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

To cite this version:

HAL Id: mnhn-03969507
https://mnhn.hal.science/mnhn-03969507
Submitted on 2 Feb 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes

Richard G. Dorrell1,3, Adrien Villain2,2, Benoît Perez-Lamarque4,2, Guillemette Audren de Kerdrel1,4, Giselle McCallum3, Andrew K. Watson2, Ouardia Ait-Mohamed5, Adriana Alberti1,4, Eriwan Corré6, Kyle R. Frischkor6,9, Juan J. Pierella Karlusich4, Eric Pelletier9,9, Hélène Morlon9, Chris Bowler9,9, and Guillaume Blanch4,1

*Institut de Biologie de l’École Normale Supérieure, Ecole Normale Supérieure, CNRS, INSERM, Université Paris Sciences et Lettres, 75005 Paris, France; 1Aix Marseille University, Université de Toulon, CNRS, Institut de Recherche pour le Développement (IRD), Mediterranean Institute of Oceanography (MIOM) UM 110, 13288 Marseille, France; 2Institut de Systématique, Evolution, Biodiversité (ISYE), Muséum National d’Histoire Naturelle, CNRS, Sorbonne Université, École Pratique des Hautes Études (EPHE), Université des Antilles (UA), 75005 Paris, France; 3Department of Biology, Concordia University, H3G 1M8 QC, Montreal, H3G 1M8 QC, Canada; 4Metabolic Genomics, Genoscope, Institut de Biologie François Jacob, Commissariat à l’Énergie Atomique, CNRS, Université Evry, Université Paris-Saclay, 91000 Evry, France; 5FRR2 CNRS, Analysis and Bioinformatics for Marine Science, Station Biologique de Roscoff, Université Pierre et Marie Curie Paris 06, 75005 Paris, France; and 6Research 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 “crow supergroups,” some of which (e.g., opisthokonts, including animals and fungi) are exclusively nonphotosynthetic, while others (e.g., archaeoplastids, 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 archaeoplastids (green algae and plants, red algae, and glaucophytes), which has subsequently been transmitted to 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-phvtoplankton of no more than a few microns in 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 genes of bacterial origin (14, 16, 17) and between 100 and >1,500 genes of bacterial HGTs predominantly impact the diatom secreted proteinome, which may contribute to the striking evolutionary success of diatoms.

Significance

The ochrophytes are an ancient and important group of eukaryotic algae, including diatoms, the most important photosynthesizers 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.


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

Published under the PNAS license.

1To whom correspondence may be addressed. Email: dorrell@bio.ens.psl.eu or guillaume.blanch@miou.ups-tlse.fr.

2Present address: BIOASTER, 69007 Lyon, France.

3Present address: Centre d’Études Prospectives et d’Informations Internationales, 75007 Paris, France.

4Present address: Immunity and Cancer Department, Institut Curie, Paris Sciences and Letters Research University, INSERM U932, 75005 Paris, France.

5Present 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 genome 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 angiosperm-ochrophyte radiation, 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. L4 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 nonstromatolite 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 phylogenetics 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. L4 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. L4 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 Hypogysteida (pelagophytes and dicytochromophytes) (Fig. L8 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 ($\chi^2$ enrichment $P < 1E-05$) (SI Appendix, Figs. S4 and S5 and Dataset S2). In contrast, this signal was evenly distributed across all haptophyte groups, including the early-diverging pavlophytes (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. S3B) and internal BLAST searches of each alignments (SI Appendix, Fig. S4) also revealed an association between pelagophytes and dictyochophytes with all dinoflagellate 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 synchronochromytes, synurophytes, and chrysophytes (SI Appendix, Fig. S3C). 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).

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. 1B). 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. L4 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. L4 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. L4 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 assignment (Dataset S1). Notably, Phatr3_1Jdraf52.p1 shows highest BLAST similarity with sequences from the nitrogen-fixing diatom endobiont *Richelia intracellularis* (27); and Phatr3_3J13257.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...
we compared the set of diatom sequence libraries, of which 7 transcriptomes were generated from 162 ochrophyte genomes and transcriptomes, including 109 phylogenetic histories compatible with prokaryotic HGTs based elsewhere in the ochrophyte tree. We identified genes that had bacterial (as opposed to archaeal, or viral) lineages (for six exemplar clusters in SI Appendix, Fig. S6). We present consensus MrBayes and RAxML topologies from 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 Planctomyceae-Verrucomicrobiaeae-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 Ricketsiales, Holosporales, Legionellales, or Chlamydiaceae branches (SI Appendix, Fig. S8). A limited number of genes (397) of viral origin were also predicted, of which the greatest (139, of predominant Phycoviridae origin) were restricted to the brown alga Ectocarpus, consistent with previous reports of integrated viral segments in this genome (29, 30) along with 495 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 Synechocystis sp. (31), and the diatom transcriptomes for Thalassiosira miniscula, Chaetoceros sp. UNCC1202, Chaetoceros sp. neogaeaceae, and Chaetoceros sp. curvisetum, and the transcriptome of the dictyochoyte Pseudopediaella 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 versus conserved nucleotides 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 composition distributions 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, Estuobaccellus 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. 1B), 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-genome 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 Atheyea 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 pinguioiophytes, 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 Hypogystea, and of pennate and polar centric diatoms, and radial centric diatoms within the Coscinodiscophyceae (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), 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. 3D). This ALE-verified dataset
contained 435 HGT protein clusters that originated up to 10 times more frequently within diatoms compared to other ochrophyte lineages \( (t \text{ test}, t_m = 32) = 3.4, P = 0.0008 \) (Fig. 3B, SI Appendix, Fig. S19B, and Dataset S4). We confirmed using randomizations that this diatom enrichment in bacterial HGTs was not due to random chance (Fig. 3C). We also found that many of the bacterial HGT proteins have been subsequently lost in diatoms (SI Appendix, Fig. S19C).

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 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. S1A). 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 menaquione (Phatr3_J47199.p1,Phatr3_J48624.p1, demethylmenaquinone methyltransferase, Phatr3_J4670.p1, 1 4-dihydroxy-2-naphthoate polyprenyltransferase) biosynthesis, reflecting the probable role of HGT for diversifying eukaryotic secondary metabolism (44, 45).

Futher pathways exceptionally annotated as being behaviorally 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 cyclo-uariclasticase, Phatr3_J24238.p1, mitochondrial carbamate kinase, and Phatr3_J3969.p1, chloroplast acetylglutamate kinase), alongside other derived from within-eukaryotic HGT (e.g., Phatr3_J29702.p1, mitochondrial urease, acquired from dinoflagellates; and Phatr3_J50577.p1, J52221.p1, chloroplast acetylornithine transaminase, acquired from haptophytes or green algae; SI Appendix, Fig. S1H), or with presumed vertical origin. These data underline 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 [as 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. 4A 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-acetylersterase 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. 4B and SI Appendix, Fig. S22). Furthermore, Phatr3_J50595.p1, a homolog of the bacterial CoA-binding protein YccU conserved across all ochrophytes (SI Appendix, Fig. S7C), as well as Phatr3_J8596.p1, a protein of unknown function shared across all ochrophytes except for chrysophytes, synechocysts, and pioniuqioptes localized to the plasma membrane (Fig. 4B).

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 phylum the enrichment was only detected in bacterial HGTs acquired at early time points in ochrophyte (Fig. 4A). 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 genomic information is available. Using RNAseq, 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. S24A). 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 \( (P < 0.05, \chi^2) \) 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. S24B and Dataset S1). However, structural inspection and 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. S24B).

Considering epigenetic marks associated with *P. tricornutum* 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; \( \chi^2 \) test, *P* < 0.01), and had higher mean reads per million kilobases.
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 \textit{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 metatranscriptomes of microarray (53) and RNA-seq (RNA-seq) data (52) from \textit{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 \textit{P. tricornutum} genome show distinct patterns of coregulation, falling within different modules of coredigated 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 ockrophyte 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 ockrophytes throughout their evolutionary history that are detectable in the genome of the model diatom \textit{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 ockrophyte 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 ockrophytes. 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 \textit{P. tricornutum}.

We note the specific importance of bacterial HGTs in the recent evolution of diatoms. Considering the diversity of HGTs detectable in \textit{P. tricornutum}, we show that HGT from bacteria has occurred continuously throughout ockrophyte 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 ockrophytes into bacteria, which may reflect transcriptional or
post-translational features that impede their functional expression in bacteria (59, 60). Considering the frequencies of bacterial HGTs identifiable across multispecies ophryophage data, we further 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 ophryophage groups (Fig. 3 and SI Appendix, Figs. S4 and S18–S21). Ultimately, our data point to a continuous bombardment of ophryophage genomes with bacterial DNA, which may have occurred more substantially in diatoms than other ophryophytes. 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 sequence data, for example from Diatom sp. RCC3090, Guinardia flaccida sp. RCC3092, Pleurosigma RCC1991, and Synedra sp. RCC2677, Tara Oceans, will be key to understanding how bacterial HGT underpins the versatility of 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 the various gene models, we project significantly higher frequencies of HGT losses in diatoms than in nondiatom ophryophytes (Fig. 3 and SI Appendix, Fig. S19C).

We additionally notice specific biases in the functions attributed to bacterial HGTs. While early bacterial HGTs in ophryophage genomes are enriched in chloroplast-targeting predictions, suggesting origins associated with the ophryophage 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 ophryophage 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 ophryophytes.

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, Pleuroneisigma sp. RCC3090, Shionodiscus biculatus RCC1991, and Synedra sp. RCC2510 (Datasets S1 and S3). These species were selected based on their high abundances in Tara Oceans metacode 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
Continuous secretory enrichment of bacterial HGTs in ochrophytes.

Cellular localizations and bacterial origins:

- Ochrophyte
- Diatoms to Coscinodiscales
- Diatoms to Probosciales
- Raphid pennate diatoms
- Polar centric diatoms
- Polar/ pennate diatom ancestor
- Skeletonema/ Thalassiosirales
- Khakista
- Ditylum ancestor
- Phatr3_J8596.p1-GFP
- Phatr3_J42871.p1-GFP

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 (proteins.ensembl.org/Phaeodactylum_tricornutum/Infelix), was searched using BLASTp against a complete library consisting of a complete version of uniprot (accessed June 2018) (72), decontaminated versions of the MMTESP (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 (amorpheza, 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 10⁻⁵⁰, were extracted and compiled to form a local reference library of up to 151 closest homologs from 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 GeneOUPS 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 (defining sequences with fewer than 50% pairwise identity across the coverage 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 PROTGAMMAJSON 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 lineage of the chloroplast-containing ochrophytes (raphidophytes, *Phaeodactylum*, eustigmatophytes, pinguiochlorates, or SC clade) to ochrophytes from Khakista and Hypogyristea, was inferred as root. Exactly 8,000 genes, which retrieved a best-scoring hit, excluding all ochrophyte orthologs, to a nonstramenopile lineage were retained for subsequent phylogenetic inspection.

All HGTs 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 (proteins.ensembl.org/Phaeodactylum_tricornutum/Infelix), was searched using BLASTp against a complete library consisting of a complete version of uniprot (accessed June 2018) (72), decontaminated versions of the MMTESP (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 (amorpheza, 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 10⁻⁵⁰, were extracted and compiled to form a local reference library of up to 151 closest homologs from 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 as root. 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 GeneOUPS 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 (defining sequences with fewer than 50% pairwise identity across the coverage 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 PROTGAMMAJSON 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 lineage of the chloroplast-containing ochrophytes (raphidophytes, *Phaeodactylum*, eustigmatophytes, pinguiochlorates, or SC clade) to ochrophytes from Khakista and Hypogyristea, was inferred as root. Exactly 8,000 genes, which retrieved a best-scoring hit, excluding all ochrophyte orthologs, to a nonstramenopile lineage were retained for subsequent phylogenetic inspection.
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 (totaling 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 Hypopyrgites (P. tricornutum and at least one of bolidophytes, pelagophytes, or dictyochophytes); and the ochrophyte ancestor (P. tricornutum, and the remaining eukaryotic sequences at 90% similarity with CD-HIT, and the 10 longest diatom, nondiatom ochrophyte, or nonochrophyte branches as described elsewhere (5, 70, 78).

An exemplar series of tree topologies are provided in SI Appendix, Fig. S1, and full assignments 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 (as annotated above), along with tree renderings and LG and WAG matrices for each HGT group was added to a previously published library of bacterial, archaean, viral, plasmid, and uncellular eukaryote genomes (32), and used in an all-versus-all BLASTp search with an expect value of $1 \times 10^{-10}$ and a minimum identity of 30%. Sequences were grouped into families if they overlapped by more than 80% of their lengths. ComposenetSearch 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 a composite gene was assessed by profiling the consensus origin (either bacteria, archaean, 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 Multigenic 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 plasmds 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, for each of the top 100 BLAST hits identified. Further sequence alignments were performed using the top 10 hits of each ochrophyte lineage (Dataset S4). Manual alignments were constructed for 65 genomes, 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, bootstrapped in MAFFT (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 multigenic 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 SI Appendix, Dataset A4.

Dorrell et al.
Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes
PNAS | 9 of 12
https://doi.org/10.1073/pnas.2009974118

Downloaded from https://www.pnas.org by Univ Pierre et Marie Curie on February 2, 2023 from IP address 134.157.183.98.
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/5zwvk/, 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 reconstructions, output the average number of speciations, duplications, interspecific transfers, and losses, 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 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 9% 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 re- and raw microscopic data, are publicly accessible through the Open Science Framework (osf) portal: https://osf.io/5zwvk/. 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) (101). 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. All supporting information, including alignments and single-gene trees of manually and automatically annotated HGTs, randomization, and subsampling controls for mathematical recombination of HGT frequencies, and in silico predictions, transcriptional abundance calculations, and raw microscopic data, are publically accessible through the Open Science Framework (osf) portal: https://osf.io/5zwvk/.

Acknowledgments. We thank Fuhai Liu, Leila Tirichine, and Catherine Cantrel (Institut de Biologie de l’Ecole Normale Supérieure) for assistance with the construction of diatom P. tricornutum gene coregulation values were calculated for P. tricornutum identified to be of bacterial origin, and for P. tricornutum transferred into or out of ochrophyte genomes through manual annotation (17); epigenetic marks and RPKM values for gene expression under N, P, and Fe stress (26, 27, 54, 55). Coregulation values were calculated for bacterial HGTs identified in published diatom microarray metatext 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, and P stress (26, 52, 54, 55). Coregulation values were calculated for bacterial HGTs identified by automated analysis, and attributed from MRCA estimates 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 Hypogryriota, or all ochrophytes. Targeting predictions and gene coregulation outputs, are publicly accessible within the Open Science Framework (OSF) library, https://osf.io/5zwvk/, 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. 44 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 × 10−8 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 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 line) 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 nuclei (SI Appendix, Fig. S23).

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 recombination of HGT frequencies, and in silico predictions, transcriptional abundance calculations, and raw microscopic data, are publically accessible through the Open Science Framework (osf) portal: https://osf.io/5zwvk/.

Data Availability. Osf library data have been deposited in the Diatom HGT paper dataset (10.17605/OSF.IO/T52WK) (101). 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 Fudan Liu, Leila Tirichine, and Catherine Cantrel (Institut de Biologie de l’Ecole Normale Supérieure) for assistance with the construction of the diatom P. tricornutum gene expression under N, P, and Fe stress (26, 27, 54, 55). Coregulation values were calculated for bacterial HGTs identified by automated analysis, and attributed from MRCA estimates 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 Hypogryriota, or all ochrophytes. Targeting predictions and gene coregulation outputs, are publicly accessible within the Open Science Framework (OSF) library, https://osf.io/5zwvk/, in the folders “Localization of HGTs” and “Co-regulation of HGTs.”

5. R. G. Dorrell et al., Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. eLife 6, e23717 (2017).

PNAS | 10 of 12 | doi:10.1073/pnas.2009974118

Dorrell et al.
Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes.
Phylogenomic fingerprinting of tempo and functions of genomic gene transfer within opohytes