Tannier, V. Daubin, B. Boussau, Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140335 (2015).
38. G. J. Szöllõsi, W. Rosikiewicz, B. Boussau, E. Tannier, V. Daubin, Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).
39. Y. Fukasawa *et al*., MitoFates: Improved prediction of mitochondrial targeting sequences and their cleavage sites. *Mol. Cell. Proteomics* **14**, 1113–1126 (2015).
40. B. Gschloessl, Y. Guermeur, J. M. Cock, HECTAR: A method to predict subcellular targeting in heterokonts. *BMC Bioinformatics* **9**, 393 (2008).
41. A. Gruber, G. Rocap, P. G. Kroth, E. V. Armbrust, T. Mock, Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant J.* **81**, 519–528 (2015).
42. P. Horton *et al*., WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* **35**, W585-7 (2007).
43. R. H. Lampe *et al*., Different iron storage strategies among bloom-forming diatoms. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E12275–E12284 (2018).
44. G. Emiliani, M. Fondi, R. Fani, S. Gribaldo, A horizontal gene transfer at the origin of phenylpropanoid metabolism: A key adaptation of plants to land. *Biol. Direct* **4**, 7 (2009).
45. N. Khaldi, J. Collemare, M. H. Lebrun, K. H. Wolfe, Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome Biol.* **9**, R18 (2008).
46. J. Levering *et al*., Genome-scale model reveals metabolic basis of biomass partitioning in a model diatom. *PLoS One* **11**, e0155038 (2016).
47. J. Prihoda *et al*., Chloroplast-mitochondria cross-talk in diatoms. *J. Exp. Bot.* **63**, 1543–1557 (2012).
48. T. Nonoyama *et al*., Metabolic innovations underpinning the origin and diversification of the diatom chloroplast. *Biomolecules* **9**, 322 (2019).
49. J. D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
50. A. Veluchamy *et al*., An integrative analysis of post-translational histone modifications in the marine diatom *Phaeodactylum tricornutum. Genome Biol.* **16**, 102 (2015).
51. A. Veluchamy *et al*., Insights into the role of DNA methylation in diatoms by genome-wide profiling in *Phaeodactylum tricornutum*. *Nat. Commun.* **4**, 2091 (2013).
52. O. Ait-Mohamed *et al*., PhaeoNet: A holistic RNAseq-based portrait of transcriptional coordination in the model diatom *Phaeodactylum tricornutum*. *Front Plant Sci* **11**, 590949 (2020).
53. J. Ashworth, S. Turkarslan, M. Harris, M. V. Orellana, N. S. Baliga, Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. *Mar. Genomics* **26**, 21–28 (2016).
54. J. K. McCarthy *et al*., Nitrate reductase knockout uncouples nitrate transport from nitrate assimilation and drives repartitioning of carbon flux in a model pennate diatom. *Plant Cell* **29**, 2047–2070 (2017).
55. S. R. Smith *et al*., Transcriptional orchestration of the global cellular response of a model pennate diatom to diel light cycling under iron limitation. *PLoS Genet.* **12**, e1006490 (2016).
56. W. F. Martin, Eukaryote lateral gene transfer is Lamarckian. *Nat. Ecol. Evol.* **2**, 754 (2018).
57. S. J. Sibbald, L. Eme, J. M. Archibald, A. J. Roger, Lateral gene transfer mechanisms and pan-genomes in eukaryotes. *Trends Parasitol.* **36**, 927–941 (2020).
58. E. Vancaester, T. Depuydt, C. M. Osuna-Cruz, K. Vandepoele, Comprehensive and functional analysis of horizontal gene transfer events in diatoms. *Mol. Biol. Evol.* **37**, 3243–3257 (2020).
59. L. Guarente, T. M. Roberts, M. Ptashne, A technique for expressing eukaryotic genes in bacteria. *Science* **209**, 1428–1430 (1980).
60. M. K. Yang *et al*., Proteogenomic analysis and global discovery of posttranslational modifications in prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E5633–E5642 (2014).
61. H. Qiu, J. M. Lee, H. S. Yoon, D. Bhattacharya, Hypothesis: Gene-rich plastid genomes in red algae may be an outcome of nuclear genome reduction. *J. Phycol.* **53**, 715–719 (2017).
62. E. Hehenberger, R. J. Gast, P. J. Keeling, A kleptoplastidic dinoflagellate and the tipping point between transient and fully integrated plastid endosymbiosis. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 17934–17942 (2019).
63. S. Sato *et al*., Genome-enabled phylogenetic and functional reconstruction of an araphid pennate diatom *Plagiostriata* sp. CCMP470, previously assigned as a radial centric diatom, and its bacterial commensal. *Sci. Rep.* **10**, 9449 (2020).
64. K. C. Haines, R. R. L. Guillard, Growth of vitamin B12-requiring marine diatoms in mixed laboratory cultures with vitamin B12-producing marine bacteria. *J. Phycol.* **10**, 245–252 (1974).
65. I. Kaczmarska, J. M. Ehrman, S. S. Bates, C. Léger, J. Harris, Diversity and distribution of epibiotic bacteria on Pseudo-nitzschia multiseries (Bacillariophyceae) in culture, and comparison with those on diatoms in native seawater. *Harmful Algae* **4**, 725–741 (2005).
66. H. Biebl *et al*., Description of *Labrenzia alexandrii* gen. nov., sp. nov., a novel alphaproteobacterium containing bacteriochlorophyll a, and a proposal for reclassification of *Stappia aggregata* as *Labrenzia aggregata* comb. nov., of *Stappia marina* as *Labrenzia marina* comb. nov. and of *Stappia alba* as *Labrenzia alba* comb. nov., and emended descriptions of the genera *Pannonibacter, Stappia* and *Roseibium*, and of the species *Roseibium denhamense* and *Roseibium hamelinense*. *Int. J. Syst. Evol. Microbiol.* **57**, 1095–1107 (2007).
67. S. Malviya *et al*., Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1516–E1525 (2016).
68. P. Tréguer *et al*., Influence of diatom diversity on the ocean biological carbon pump. *Nat. Geosci.* **11**, 27–37 (2018).
69. F. M. Ibarbalz *et al*.; Tara Oceans Coordinators, Global trends in marine plankton diversity across kingdoms of life. *Cell* **179**, 1084–1097.e21 (2019).
70. E. Kopylova, L. Noé, H. Touzet, SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
71. B. J. Haas *et al*., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
72. B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, C. H. Wu, UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

73. E. J. Carpenter *et al*., Access to RNA-sequencing data from 1,173 plant species: The 1000 plant transcriptomes initiative (1KP). *Gigascience* **8**, giz126 (2019).
74. K. Katoh, J. Rozewicki, K. D. Yamada, MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* (2017).
75. M. Kearse *et al*., Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
76. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
77. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
78. A. O. Marron *et al*., The evolution of silicon transport in eukaryotes. *Mol. Biol. Evol.* **33**, 3226–3248 (2016).
79. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
80. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
81. L. Li, C. J. Stoeckert, Jr, D. S. Roos, M. C. L. Ortho, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
82. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
83. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
84. M. A. Miller *et al*., A RESTful API for access to phylogenetic tools via the CIPRES Science gateway. *Evol. Bioinform. Online* **11**, 43–48 (2015).
85. F. Ronquist *et al*., MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
86. J. S. Pathmanathan, P. Lopez, F. J. Lapointe, E. Bapteste, CompositeSearch: A generalized network approach for composite gene families detection. *Mol. Biol. Evol.* **35**, 252–255 (2018).
87. V. Hampl *et al*., Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3859–3864 (2009).
88. C. A. Cummins, J. O. McInerney, A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Syst. Biol.* **60**, 833–844 (2011).
89. E. C. Yang *et al*., Supermatrix data highlight the phylogenetic relationships of photosynthetic stramenopiles. *Protist* **163**, 217–231 (2012).
90. U. Sorhannus, M. G. Fox, Phylogenetic analyses of a combined data set suggest that the *Attheya* lineage is the closest living relative of the pennate diatoms (Bacillariophyceae). *Protist* **163**, 252–262 (2012).
91. I. Riisberg *et al*., Seven gene phylogeny of heterokonts. *Protist* **160**, 191–204 (2009).
92. E. C. Theriot, M. P. Ashworth, T. Nakov, E. Ruck, R. K. Jansen, Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. *Mol. Phylogenet. Evol.* **89**, 28–36 (2015).
93. E. C. Theriot, M. Ashworth, E. Ruck, T. Nakov, R. K. Jansen, A preliminary multigene phylogeny of the diatoms (Bacillariophyta): Challenges for future research. *Plant Ecol. Evol.* **143**, 278–296 (2010).
94. N. Lartillot, H. Philippe, A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
95. S. Q. Le, O. Gascuel, An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
96. M. Groussin *et al*., Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun.* **8**, 14319 (2017).
97. E. Afgan *et al*., The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3–W10 (2016).
98. A. Falciatore, R. Casotti, C. Leblanc, C. Abrescia, C. Bowler, Transformation of non-selectable reporter genes in marine diatoms. *Mar. Biotechnol. (NY)* **1**, 239–251 (1999).
99. M. Siaut *et al*., Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene* **406**, 23–35 (2007).
100. R. Dorrell *et al*., Diatom HGT paper dataset. Open Science Framework. https://doi.org/10.17605/OSF.IO/T52WK. Deposited 19 December 2020.
101. R. Dorrell, G. de Kerdrel, S. Smith, A. Allen, C. Bowler, Stramenopile multigene tree dataset. https://doi.org/10.17863/CAM.6608. Deposited 4 December 2016.