Continuous gene flow contributes to low global species abundance and distribution of a marine model diatom

Achal Rastogi[1], Anne-Flore Deton-Cabanillas[1], Fabio Rocha Jimenez Vieira[1], Alaguraj Veluchamy[1,$], Catherine Cantrel[1], Gaohong Wang[2], Pieter Vanormelingen[3], Chris Bowler[1], Gwenael Piganeau[4,5], Leila Tirichine[1*] and Hanhua Hu[2]

[1]IBENS, Département de Biologie, Ecole Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005, Paris, France

[$]Present Address: Biological and Environmental Sciences and Engineering Division, Center for Desert Agriculture, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

[2]Key Laboratory of Algal Biology, Institute of Hydrobiology, Donghu south road, Wuchang district, Wuhan, Hubei Province, China

[3]Ghent University, Department of Biology, Research Group Protistology and Aquatic Ecology Krijgslaan 281/S8 9000 Gent, Belgium

[4]CNRS, UMR 7232, Observatoire Océanologique

[5]UPMC University Paris 06, Observatoire Océanologique, Sorbonne Universités

*Corresponding author: tirichin@biologie.ens.fr

Keywords

Heterozygosity, Population genetics, Admixture, Phytoplankton, Diatoms, *Phaeodactylum tricornutum*

Abstract

Unlike terrestrial ecosystems where geographical isolation often leads to a restricted gene flow between species, genetic admixing in aquatic micro-eukaryotes is likely to be frequent. Diatoms inhabit marine ecosystems since the Mesozoic period and presently constitute one of the major primary producers in the world's ocean. They are a highly diversified group of eukaryotic phytoplankton with estimates of up to 200,000 species. Since decades, *Phaeodactylum tricornutum* is used as a model diatom species to characterize the functional pathways, physiology and evolution of diatoms in general. In the current study, using whole genome sequencing of ten *P. tricornutum* strains, sampled at broad geospatial and temporal scales, we show a continuous dispersal and genetic admixing between geographically isolated strains. We also describe a very high level of heterozygosity and propose it to be a consequence of frequent ancestral admixture. Our finding that *P. tricornutum* sequences are plausibly detectable at low but broadly distributed levels in the world's ocean further suggests that high admixing between geographically isolated strains may create a significant bottleneck, thus influencing their global abundance and distribution in nature. Finally, in an attempt to understand the functional implications of genetic diversity between different *P. tricornutum* ecotypes, we show the effects of domestication in inducing changes in the selection pressure on many genes and metabolic pathways. We propose these findings to have significant implications for understanding the genetic structure of diatom populations in nature and provide a framework to assess the genomic underpinnings of their ecological success.

Introduction

Diatoms are unicellular predominantly diploid and obligate photosynthetic eukaryotes. They belong to a large group of heterokonts, constituents of the chromalveolate [or SAR (Stramenopila, Alveolate, Rhizaria)] group, which are believed to have evolved from serial endosymbiosis involving green and red algal symbionts (Bowler et al. 2008; Moustafa et al. 2009; Dorrell et al. 2017). Diatoms were first discovered by Ehrenberg in the 19th century in dust samples collected by Charles Darwin in the Azores. According to the earliest fossil records, they are believed to be in existence since at least 190 million years (Armbrust 2009) and their closest sister group are the Bolidomonads.

Diatoms are a highly diversified group of eukaryotic phytoplankton (Armbrust 2009), and exist in a wide range of shapes and sizes (Tirichine et al. 2017). Multiple ecological factors, including geographical isolation and competitive displacement, are proposed to account for the current global diversity of diatoms (Rabosky and Sorhannus 2009). These studies have further improved our understanding of the abundance and mosaic distribution of diatom species in the world's ocean, supporting the idea of their continuous and unrestricted dispersal (Finlay 2002; Cermeno and Falkowski 2009), and suggesting that geographical isolation may not have a significant impact on gene-flow. However, our understanding of the mechanisms and evolutionary forces that stabilized/regulated this continuous mixing within a resident population, generating genetic and phenotypic diversity, is limited.

A sexual stage is considered obligatory in most diatom species (Chepurnov et al. 2004). However, they predominantly reproduce asexually and only a few species (Davidovich and Bates 1998; Chepurnov et al. 2002; Mouget et al. 2009; Davidovich et al. 2012; Godhe et al. 2014) have actually been observed as having a sexual stage in their life cycle. *Phaeodactylum tricornutum* is a non-abundant coastal diatom species found under highly unstable environments like estuaries, rock-pools, etc. and has never been reported to undergo sexual reproduction. However, factors like small cell size, discontinuous sexual phases in diatoms,

and the observation that their sexual reproduction is sensitive to many nonspecific abiotic components (Mouget et al. 2009; Godhe et al. 2014), limit our ability to constrain the sexual cycle of these organisms. Despite its low abundance in the open ocean (Malviya et al. 2016), *P. tricornutum* is extensively used as a model diatom to characterize their metabolism (Bowler et al. 2008; Allen et al. 2011; Huysman et al. 2013; Morrissey et al. 2015; Tanaka et al. 2015; Fortunato et al. 2016), and to understand their evolution (Bowler et al. 2008). *P. tricornutum* is among the few diatom species with a whole genome sequence available to the community (Tirichine et al. 2017), and the only diatom for which state-of-the-art functional and molecular tools have been developed over the past few decades (Falciatore et al. 1999; Siaut et al. 2007; De Riso et al. 2009; Maheswari et al. 2009; Huysman et al. 2010; Maheswari et al. 2010; Veluchamy et al. 2013; Kaur and Spillane 2015; Veluchamy et al. 2015; Diner et al. 2016; Nymark et al. 2016; Rastogi et al. 2016). These resources have advanced *P. tricornutum* as a model diatom species and provided a firm platform for future genome-wide structural and functional studies.

Since the discovery of *P. tricornutum* by Bohlin in 1897 and the characterization of different morphologies or morphotypes, denoted fusiform, triradiate, oval, round and cruciform , 10 strains from 9 different geographic locations (sea shores, estuaries, rock pools, tidal creeks, etc.) around the world, from sub-polar to tropical latitudes, have been isolated [well described in (De Martino 2007)]. These ecotypes have been collected within the time frame of approximately one century, from 1908 (Plymouth strain, Pt2/3) to 2000 (Daylan strain, Pt10) (De Martino 2007). All the strains have been maintained either axenically or with native bacterial populations in different stock centers and have been cryopreserved after isolation. Previous studies have reported distinct functional behaviors of different ecotypes as adaptive responses to various environmental cues (Stanley 2007; Bailleul et al. 2010; Abida et al. 2015; Taddei et al. 2016) but very little is known about their genetic diversity. Based on sequence similarity of the ITS2 region within the 28S rDNA repeat sequence, the ecotypes can be divided

into four genotypes (Genotype A: Pt1, Pt2, Pt3 and Pt9; Genotype B: Pt4; Genotype C: Pt5 and Pt10; Genotype D: Pt6, Pt7 and Pt8), with genotypes B and C being the most distant (De Martino 2007). These genotype clusters neither appear to correlate with their geographic sampling locations, nor with morphotype characteristics, or with the sampling year (Fig S1). This indicates low genetic distance between ecotypes that are geographically isolated, which is maintained across a long time scale and that trait diversity observed within *P. tricornutum* populations is largely independent of genotype.

The accumulated effect of diverse evolutionary and ecological forces such as recombination, mutation, selection, drift and admixture has been found to dictate the structure and diversity of genomes in a wide range of species (Liti et al. 2009; Cao et al. 2011; Flowers et al. 2015). Such studies within diatoms are rare and estimates of genetic diversity within diatom populations are mostly inferred using microsatellite-based genotyping approaches (Harnstrom et al. 2011; Whittaker and Rynearson 2017). Although these techniques have revealed a wealth of information about diatom evolution, dispersal and reproductive physiology, additional insights can likely be obtained using state-of-the-art whole genome comparative analysis techniques. Deciphering the standing genomic variation of *P. tricornutum* across different ecotype populations is an important first step to assess the role of various evolutionary forces in regulating the adaptive capacities of diatoms in general (e.g.(Matuszewski et al. 2015)).

In order to understand the underlying genomic diversity within different ecotypes of *P. tricornutum* and to establish the functional implications of such diversity, we performed deep whole genome sequencing of the 10 most studied ecotypes. Using reference-based population genomics approaches, we present a genome-wide diversity map and the population genetics structure of geographically isolated ecotypes, revealing the impact of continuous admixing on the evolution of diatom populations. Further, while deciphering multiple haplotypes at the whole genome level using compensatory base changes (CBC)

analysis we show that these haplotypes are not reproductively isolated. Like in any model organism, this work further provides the community with whole genome sequences of the ten most studied ecotypes, which will be a valuable genetic resource for functional studies of ecotype-specific ecological traits in the future.

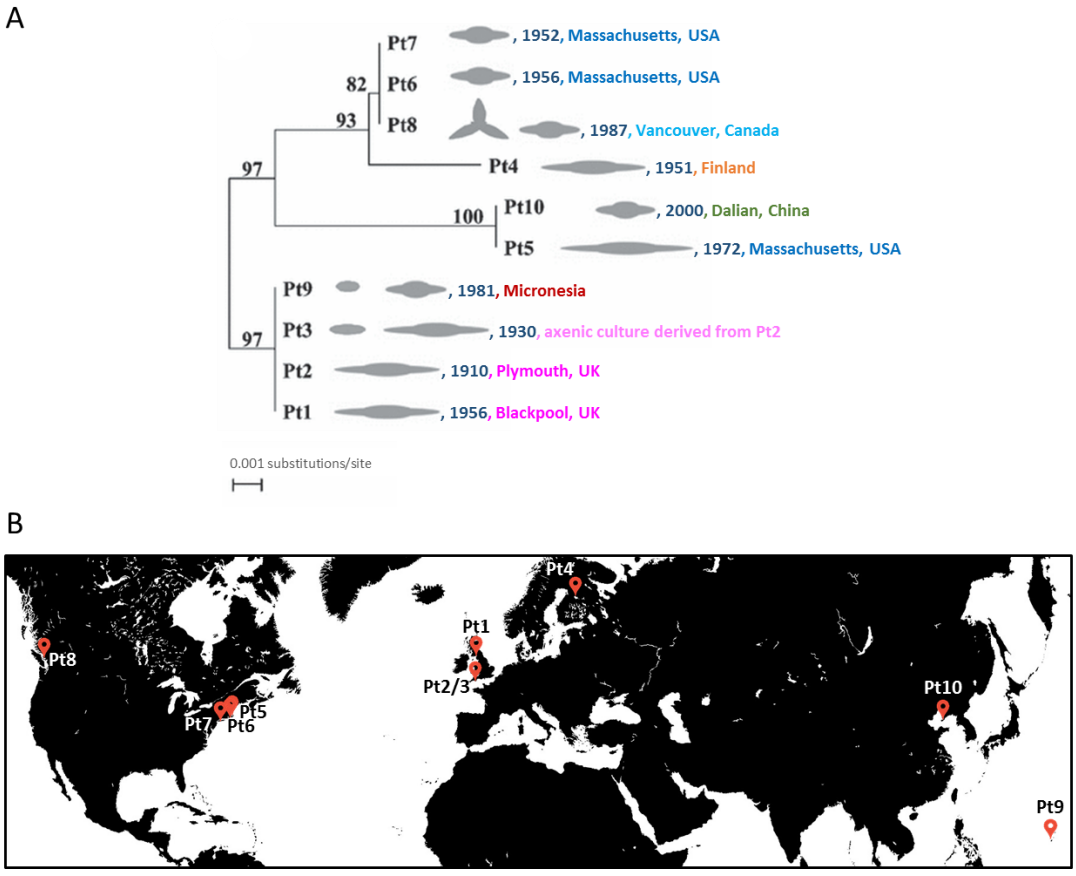


Figure S1. Origins of *P. tricornutum* ecotypes used in this study. (A) ITS2 tree derived from Martino et al. 2007 showing the dominant morphology, geographic location and year of sampling. (B) Ecotype map represents geographical sampling locations of all the *P. tricornutum* ecotypes used in the study.

Results

Heterozygous alleles account for most of the genetic diversity within *P. tricornutum* ecotypes

We sequenced the genomes of 10 isolates of *P. tricornutum* and performed a reference-based assembly using the genome sequence of the reference strain Pt1 8.6 (Bowler et al. 2008).

Across all ecotypes, the alignment depth ranged from 26X to 162X covering 92% to 98% of the genome (Table 1). The percentage of sequence reads mapped on the reference genome ranged between ~65% to ~80% (Table 1), which is independent of the size of the sequence library, as the latter does not correlate with genome coverage (Table 1). Because genome coverage is high, a portion of unmapped reads is likely a consequence of the incomplete reference assembly, which contains several gaps (Bowler et al. 2008). Also, many regions on the reference genome that are observed as being unmapped by reads from individual ecotypes are annotated as being rich in transposable elements (TEs) (Fig S2). Moreover, across all the ecotypes, the repeated proportion of unmapped reads varies between ~38% (Pt1) to 75% (Pt4), with >90% similarity. Further, using a normalized measure of read depth (see Materials and Methods), we found that 259 and 590 genes, representing ~2% and ~5% of the total gene content, have been lost or exhibit copy number variation (CNV), respectively, across the 10 ecotypes with respect to the reference Pt1 8.6 (Fig 1A, Fig S3A) (File S1). Further, 21 randomly chosen loci were validated by PCR for their loss from certain ecotypes compared to the reference strain Pt1 8.6 (Fig S4).

Approximately 70% of the genes that were either lost within ecotypes or present in many copies are shared among multiple ecotypes (Fig 1A). In addition, we discovered 207 transposable elements (TEs) (~6% of the total annotated TEs) (File S2) to exhibit CNV across one or more ecotypes (Fig S3B). More precisely, 80% of all TEs showing CNVs are shared among two or more ecotypes, with Pt10 possessing the maximum number of ecotype-specific TEs exhibiting CNV (Fig S3B). Not surprisingly, across all the ecotypes, class I-type TEs, which undergo transposition via a copy-and-paste mechanism, show more variation in the estimated number of copies than class II-type TEs, which are transposed by a cut-and-paste mechanism (Fig S3C, S3D). Therefore, in light of the highly repetitive nature of unmapped reads and large structural variations within each ecotype genome, a major proportion of reads that remain unmapped are likely indicative of gene duplication events, transpositions and/or

translocations within individual ecotype genomes compared to the reference genome. This can be further explained with our observations from the Pt10 ecotype, whose WGS reads were only able to map ~92% of the reference genome (Table 1). As expected, large structural variations like CNV and gene loss analysis revealed the maximum number of genes being lost and maximum number of specific genes showing high copy numbers in Pt10 compared to the other ecotype genomes (Fig 1A).

Overall, each ecotype can be characterized by specific genetic features (~0.3% to ~28% ecotype-specific CNVs, Fig1A), possibly linked to the explicit functional behavior of some ecotypes in response to various environmental cues, as reported previously (Stanley 2007; Bailleul et al. 2010; Abida et al. 2015). Biological processes can only be traced, on average, for 40% of the genes exhibiting ecotype-specific CNVs. Among all the enriched biological processes (chi-square test, $P<0.01$) that are associated to genes exhibiting ecotype-specific CNVs (File S3), a gene associated to nitrate assimilation (Phatr3_EG02286) is observed to have higher copy number in Pt4. Nitrate assimilation was shown to be regulated extensively under low light or dark conditions to overcome nitrate limitation of growth in *Thalassiosira weissflogii* (Clark et al. 2002).

Next, we discovered 462,514 (depth >= 4x) single nucleotide variants (SNVs), including ~25% singleton sites, 573 insertions (of length 1 bp to 312 bp) and 1,801 deletions (of length 1 bp to 400 bp) (Fig 1B). Interestingly, we found that most of the SNVs are heterozygous (Fig 1B) and shared between different ecotypes (Fig 1C). The proportion of heterozygous alleles across all the ecotypes varies between ~45% (in Pt5 and Pt10) to ~98% (in Pt1, Pt2 and Pt3). Thus, homozygous SNVs are observed much less frequently (on average once every 8,314 bp) within ecotypes having high levels of heterozygosity (>90%; Pt1, Pt2, Pt3, Pt9) and more frequently (on average 1 out of 319 bp) in ecotypes where heterozygosity was comparatively low (<65%; Pt4, Pt5, Pt6, Pt7, Pt8, Pt10). Similarly, most INDELS (insertions and deletions) are also shared

between different ecotypes, except for Pt4, which possesses the highest proportion of specific INDELS (~75%) and SNVs (~35%) (Fig 1C). With an average transition to transversion ratio of ~1.6, the spectrum of SNVs across all the ecotypes reveals a higher rate of transitions over transversions. In total, compared to the reference allele, six possible types of single nucleotide changes could be distinguished, among which, G:C -> A:T and A:T -> G:C, accounted for more than ~60% of the observed mutations (Fig 1D). Interestingly, this observation is consistent across all the ecotypes, regardless of their striking difference in levels of heterozygosity. This might be mediated by DNA methylation, which is known to cause a high rate of C to T transitions.

Library Name	Origin	Year of Isolation	Mapped Read-Pairs	% Mapped Read-Pairs	Alignment Depth (X)	Genome Coverage (%)
Pt1	Blackpool, UK	1956	3, 642,044	79.41	26.5	98.0
Pt2	Plymouth, UK	Prior to 1910	6, 016,241	78.23	43.8	98.0
Pt3	Plymouth, UK	1930s	6, 373,591	65.62	46.4	98.3
Pt4	Finland	1951	15, 583,665	67.31	113.5	94.0
Pt5	West Dennis, MA, USA	1972	5, 346,009	75.50	38.9	93.2
Pt6	MA, USA	1956	3, 922,830	64.50	28.5	94.1
Pt7	Long Island, NY, USA	1952	4, 937,516	67.30	35.9	94.9
Pt8	Vancouver, Canada	1987	22, 235,170	78.36	162.1	94.4
Pt9	Guam, Micronesia	1981	7, 551,099	74.68	55.2	97.5
Pt10	Dalian, China	2000	5, 436,057	72.59	39.6	92.1

Table 1. Reference-assisted mapping statistics. The table summarizes the origin and year of sampling of each isolate of *P. tricornutum* along with the number of total reads mapped on the reference. Average depth (X=average number of reads aligned on each base covered across the entire genome) was estimated using the number of mapped read pairs and the horizontal coverage (aka. coverage breadth) across the whole genome.

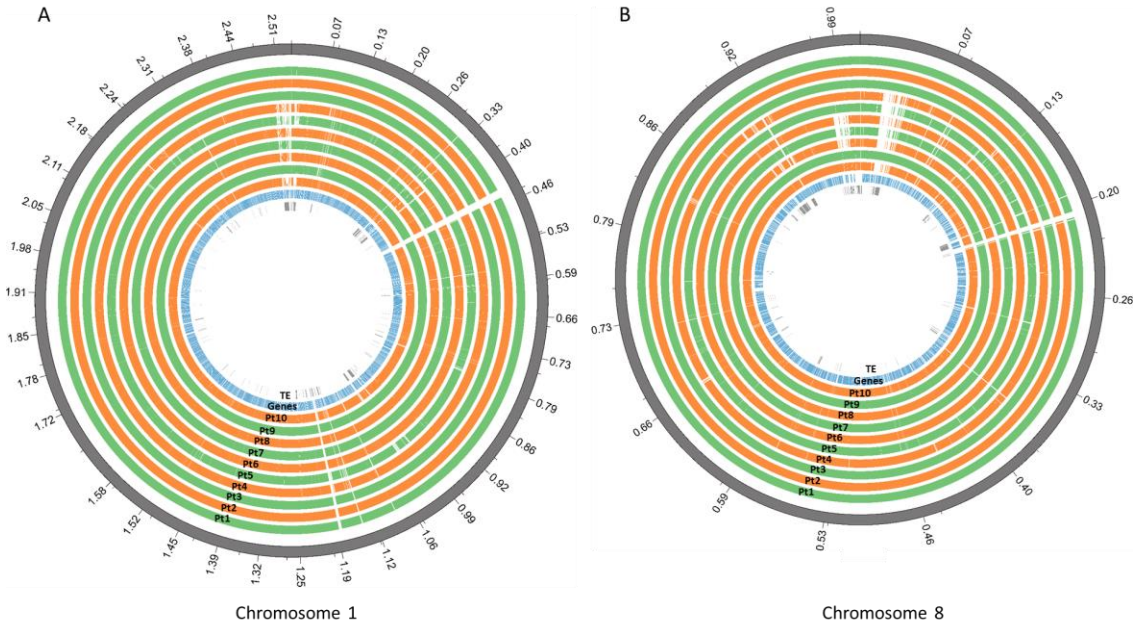


Figure S2. Reference genome coverage. CIRCOS plot showing genomic coverage of chromosomes 1 and 8 by sequence alignment of reads corresponding to individual ecotypes. The outermost circles represent chromosome 1 (A) and chromosome 8 (B) histograms. The two innermost circles represent the genomic regions annotated as genes and transposable elements (TEs), respectively, in the reference genome.

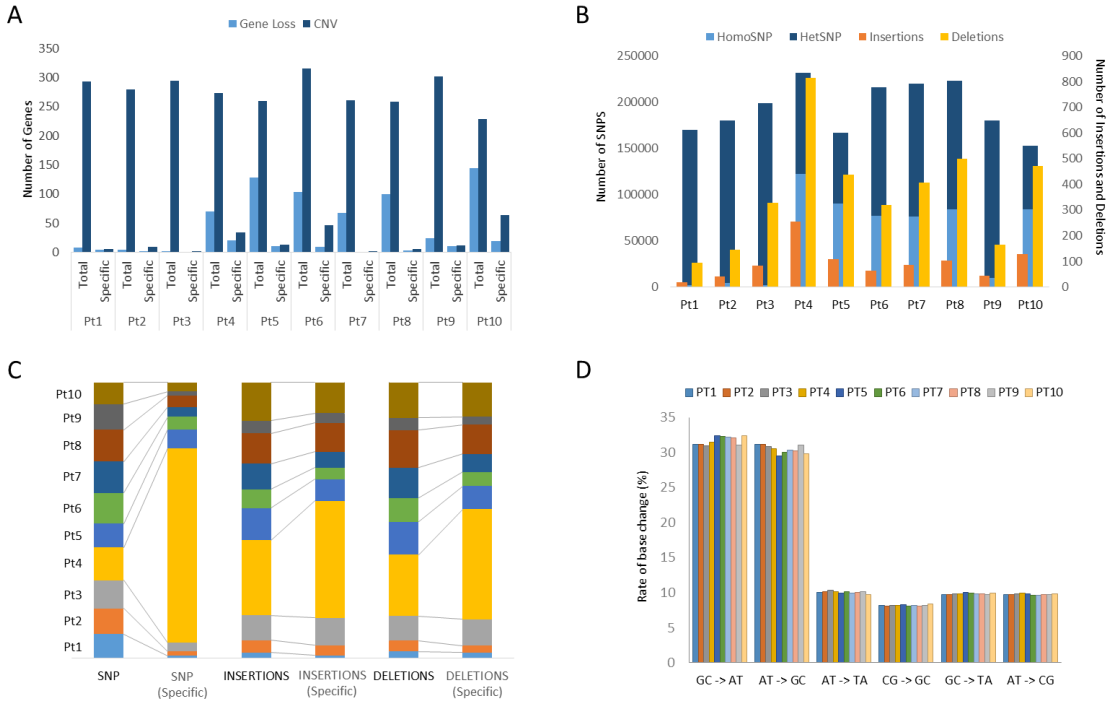


Figure 1. Ecotype genome diversity. (A) The bar plots represent the total and specific numbers of genes, denoted on Y-axis, that exhibit a loss or multiple copies (CNV) within one or more ecotypes. (B) The bar plot represents total number of discovered SNPs, with the proportion of heterozygous SNPs (dark blue) and homozygous SNPs (light blue), INSERTIONS (orange)

and DELETIONS (yellow) in each ecotype compared to the reference genome. (C) The stack bar plot represents the proportion of total vs specific polymorphic variant sites, including SNPs, insertions and deletions (from left to right, respectively) across all the ecotypes. (D) The bar plot represents the mutational spectrum of all the SNPs discovered across the ecotypes. Y-axis denotes the total percentage of individual mutations observed as denoted on X-axis. Colors used in the figure are chosen randomly and have no biological significance.

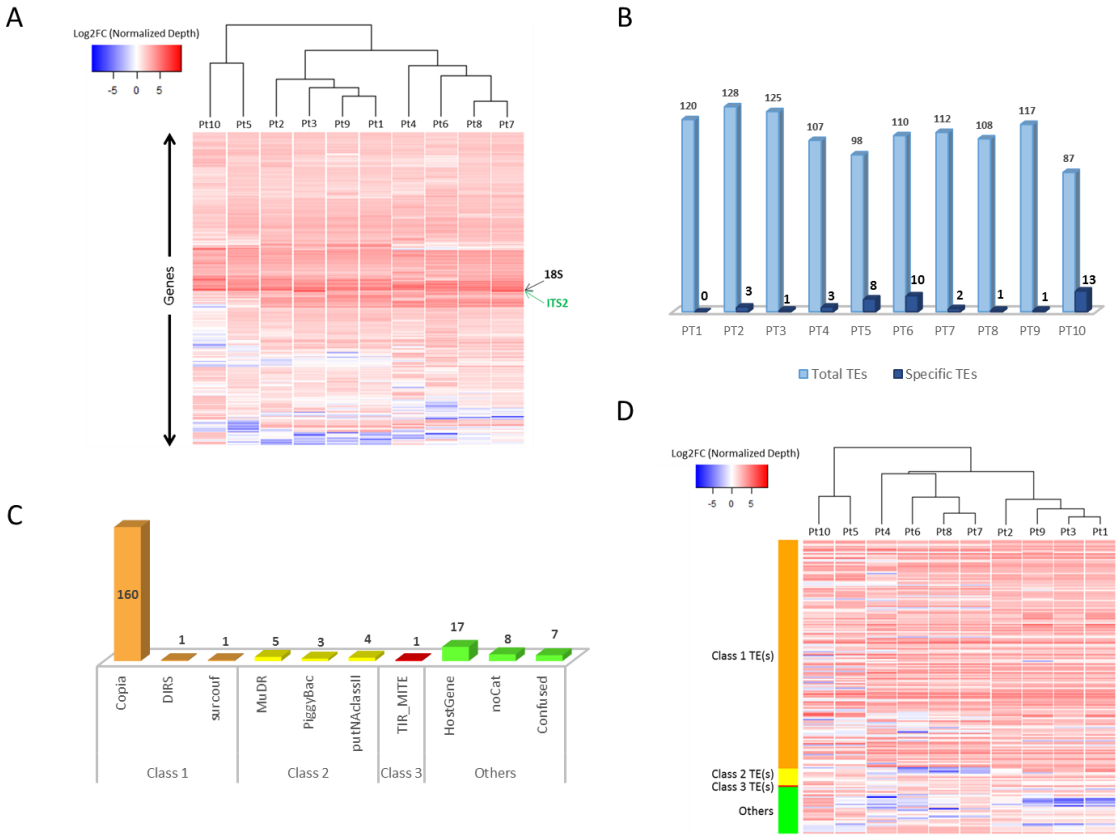


Figure S3. Large structural variations within ecotypes. (A) The heat-map displays the fold-change (FC) of read depth between each reference gene and median of read depth of all the reference genes, within each ecotype. Using Z-score as a measure of normalized read depth, log2 fold change (FC) is calculated as a ratio of Z-score per gene to the average normalized read depth of all the genes per ecotype. Low to high log2FC is represented by a blue to red color gradient in the heat-map. From all the ecotypes only those genes are plotted where log2FC is more than 2 in at least one of the ecotypes and are considered to exhibit copy number variation (CNV). (B) The bar plots represent the number of total- and ecotype-specific transposable elements (TEs) exhibiting CNV across one or more ecotypes. (C) The bar plot represents the absolute numbers of different types of TEs, grouped as Class 1, 2, 3 and others, exhibiting CNV. (D) With similar principle aesthetics as Figure S3A, the heat-map shows the patterns of log2FC only across all the ecotypes of those TEs exhibiting CNV in at least one of the ten ecotypes studied.

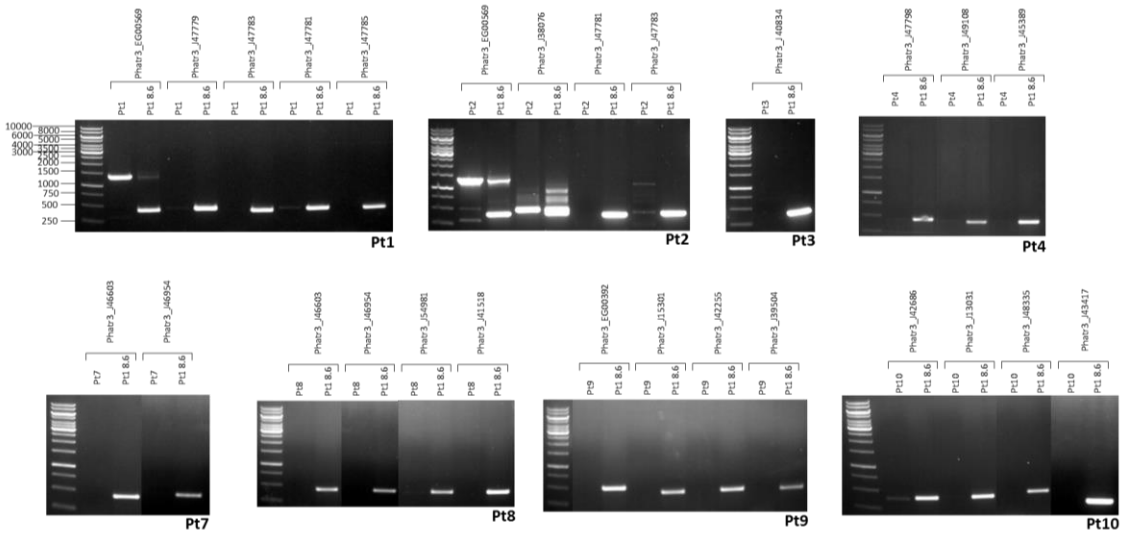


Figure S4. Gene loss validation. PCR validation of 21 candidate genes (denoted by their Phatr3 gene assignment codes) found to be absent in different ecotypes compared to the reference strain Pt1 8.6. The list of primers used in the experiment is provided in Table S1.

Genetic diversity between the ecotypes reveals the presence of four haplogroups

With an exception of Pt4, where we found the maximum number of variant alleles to be ecotype specific, most of the variant alleles within other ecotypes were shared between at least two ecotypes, indicating close genetic relatedness (File S3). In order to cluster the ecotypes based on their genetic distance we therefore estimated the Fixation Index (F_{ST}) as a measure of genetic differentiation among all possible pairs of ecotypes, which ranges between 0.03 and 0.5 and suggests four haplogroups (Fig. 2A). These haplogroups are in broad agreement with 18S gene diversity and previous reports of ecotype diversity based on internal transcribed spacer 2 (ITS2) sequences (De Martino 2007) (Fig S5A and S5B), and were thus denoted haplogroup A (Pt1, Pt2, Pt3 and Pt9), haplogroup B (Pt4), haplogroup C (Pt5 and Pt10), and haplogroup D (Pt6, Pt7 and Pt8) (Fig 2A). Furthermore, the clustering of ecotypes within each haplogroup was consistent at the whole genome scale, as inferred by a phylogenetic tree generated using maximum likelihood algorithm based on all (Fig. 2B) and

only homozygous polymorphic sites (SNVs and INDELS) (Fig. 2B and S5C, respectively), across all the ecotypes.

Geographic isolation often leads to the delineation of isolated populations into different species that can be distinguished based on their ribosomal DNA sequences (Chu et al. 2013). Inspection of the 18S and ITS2 rDNA gene sequences across different haplogroups indicated the presence of multiple variations, including both heterozygous and homozygous variant alleles (Fig S5D and S5E). Because the ribosomal DNA region including 18S and ITS2 is highly repetitive, which is on average ~4 times more than non-ribosomal genes (Fig S2A), these variations can be understood as intra-genomic variations within the genome. However, taxonomists and ecologists use differences within 18S gene sequences as a prominent measure of species assignation and to estimate species delineation (Malviya et al. 2016). This latter practice has been shown to be very conservative as no differences in the 18S gene between reproductively isolated species is expected in species with large effective population sizes (Piganeau et al. 2011). Alternatively, the possibility of sub-populations cannot be ignored. Moreover, there are reports of cryptic speciation within planktonic foraminifers (de Vargas et al. 1999) and coccolithophores (Saez et al. 2003). Therefore, we first considered examining the effect of subpopulations on the 18S gene heterozygosity within the ecotype cultures. We confirmed the expression of all the heterozygous alleles within the 18S rDNA gene using whole genome and total-RNA sequencing of a monoclonal culture isolated from Pt8 population (constituent of haplogroup D), referred to as Pt8Tc (Fig S5D), indicating that the cultures are single population.

To strengthen our understanding of the possibility of speciation, relating observed polymorphisms within 18S ribosomal marker gene to the presence of multiple species, we analyzed the existence of compensatory base changes (CBCs) within secondary structures of the ITS2 gene between all the ecotypes. The presence of CBCs within ITS2 gene has been recently shown to account for reproductive isolation in multiple plant species (Wolf et al.

2013) and between diatom species (Kaczmarska et al. 2014). By comparing the ITS2 secondary structure from all the ecotypes, we did not find any CBCs between any given pair of ecotypes (Fig S6). As a control, we compared the ITS2 secondary structure of all the *P. tricornutum* ecotypes with the ITS2 sequence of other diatom species (*Cyclotella meneghiniana*, *Pseudo-nitzschia delicatissima*, *Pseudo-nitzschia multiseries*, *Fragilariopsis cylindrus*) that have significant degree of evolutionary divergence as depicted previously using multiple molecular marker genes (Medlin 2015; Tirichine et al. 2017), and found multiple CBCs in them (Fig S6). The results thus reject the hypothesis of haplogroups being different species and suggest theoretical sexual compatibility between different geographically isolated ecotypes.

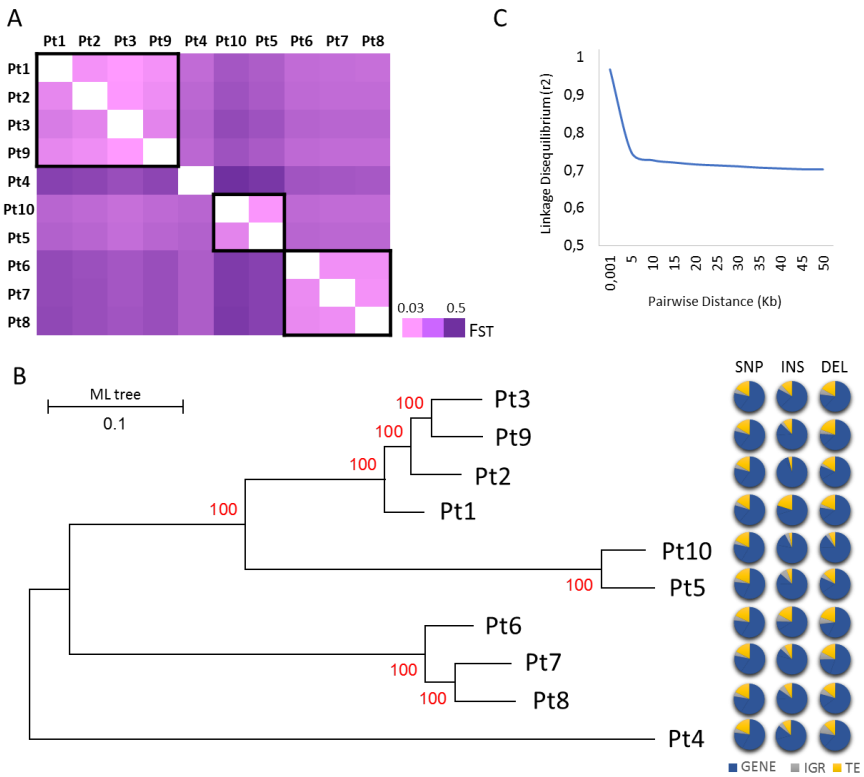
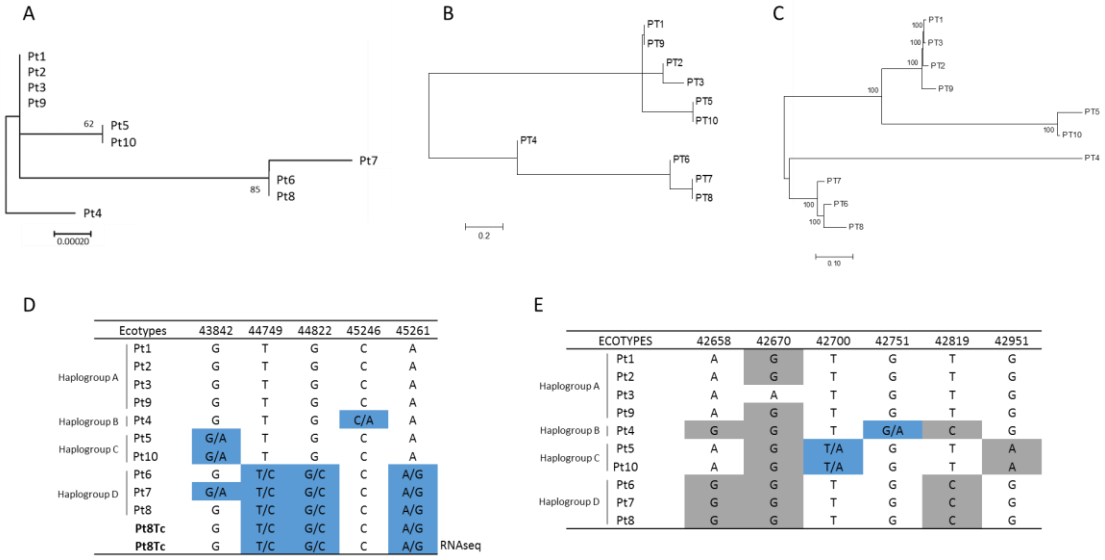


Figure 2. Genetic diversity between the ecotypes. (A) The heat-map measures the genetic differentiation or association between all possible pairs of ecotypes. The colors indicate F_{ST} values, which range from 0.03 to 0.5, with a color gradient from pink to violet, respectively. Values closer to 0 signify high genetic exchange and 1 indicates no exchange between the populations. (B) Phylogenetic association of the ecotypes based on 468,188 genome-wide polymorphic sites (including SNP and INDELS) using a maximum likelihood approach. The numbers on the branches indicate the bootstrap values. Pie charts adjacent to each node of the whole genome tree corresponds to the proportion of SNPs and INDELs over all functional

features of the genome; GENEs (blue), TEs (Transposable Elements, represented in yellow), IGRs (Intergenic Regions, represented in grey). (C) The line plot represents the linkage disequilibrium (LD) decay (r2) across all the haplogroups with pairwise distance between any given pair of homozygous polymorphic alleles.



Ecotypes	43842	44749	44822	45246	45261
Haplogroup A: Pt1	G	T	G	C	A
Haplogroup A: Pt2	G	T	G	C	A
Haplogroup A: Pt3	G	T	G	C	A
Haplogroup A: Pt9	G	T	G	C	A
Haplogroup B: Pt4	G	T	G	C/A	A
Haplogroup C: Pt5	G/A	T	G	C	A
Haplogroup C: Pt10	G/A	T	G	C	A
Haplogroup D: Pt6	G	T/C	G/C	C	A/G
Haplogroup D: Pt7	G/A	T/C	G/C	C	A/G
Haplogroup D: Pt8	G	T/C	G/C	C	A/G
Pt8Tc	G	T/C	G/C	C	A/G
Pt8Tc	G	T/C	G/C	C	A/G

ECOTYPES	42658	42670	42700	42751	42819	42951
Haplogroup A: Pt1	A	G	T	G	T	G
Haplogroup A: Pt2	A	G	T	G	T	G
Haplogroup A: Pt3	A	A	T	G	T	G
Haplogroup A: Pt9	A	G	T	G	T	G
Haplogroup B: Pt4	G	G	T	G/A	C	G
Haplogroup C: Pt5	A	G	T/A	G	T	A
Haplogroup C: Pt10	A	G	T/A	G	T	A
Haplogroup D: Pt6	G	G	T	G	C	G
Haplogroup D: Pt7	G	G	T	G	C	G
Haplogroup D: Pt8	G	G	T	G	C	G

Figure S5. Haplotype analysis. Phylogenetic association of the ecotypes based on (A) 18S and (B) ITS2 sequence alignment across 10 studied ecotypes, using maximum likelihood. (C) Phylogenetic association of the ecotypes based on only homozygous genome-wide SNPs, using maximum likelihood. Numbers on the branches of phylogenetic trees indicate bootstrap values. Multiple sequence alignment of 18S and ITS2 sequences across the ten ecotypes revealed multiple polymorphic positions as indicated in (D) and (E), respectively. Homozygous SNPs are shown in grey, while heterozygous SNPs are shown in blue with all possible alleles.

A

	Cm	Pde	Pm	Fc
Cm		3	4	3
Pde	3		6	5
Pm	4	6		8
Fc	3	5	8	

Centric Diatom / Pennate Diatoms

B

	Fc	Pt1	Pt2	Pt3	Pt4	Pt5	Pt6	Pt7	Pt8	Pt9	Pt10
Fc		1	1	1	3	1	1	1	3	1	1
Pt1	1		0	0	0	0	0	0	0	0	0
Pt2	1	0		0	0	0	0	0	0	0	0
Pt3	1	0	0		0	0	0	0	0	0	0
Pt4	3	0	0	0		0	0	0	0	0	0
Pt5	1	0	0	0	0		0	0	0	0	0
Pt6	1	0	0	0	0	0		0	0	0	0
Pt7	1	0	0	0	0	0	0		0	0	0
Pt8	3	0	0	0	0	0	0	0		0	0
Pt9	1	0	0	0	0	0	0	0	0		0
Pt10	1	0	0	0	0	0	0	0	0	0	

C

	Pm	Pde	Pt1	Pt2	Pt3	Pt4	Pt5	Pt6	Pt7	Pt8	Pt9	Pt10
Pm		4	2	2	2	1	2	2	2	2	2	2
Pde	4		4	4	4	3	4	4	4	4	4	4
Pt1	2	4		0	0	0	0	0	0	0	0	0
Pt2	2	4	0		0	0	0	0	0	0	0	0
Pt3	2	4	0	0		0	0	0	0	0	0	0
Pt4	1	3	0	0	0		0	0	0	0	0	0
Pt5	2	4	0	0	0	0		0	0	0	0	0
Pt6	2	4	0	0	0	0	0		0	0	0	0
Pt7	2	4	0	0	0	0	0	0		0	0	0
Pt8	2	4	0	0	0	0	0	0	0		0	0
Pt9	2	4	0	0	0	0	0	0	0	0		0
Pt10	2	4	0	0	0	0	0	0	0	0	0	

Figure S6. CBC analysis. Each matrix represents the number of compensatory base changes (CBC) found between any given pair of species upon comparing their ITS2 secondary structures. A color gradient from red to green indicates lowest to highest numbers of CBCs found between each pair. Cm denotes *Cyclotella meneghiniana*, Pde denotes *Pseudo-nitzschia delicatissima*, Pm denotes *Pseudo-nitzschia multiseries*, Fc denotes *Fragilariopsis cylindrus*, and Pt denotes *Phaeodactylum tricornutum*. Cm is a centric diatom species, while Pde, Pm, Fc and Pt are pennate diatom species.

Population genetics analysis reveals heterozygosity as a measure of continuous admixing and unstable genetic population

We further wished to determine the genome wide nucleotide diversity across all the ecotypes. With low genetic differentiation at the intra-haplogroup level compared to other haplogroups, pairwise nucleotide diversity (π) estimated in non-overlapping 1 kb windows across all the ecotypes is 0.002 ± 0.001 per site on average. This indicates that any two homologous sequences taken at random across different populations will on average differ by ~0.2%, which is remarkably low in comparison with polymorphism estimates in other unicellular eukaryotes (Blanc-Mathieu et al. 2014; Flowers et al. 2015; Liti 2015). It is also slightly lower than dimorphic fungi such as *Candida albicans* (Hirakawa et al. 2015). Linkage disequilibrium (LD) analysis using only homozygous SNV sites revealed, on average, high linkage disequilibrium (LD > 0.7) over pairs of polymorphisms, as a consequence of the population structure. Within 5 kb, LD declines with the increase in pairwise distance between sites (Fig 2C).

Considering high heterozygosity in the dataset, we investigated whether we could find evidence of admixing between ecotypes. Therefore, we used ADMIXTURE (Alexander et al. 2009) to estimate the number of ancestral populations (K) of each ecotype. This allele-based, unsupervised clustering algorithm uses a cross-validation error rate (CVE) (Alexander and Lange 2011) to predict the most probable number of ancestral genomes that have influenced the genetic makeup of present individuals or populations. The best-fit maximum likelihood

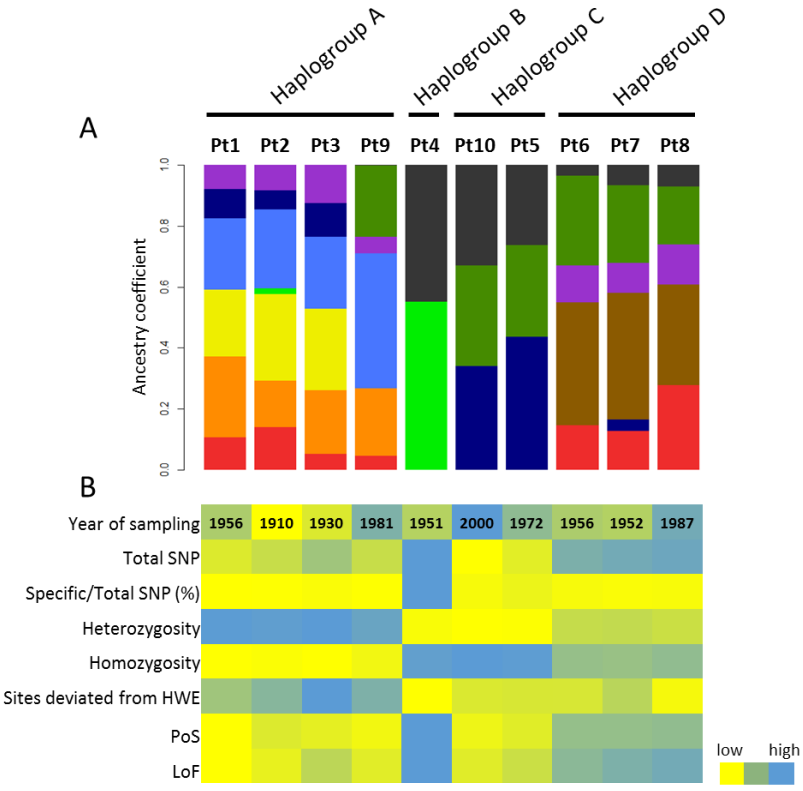
estimates of individual ancestry, quantified as K, is expected to exhibit a low cross-validation error compared to other values of K. Using the algorithm, we traced the influence of ten ancestral populations, as signified by lowest CV error value, on the genetic structure of the ecotypes (Fig S7). The cross-validation error deviation between K=6 through K=10 is relatively low indicating the insensitivity of cross-validation in population stratification when F_{ST} between the studied populations are low (Fig 2A), as also reported previously (Alexander and Lange 2011). However, although individual haplogroups maintain similar ancestral admixtures across K = 6 through K = 9 (Fig S7), at K = 10 the ancestry assignment differentiated the Pt4 specific composition, consistent with the high proportion of non-shared regions in the Pt4 genome (Fig 1A and Fig 1C). Further, the proportion of individual ancestral populations was found to be variable in each ecotype (Fig 3A), which could be a consequence of selection to local environments. The admixing does not correlate with the geographic distribution of most ecotypes (Fig 3A), suggesting non-restricted dispersal of *P. tricornutum* even though it is a coastal species. Furthermore, the admixtures are consistently maintained in ecotypes that were sampled in different years of the last century, indicating continuous genetic exchange between their ancestors (Fig 3B).

From the four distinguishable haplogroups, haplogroup A (Pt1, Pt2, Pt3 and Pt9) shows maximum admixture while haplogroup B (Pt4) has the least admixing with only two major ancestral admixtures (Fig 3A). Interestingly, this pattern of admixing is consistent with the high levels of heterozygosity within haplogroup A ecotypes (Pt1, Pt2, Pt3 and Pt9), where most of the alleles are also deviated from Hardy-Weinberg equilibrium (HWE) (Fig 3B). This suggests that the ecotypes are under isolate-breaking effect. Isolate breaking is a phenomenon where heterozygosity temporarily increases in the population when distinct populations/subpopulations interact and/or interbreed (Dorak 2014). Supporting the latter, most of the variant alleles in the Pt4 ecotype, which displays the lowest admixing, are

396 homozygous with relatively low proportions of heterozygous alleles deviating from HWE (Fig
397 3B). This suggests that most of the heterozygosity is due to frequent mixing of the
398 geographically isolated strains and that not all of the heterozygous alleles can be explained as
399 being selected under balancing selection.

400 Genetic admixing between distant populations can lead to their extinction (Wecek et al.
401 2016). This can account for the low abundance of *P. tricornutum* in nature, as genetic admixing
402 between the strains is continuous. Therefore, taking advantage of *Tara* Oceans (Karsenti et al.
403 2011) meta-genomics (MetaG) and meta-transcriptomics (MetaT) data that resulted from a
404 broad geospatial sampling of microeukaryotes (Carradec et al. Under revision), we attempted
405 to re-evaluate the abundance of *P. tricornutum* in nature, even though it has not been found
406 in 18S-based metabarcoding data from this project (Malviya et al. 2016). By following the
407 lowest common ancestor algorithm (LCA) and using PhyloDB (Dupont et al. 2015) as reference
408 database, we were able to track a wide albeit very low abundant distribution of *P. tricornutum*
409 (Fig 3C) in the world's ocean by taxonomically assigning *Tara* Oceans unigenes (Carradec et al.
410 *Under revision*) to *P. tricornutum*. From the set of unigenes assigned as unclassified diatoms
411 in the global ocean atlas of eukaryotic genes (Carradec et al. *Under revision*), we were able to
412 assign 70 unclassified diatom unigenes to best represent *Phaeodactylum* species complex (File
413 S5sheetA), with an average coverage of 80% - >95% and 60% - 80% DNA sequence level
414 identity (Fig S9A) over 118 PhyloDB reference genes of *P. tricornutum*. Most of the 118
415 reference genes, onto which the unigenes have the best hits, were specific to *P. tricornutum*
416 (Rastogi et al. Submitted) with highest (on average >94%) unigene coverage over them (Fig
417 S9B) (File S5sheetB). Since the percentage identity of the unigene sequences with *P.*
418 *tricornutum* references are quite low, the assignation might be biased towards
419 *Phaeodactylum* and the unigenes could be derived from unsequenced species, branching
420 between *P. tricornutum* and the next closest sequenced species. However, compared to *P.*

421 *tricornutum* reference genes that are shared broadly among the diatom clade (diatom specific
422 genes), the unigene coverage over *P. tricornutum* specific genes is very high (Fig S9B) and
423 define these genes as potential references for further experimental assessment of *P.*
424 *tricornutum* distribution in the open ocean. We then determined the relative expression of *P.*
425 *tricornutum* assigned unigenes using meta-transcriptomics (MetaT) data, which further
426 correlated with the distribution of *P. tricornutum* depicted by MetaG (Fig. S8). However,
427 relative to *Skeletonema* and *Pseudo-nitzschia*, genera that can also reproduce asexually with
428 limited admixing preferences (Casteleyn et al. 2010; Harnstrom et al. 2011), the abundance
429 of *P. tricornutum* is very low and is much less widely distributed (Fig. 3D, 3E, and S8).



430

431 **Figure 3. Population structure of the ecotypes.** (A) Admixture analysis of the 10 studied *P.*
432 *tricornutum* ecotypes using ADMIXTURE. The bar plot represents individual ancestries within
433 each ecotype and estimated using an unsupervised clustering algorithm employed in
434 ADMIXTURE, which predicted 10 (K=10) ancestral populations, represented with 10 different
435 colors. Y-axis represents the proportion of individual ancestry within each ecotype (represented
436 on top X-axis). (B) The color gradient from yellow to blue indicates low to high numerical
437 values across each ecotype (indicated on top X-axis of panel A) within different functional
438 categories indicated on Y-axis. These include (from top to bottom), Year of sampling = Year
439 in which the respective ecotype was sampled, Total SNP = Absolute number of SNPs found in

each ecotype, Specific/Total SNP (%) = percentage of ecotype specific SNPs, Heterozygosity = Number of heterozygous SNPs from a set of total SNPs within each ecotype, Homozygosity = Number of homozygous SNPs from a set of total SNPs within respective ecotype, Sites deviated from HWE = Number of SNP sites been predicted to be deviated from Hardy-Weinberg equilibrium (HWE), PoS = Number of genes under Positive Selection (PoS), LoF = Number of genes localizing Loss of Function variant sites.

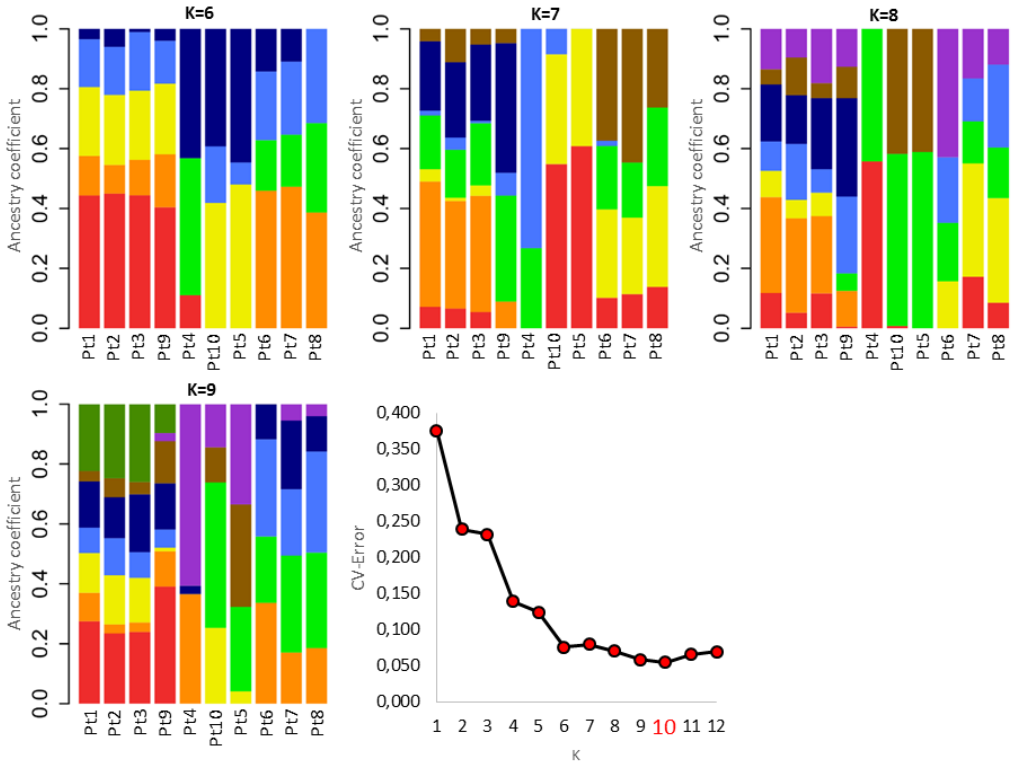


Figure S7. Allele-based estimates of ancestry in ten ecotypes of *P. tricornutum* as inferred by ADMIXTURE. Unsupervised results for K = 6-9 are displayed with the distribution of cross-validation error rate (represented on Y-axis of the line plot) across different values of K (represented on X-axis of the line plot). 5-folds cross validation indicated K = 10 as the best fit (represented in Fig. 3A) but given the limitation of ADMIXTURE in classifying populations with low Fst, and a stable distribution of CV-Error values from K = 6 through K = 10 compared to its distribution across K = 1:5, we therefore predictably focused on K = 6:10 ancestral populations.

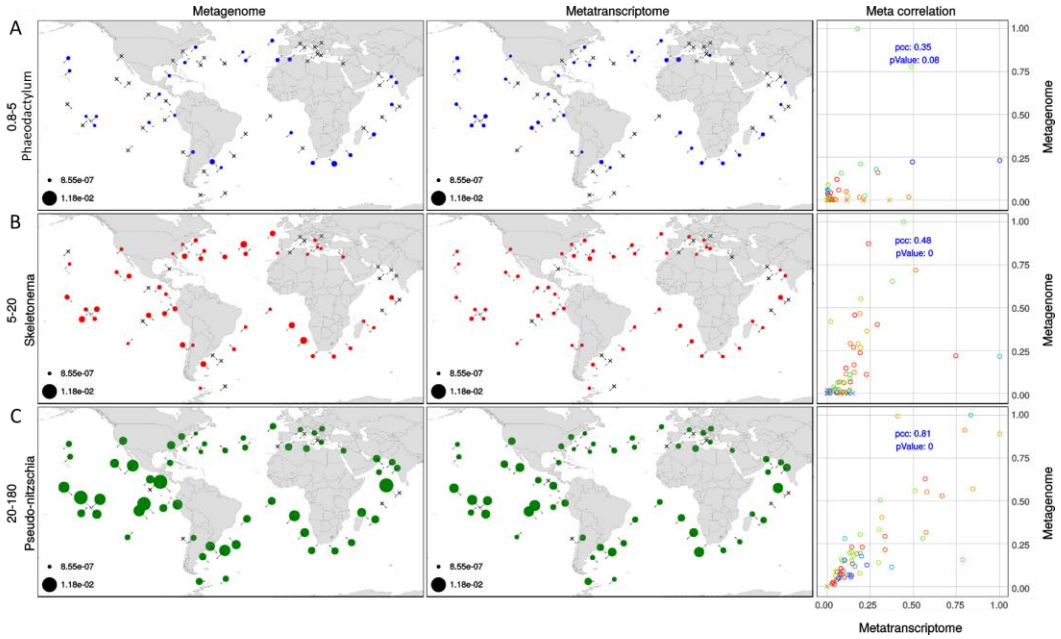


Figure S8. Global abundance and distribution of *Phaeodactylum*, *Skeletonema* and *Pseudo-nitzschia* genera based on *Tara* Oceans metagenomics and metatranscriptomics data. Unigene relative abundance from *Phaeodactylum* (A), *Skeletonema* (B) and *Psuedo-nitzschia* (C) detected within *Tara* Oceans samples are represented. Based on the known morphologies of the species within each genus, we investigated the abundance and distribution of these three genera within size fraction 0.8-5, 5-20 and 20-180 micrometers, respectively (Y-axis). The levels of relative abundance (Metagenome, MetaG) and expression (Metatranscriptome, MetaT) of each clade are represented by the disk area. We highlighted by "X", the *Tara* Oceans stations where a clade was not detected. Meta correlations signify the relation between metagenome and metatranscriptome levels, expressed as a ratio of the total abundance and expression, respectively. Scatter plots indicate Pearson correlation coefficients (pcc) and p-values in blue.

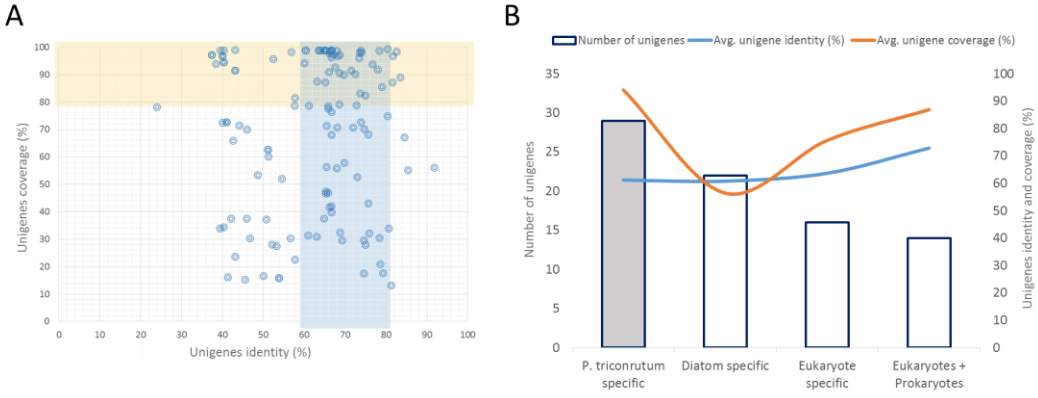


Figure S9. Conservation assessment of *P. tricornutum* genes across various taxonomic groups in the tree of life depicts maximum coverage of unclassified diatom unigenes over *P. tricornutum* specific genes. From our analysis of sequence divergence of *P. tricornutum* genes along the tree of life, we looked into the nature of *P. tricornutum* reference genes aligned to the *Tara* Oceans unigenes, which were taxonomically assigned as unclassified diatoms in previous study (Carradec et al. Under revision), with variable identity. (A) The scatter plot represents the percent coverage (Y-axis) and identity (X-axis) of all the unigenes mapped onto

the *P. tricornutum* reference genes. Most of the unigenes assigned as *Phaeodactylum* achieved 60 to 80% of DNA sequence identity (highlighted with blue) and 80% to >95% of coverage (highlighted in dusty yellow). (B) The bar plot represents the total number of reference genes with different conservation patterns (Left Y-axis). The line in the plot indicates (right Y-axis) the average percent coverage (Orange) and sequence identity (Blue) of all the unigenes on the reference genes clustered based on their conservation patterns (X-axis).

Functional characterization of polymorphisms suggests adaptation to laboratory conditions

Species are under continuous pressure to adapt to a changing environment over time. We therefore wanted to understand the functional consequences of the genetic diversity between the ecotypes. Localization of the polymorphic sites over genomic features (genes, transposable elements, and intergenic regions) revealed the highest number of variants over genes (Fig 2B), specifically on exons, and was consistent across all the studied ecotype populations. An average non-synonymous to synonymous variant ratio (N/S) was estimated to be ~0.87, which is higher than in *Chlamydomonas reinhardtii*, N/S = 0.58 (Flowers et al. 2015). We further identified genes within different haplogroups experiencing strong selection pressure based on their high Ka/Ks (dN/dS) ratios. Since decades, this ratio is widely adopted as a measure of selective pressure in a wide range of species (Nielsen and Yang 1998; Yang et al. 2000). Across all the ecotypes, 128 genes displaying positive selection (PoS) could be detected, among which 47% are specific to one or more haplogroups (Fig 4A). Furthermore, many genes (902) were found to have loss-of-function (LoF) variant alleles (Fig 4A), including frame-shift mutations and mutations leading to theoretical start/stop codon loss or gain of premature start/stop codons. Consistent with our observations of high admixing leading to high heterozygosity in haplogroup A (Pt1, Pt2, Pt3 and Pt9) compared to other haplogroups such as haplogroup B (Pt4), hence making selection of loci limiting, we observed very few loci to be under natural selection (PoS) within haplogroup A ecotypes compared to haplogroup B (Fig 3B).

Based on the presence of functional domains, all *P. tricornutum* annotated genes (Phatr3, http://protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index/) were grouped into 3,020 gene families. These families can be as large as the reverse transcriptase gene family, which is also highly abundant in marine plankton (Lescot et al. 2016), representing 149 candidate genes having reverse transcriptase domains, or as small as families that constitute single gene candidates. Across all the ecotypes, we observed that the majority of genes experiencing LoF mutations belong to large gene families (Fig 4B). This is consistent with a previous observation of the existence of functional redundancy in gene families as a balancing mechanism for null mutations in yeast (Gu et al. 2003). Therefore, to estimate an unbiased effect of any evolutionary pressure (LoF allele or Positive selection mutations) on different gene families, we calculated a ratio, named the effect ratio (EfR, see Methods), which normalizes the fact that if any gene family has enough candidates to buffer the effect on some genes influencing evolutionary pressure, it will be considered as being less affected compared to those for which all or most of the constituents are under selection pressure. From this analysis, each haplogroup displayed a specific set of gene families to be under selection pressure (Fig S9). Significantly enriched biological processes (chi-squared test; P-value<0.05) associated to haplogroup A-specific gene families that are under selection included chlorophyll biosynthetic process, DNA intergration, fructose 6-phosphate metabolic process and pteridine-containing compound metabolic processes. Similarly, haplogroup B-specific gene families that are under strong adaptive selection exhibit significant enrichment of processes such as posttranslational protein targeting to membrane, translocation, RNA 3'end processing involving polyadenylation, and terpenoid biosynthetic process. Haplogroup C gene families that are under selection include the enrichment of processes like DNA-templated transcription, histone acetylation and vesicle-mediated transport. Likewise, haplogroup D-specific gene families that are under selection include biological processes such as arginyl-tRNA aminoacylation, intracellular signal transduction, lipid catabolic process and N-glycan

541 processing. Apart from the haplogroup specific families that are under selection pressure,
542 across all the ecotypes a group of gene families associated to methionine biosynthesis (MetH,
543 Phatr3_J23399) was also observed as experiencing strong adaptive selection (Fig S9). All
544 enriched biological processes associated to genes under positive selection in each haplogroup
545 can be found in File S3.
546 In *P. tricornutum*, MetE (cobalamin-independent methionine synthase) and MetH (cobalamin-
547 dependent methionine synthase) are known to catalyze homocysteine to methionine in the
548 presence of symbiotic bacteria and vitamin B12 in the growth media, respectively. Previous
549 reports have suggested that growing axenic cultures in conditions of high cobalamin
550 availability results in repression, leading to the loss of MetE function and high expression of
551 the MetH gene in *P. tricornutum* and *C. reinhardtii* (Helliwell et al. 2011; Bertrand et al. 2012;
552 Helliwell et al. 2015). In accordance with these results, we observed a high expression of *MetH*
553 in axenically grown laboratory cultures (Fig 4C) and thus a strong selection signal over the
554 *MetH* gene. We speculate this to happen because of the high availability of cobalamin in the
555 laboratory growth media used to maintain all the ecotype strains over the last decades, which
556 might be due to evolution triggered by laboratory culture conditions. However, we were not
557 able to trace any significant signature for the loss of *MetE* gene although its expression is
558 significantly lower in axenic cobalamin containing cultures (Fig 4D), suggesting that its loss
559 might require further generations, or that adaptation to new conditions requires the silencing
560 of *MetE* without its complete loss which likely involves epigenetic mediated regulation. Similar
561 observations were obtained for *CBA1* and *SHMT* genes in *P. tricornutum* and *T. pseudonana*
562 (Bertrand et al. 2012), which under cobalamin scarcity enhance cobalamin acquisition and
563 manage reduced methionine synthase activity, respectively (Fig 4E and 4F).
564 Considering all pairwise correlated gene families exhibiting similar selection signals (Positive
565 selection (PoS) and Loss of function mutations (LoF)), measured using EfR, among the 10
566 ecotypes, we used hierarchical clustering to examine the functional closeness of ecotype

populations with one another. Consistent with the population's genetic structure, ecotypes within individual haplogroups were more closely related than the ecotypes belonging to other haplogroups (Fig S10A and S10B), suggesting variation in functional relatedness between different proposed haplogroups.

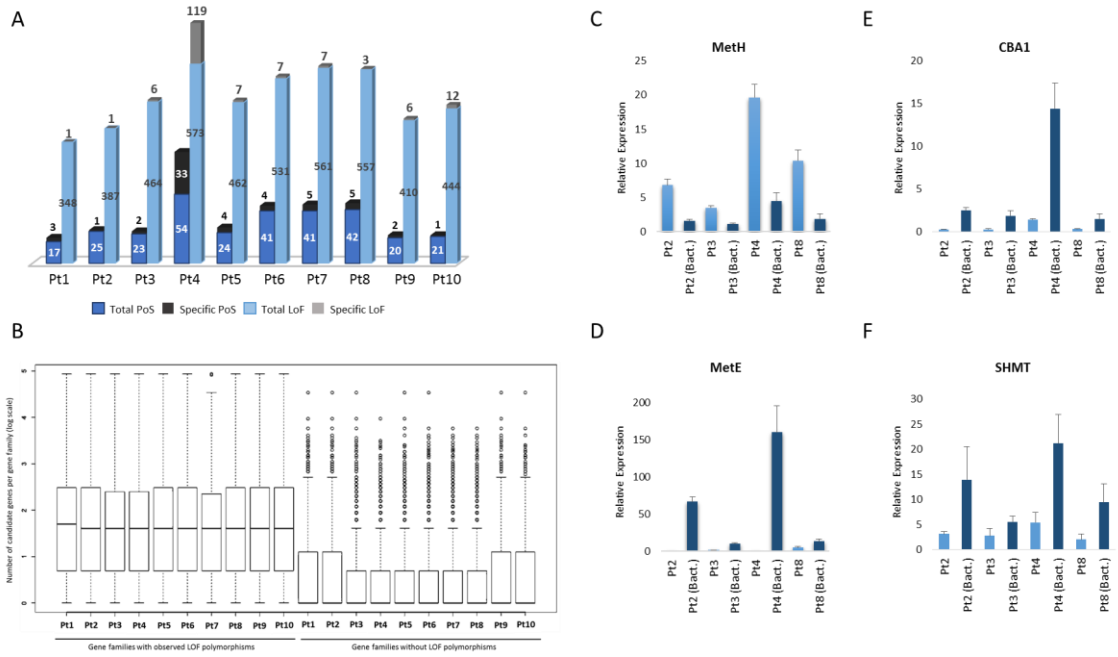
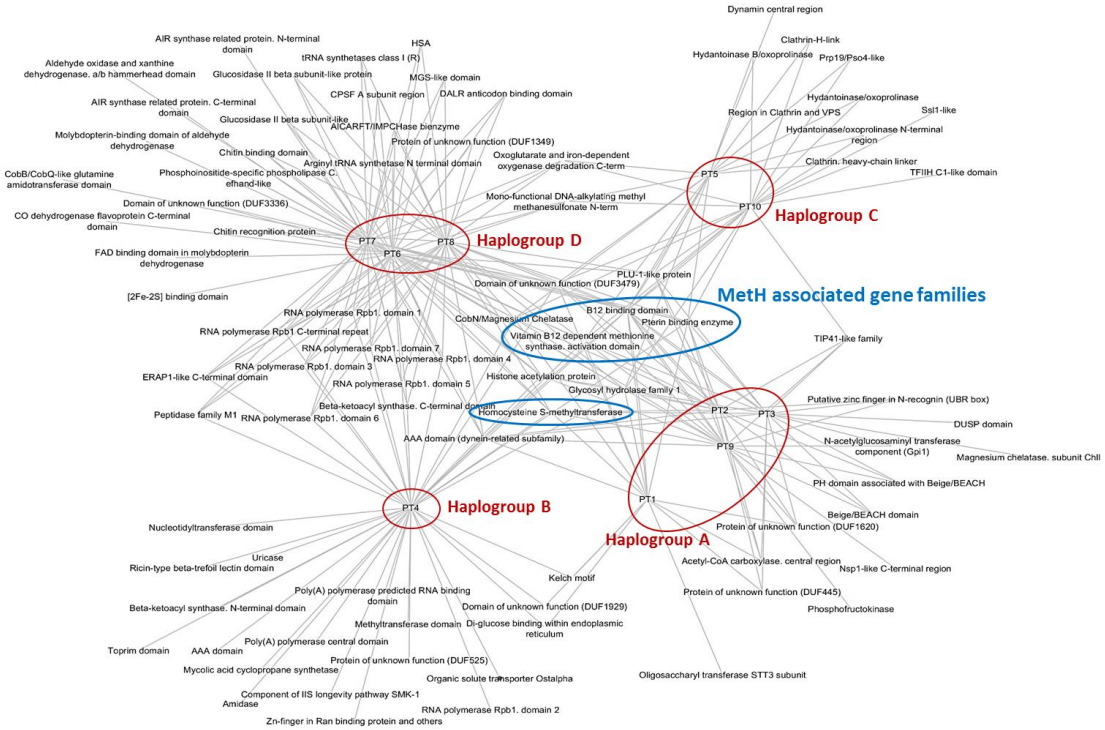
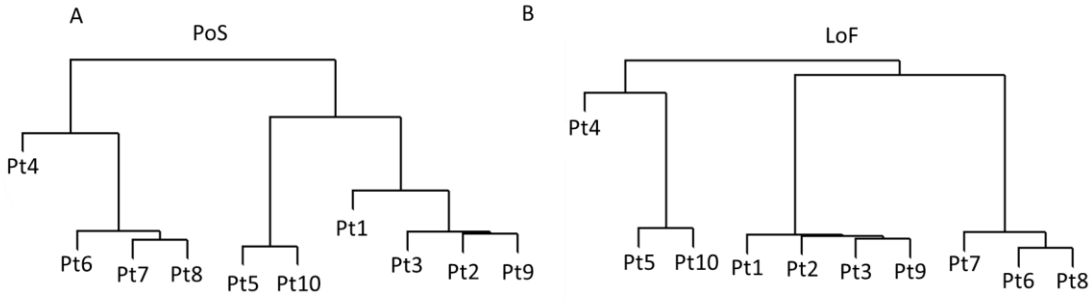


Figure 4. Evolutionary and functional consequences of polymorphisms. (A) The bar plot represents total and specific number of genes that are subject to positive selection, or experiencing loss-of-function (LoF) mutations. For each category, the ecotypes are plotted as stack plot with total and specific numbers of genes. Numbers of each gene in each category is indicated. (B) The box plot represents the number of gene families being affected by loss-of-function (LOF) mutations and suggests a bias of such mutations on the genes belonging to large gene families. Y-axis represents, as log scale, the number of genes in the gene families vs those that are not affected by LOF mutations. (C, D, E and F) The bar plots represent relative expression of *MetH*, *MetE*, *CBA1* and *SHMT* genes in four (Pt2, Pt3, Pt4 and Pt8) of the ten ecotypes with the presence of vitamin B12 in axenic cultures (light blue bars) and bacteria with no vitamin B12 in the growing media (dark blue bars).



585

586 **Figure S10. Natural selection within Haplogroups.** Based on the EfR, the network displays
587 highly affected gene families experiencing positive selection. Gene families associated to MetH
588 genes under positive selection in all the ecotypes are indicated within blue circle.
589



590
591 **Figure S11. Functional clustering of the ecotypes.** Hierarchical clustering using Pearson
592 pairwise correlation of effect ratio (EfR) measured for affected gene families within each
593 functional category [(A) Positive Selection (PoS); (B) Loss of Function (LoF)] to examine the
594 functional congruency between all the ecotypes.
595

596

597 **Discussion**

598 Using whole genome sequence analysis of *P. tricornutum* ecotypes sampled over a time span
599 of a century, this study reveals continuous gene flow between geographically isolated
600 populations, in line with previous observations indicating that dispersal is not limited within

601 diatoms (Finlay 2002; Cermeno and Falkowski 2009). It further sheds light on the subsequent
602 effects of non-restricted dispersal and admixing of the genetic makeup of each ecotype,
603 revealing heterozygosity as an indicator of unstable genetic structure of a diatom population.
604 Furthermore, considering the low abundance of *P. tricornutum* in the ocean, we propose that
605 dynamic environmental niche coupled with repeated gene flow between geographically
606 isolated populations act as strong bottlenecks on species richness possibly through a gradual
607 decrease in the effective population size coupled with Muller's ratchet (Crow 2005). Finally,
608 the study points out the effect of domestication on restructuring of the functional pathways
609 within diatoms, leading to the selection of a change in functional phenotype.
610 Following reference-based assembly, a high (>90%) coverage of the reference genome could
611 be mapped by sequencing reads from individual ecotype WGS libraries. The unmapped
612 reference genome, however, is a consequence of partial reference genome assembly,
613 transpositions and copy number variations within individual ecotype genomes. From the
614 study, we found heterozygous alleles to account for most of the genomic diversity between
615 the ecotypes. Despite high variability in the levels of heterozygosity between different
616 ecotypes, the mutational spectrum, compared to the reference and across all the ecotypes
617 consisted of high G:C -> A:T and A:T -> G:C transitions. Deamination of cytosines dominantly
618 dictates C to T transitions in both plants and animals (Becker et al. 2011; Rahbari et al. 2016),
619 and CpG methylation potential of the genome is greatly influenced by heterozygous SNPs in
620 the CpG dinucleotides (Shoemaker et al. 2010). Previous studies have demonstrated low DNA
621 methylation in *P. tricornutum,* using Pt1 8.6, a monoclonal strain isolated from Pt1 single cell,
622 as a reference. Because the Pt1 ecotype attains maximum levels of heterozygous variant
623 alleles, testing for DNA methylation patterns across different ecotypes, maintaining variable
624 levels of heterozygosity, may provide an opportunity to dissect cross-talk between loss of
625 heterozygosity and DNA methylation in the selection of certain traits (Kanai et al. 2000).

626 Heterozygosity is higher in ecotypes sharing numerous ancestral admixtures (Pt1, Pt2, Pt3,
627 Pt9, Pt6, Pt7 and Pt8) compared to the ecotypes with low admixtures (Pt4, Pt5 and Pt10)
628 indicating a prominent role of continuous gene flow in retaining high heterozygosity.
629 However, heterozygosity may also be maintained under balancing selection (Sellis et al. 2011;
630 Sellis et al. 2016) where heterozygous loci are more advantageous than homozygous loci
631 (Ferreira et al. 2011). This phenomenon has recently been studied in diatoms where
632 maintenance of bi-allelic expression of numerous loci has been demonstrated in the cold-
633 adapted diatom species *Fragilariopsis cylindrus* (Mock et al. 2017). Furthermore, continuous
634 and non-restricted admixing between *P. tricornutum* populations can account for its low
635 abundance in nature. As with fast reproducing rates and dynamic ecological niche in locations
636 where population size is small, frequent admixing can result in selecting unfavorable alleles in
637 the resident population, leading to extinction (Wecek et al. 2016). The phenomenon is more
638 plausible relative to *Skeletonema marinoi* and *Pseudo-nitzschia pungens*, which are abundant
639 diatom species (Kooistra et al. 2008) and yet, regardless of the frequent dispersal of
640 geographically structured strains, they refrain from genetic admixing (Casteleyn et al. 2010;
641 Harnstrom et al. 2011).
642 Further, based on low genetic diversity across all the ecotypes and the presence of multiple
643 variant positions within the 18S and ITS2 reference sequences, we can cluster all ecotypes
644 into four haplogroups, which is in consensus with previous studies (De Martino 2007). The
645 topology of association between the haplogroups was further confirmed to exhibit coherency
646 at the whole genome level. Hierarchical clustering based on the variation in the number of
647 copies of different genes and TEs shared among ecotypes further confirms the associations of
648 ecotypes at both inter and intra haplogroup level. Most of the shared genetic structure
649 between the ecotypes is independent of their geographical distribution and their year of
650 sampling, suggesting unceasing and non-restricted dispersal and genetic exchange between

651 different geographically isolated ecotypes, and only a limited impact of laboratory culturing
652 on genome rearrangements.
653 The mechanisms and conditions that favor admixture are not clear because low
654 recombination signatures were detected between the ecotypes, consistent with the observed
655 high levels of LD between the alleles. However, various components (genes) of meiosis
656 pathway are conserved in *P. tricornutum* as well as in other diatom species known to undergo
657 sexual reproduction (Patil et al. 2015). These components include genes that have other
658 known functions outside meiosis along with many genes whose functional role is limited to
659 meiosis. This suggests the possibility of sexual reproduction within *P. tricornutum* and can
660 explain the mechanism of genetic exchange between different ecotypes. Further, the absence
661 of contemporary base changes (CBC) within ITS2 gene secondary structure between all the
662 ecotypes, compared to the presence of many CBCs between *P. tricornutum* ecotypes and
663 other diatom species, suggests that the ecotypes may be able to reproduce sexually. Our
664 analysis reveals broad geo-spatial *P. tricornutum* distribution in the world's ocean suggesting
665 that the dispersal of ecotypes to different localities is occurring and may be fostered by ocean
666 currents (Whittaker and Rynearson 2017), human activities like rafting, ballasting (Thiel 2005;
667 Nikula et al. 2013), and migration of birds (Schlichting 1960; Proctor 1966; Foissner 2006).
668 Consistent with the nearly-neutral theory of molecular evolution (Kimura 1983), only a few
669 genes were found to be under strong selection pressure across all the ecotypes. A remarkably
670 high expression and selection pressure of the MetH gene suggests a strong adaptive selection
671 within laboratory-maintained ecotype strains. Since all the ecotype strains share similar
672 microenvironments in their respective natural niches, most of the genes that are under strong
673 selection pressure are shared among different ecotypes. However, a few characteristic gene
674 families are observed to be under selection specifically within individual haplogroups. These
675 species complexes are further supported by functional specialization of individual groups,
676 nicely illustrated with Pt4 in haplogroup B. Pt4 shows a low non-photochemical quenching

677 capacity (NPQ) (Bailleul et al. 2010), which is suggested as an adaptive trait to low light
678 conditions. Specifically, this ecotype has been proposed to have established an upregulation
679 of a peculiar light harvesting protein LHCX4 in extended dark conditions (Bailleul et al. 2010;
680 Taddei et al. 2016). In line with these observations, a gene involved in nitrate assimilation
681 (Phatr3_EG02286) in haplogroup B shows high copy numbers, suggesting an altered mode of
682 nutrient acquisition within this haplogroup. Nitrate assimilation was shown to be regulated
683 extensively under low light or dark conditions to overcome nitrate limitation of growth in
684 *Thalassiosira weissflogii* (Clark et al. 2002). Pt4 is well adapted to its low light ambient
685 environment which may well affect nitrate assimilation capacity (Ivanikova et al. 2005;
686 Weiguo Li 2011) and thus the growth rate of the strain in different conditions. Similarly, a gene
687 encoding an amino acid transporter (Phatr3_J50146) was found to be positively selected in
688 Pt4, further suggesting a role in nutrient uptake. Interestingly, a predicted seven
689 transmembrane receptor (Phatr3_J11183) belonging to the rhodopsin gene family is also
690 observed to be under positive selection within Pt4 (haplogroup B) ecotype. It is tempting to
691 speculate about its role in light perception and photo sensing in the low light environments at
692 high latitudes. Furthermore, haplogroup C (Pt5 and Pt10) contains a protein with a possible
693 role in adhesion, which is in line with the high adherence reported in Pt5 (Stanley 2007).
694 Additional functions emerging from this haplogroup include vacuolar sorting and vesicle-
695 mediated transport which could be an indication of altered intracellular trafficking (Pickett-
696 Heaps and Forer 2001).
697 In conclusion, the study brings new insights to our understanding of diatom ecology and
698 evolution. The study reveals the global distribution map of the best-studied model diatom
699 species *P. tricornutum* and recovered global patterns of ancestral admixing between
700 geographically distant ecotypes sampled at a broad temporal scale. As strains maintain high
701 levels of heterozygosity, with possible selective functional preference of one allele over the
702 other under different environmental conditions, the current study will be useful in

703 deciphering the mechanisms underpinning allele divergence and selection within diatoms,
704 and help understand the genetic basis of their success in diverse ocean ecosystems. This study
705 further provides the community with genomic sequences of *P. tricornutum* natural accessions
706 that are valuable and will be undoubtedly used for functional studies.

707

708 **Methods**

709

710 **Sample preparation, sequencing and mapping**

711

712 Ten different accessions of *P. tricornutum* were obtained from the culture collections of the
713 Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP,
714 Pt1=CCMP632, Pt5=CCMP630, Pt6=CCMP631, Pt7=CCMP1327, Pt9=CCMP633), the Culture
715 Collection of Algae and Protozoa (CCAP, Pt2=CCAP 1052/1A, Pt3= CCAP 1052/1B, Pt4= CCAP
716 1052/6), the Canadian Center for the Culture of Microorganisms (CCCM, Pt8=NEPCC 640), and
717 the Microalgae Culture Collection of Qingdao University (MACC, Pt10=MACC B228). All of the
718 accessions were grown axenically in batch cultures with a photon fluency rate of 75 µmol
719 photons m-2 s-1 provided by cool-white fluorescent tubes in a 12:12 light: dark (L:D)
720 photoperiod at 20 °C. Exponentially growing cells were harvested and total DNA was extracted
721 with the cetyltrimethylammonium bromide (CTAB) method. At least 6 µg of genomic DNA
722 from each accession was used to construct a sequencing library following the manufacturer's
723 instructions (Illumina Inc.). Paired-end sequencing libraries with a read size of 100 bp and an
724 insert size of approximately 400 bp were sequenced on an Illumina HiSeq 2000 sequencer at
725 Berry Genomics Company (China). Low quality read-pairs were discarded using FASTQC with
726 a read quality (Phred score) cutoff of 30. Using the genome assembly published in 2008 as
727 reference (Bowler et al. 2008), we performed reference-assisted assembly of all the ecotypes.

We used BOWTIE (-n 2 –X 400) for mapping the high quality NGS reads to the reference genome followed by the processing and filtering of the alignments using SAMTOOLS and BEDTOOLS. Detailed methods are provided in File S6.

Discovery of small polymorphisms and large structural variants

GATK (McKenna et al. 2010), configured for diploid genomes, was used for variant calling, which included single nucleotide polymorphisms (SNVs), small insertions and deletions ranging between 1 and 300 base pairs (bp). The genotyping mode was kept default (genotyping mode = DISCOVERY), Emission confidence threshold (-stand_emit_conf) was kept 10 and calling confidence threshold (-stand_call_conf) was kept at 30. The minimum number of reads per base, to be called as a high quality SNV, was kept to 4 (read-depth >=4x).

Next, considering Z-score as a normalized measure of read-depth, gene and TE candidates showing multiple copies (representing CNV) or apparently being lost (representing gene loss) were determined. For TE CNV analysis, TEs (from current annotation version Phatr3, (Rastogi et al. Submitted)) that are more than 100 bp lengths were considered. We measured the fold-change (Fc) by dividing normalized read depth per genomic feature (Z-score per gene or TE) by average of normalized read depth of all the genes/TEs(average Z-score), per sample. Later genes or TEs with log2 scaled fold change >=2 were reported and considered to exist in more than one copy in the genome. Genes where the reads from individual ecotype sequencing library failed to map on the reference genome were considered as potentially lost within that ecotype and reported. Detailed method is provided in File S6. Later, some randomly chosen loci were picked and validated to be lost in the ecotypes compared to the reference genome by qPCR analysis.

Validation of gene loss and quantitative PCR analysis

In order to validate gene loss, DNA was extracted from all the ecotypes as described previously (Falciatore et al. 1999) and PCR was performed with the primers listed in Table S1. PCR products were loaded in 1% agarose gel and after migration gels were exposed to UV light and photographs were taken using a gel documentation apparatus to visualize the presence and absence of amplified fragment. To assess gene expression, RNA was extracted as described in (Siaut et al. 2007) from ecotypes grown axenically in Artificial Sea Water (ASW) (Vartanian et al. 2009) supplemented with vitamins as well as in the presence of their endemic bacteria in ASW without vitamins. qPCR was performed as described previously (Siaut et al. 2007).

gene ID	forward primer	reverse primer	fragment length (bp)
Phatr3_EG00392	AGCATTGTAAATTGCGGAAC	GAAGATCTCTTCCGGGACTC	509
Phatr3_J13031	ATTCACAGACAATGCCGAAT	GGGGGCTAAGAGCTTACAAC	411
Phatr3_J15301	CCGTCATCAACAAAAACACA	ATGCAAGGAGCATTTTTCAG	422
Phatr3_J39504	TCGCGAATAACTCACAGTCA	TCTTGGAATACTTCGGCTTG	412
Phatr3_J40834	AGTGCATCGAAAGTCTGGAG	GTTACAGGGTGCGCTTTTTA	403
Phatr3_J41518	GACAAACTTGTCGTGGCTTT	CAAAGTACTCGGCTCCTTCA	427
Phatr3_J42255	CGTTCCTGGATTGAAAAATG	CCAATGAAGCTGCAGAAGAT	433
Phatr3_J42686	GCAGTCTTTTCAAGCGAGTC	CTAACCTTTGCGACGAACAT	409
Phatr3_J43417	TGAGTACGAGGCAAGTGTCA	CCAAAGGAATTGAGCGTAGA	429
Phatr3_J45389	CAGAAGTAGTTTCCCGACGA	GCTCTGGCTTTGTCTGCTAC	418
Phatr3_J46603	ACCAACGGCTGTATGTGTTT	TCCAGCTCCGTTTTGTAAAG	429
Phatr3_J46954	TTACATGATGGCTGGGAAGT	GACTCAACAACATCCCGTTC	443
Phatr3_J47779	GGACAATCAGACCCATTACG	GACCGTTCCATCATTCTGAG	443
Phatr3_J47781	AGCATGATTCTAGCCGACAC	CCCAATCATTTGATGAAAGC	414
Phatr3_J47783	CTCATTCCTTAGCCGACGTA	CATTGTTTACCCGAAACGAC	441
Phatr3_J47785	GATTGGCAGGATTGCTCTAA	CTATGGCTTGCACTTCCTGT	445
Phatr3_J47798	GCAGATCTGGAAGAAAACGA	ATGTTCTCCGCATCCAATAA	435
Phatr3_J48335	AAGCCTTTCACATGCTTCAC	ACTAACGTGCCATTGAGAGC	430
Phatr3_J48801	AGGAATTCCCTGTAGGAACG	CCTGCATAGCCTTGTATGCT	411
Phatr3_J49108	CTCACACCTTCGCAAAAAGT	AATGGATTTCTTCCCTTTGG	402
Phatr3_J54981	GACGCGACTTTCAAAACAGT	CATGAAGCTAAGGGCGTAAA	412

Table S1. List of PCR primers used to validate the gene loss candidates (listed in column 1).

P. tricornutum population structure

***Haplotype analysis*:** First, to cluster the ecotypes as haplogroups, ITS2 gene (chr13: 42150-43145) and 18S gene (chr13: 43553-45338) were used. Polymorphic sites across all the ecotypes within ITS2 and 18S genes were called and used to generate their corresponding

ecotype specific sequences, which were then aligned using CLUSTALW. The same approach was employed to perform haplotype analysis at the whole genome scale. Later, a maximum likelihood algorithm was used to generate the 18S, ITS2 and, whole genome tree with bootstrap values of 1,000. We used MEGA7 (Kumar et al. 2016) to align and deduce the phylogenetic trees.

***CBC analysis*:** CBC analysis was done by generating the secondry structure of ITS2 sequences, using RNAfold (Lorenz et al. 2011), across all *P. tricornutum* ecotypes and other diatom species. The other species include one centric diatom species *Cyclotella meneghiniana* (AY906805.1), and three pennate diatoms *Pseudo-nitzschia delicatissima* (EU478789.1), *Pseudo-nitzschia multiseries* (DQ062664.1), *Fragilariopsis cylindrus* (EF660056.1). The centroid secondry structures of ITS2 gene with lowest minimum free energy was used for CBC analysis. We used 4SALE (Seibel et al. 2006) for estimating the presence of CBCs between the secondry structure of ITS2 gene across all the species.

***Population genetics*:** Further, we measured various population genetic functions to estimate the effect of evolutionary pressure in shaping the diversity and resemblance between different ecotype populations. Within individual ecotypes, by using approximate allelic depths of reference/alternate alleles, we calculated the alleles that are deviated from Hardy Weinberg equilibrium (HWE). We used chi-square estimation to evaluate alleles observed to deviate significantly (P-value < 0.05) from the expected proportion as per [p2 (homozygous) + 2pq (heterozygous) +q2 (homozygous) =1) and should be 0.25% + 0.50% + 0.25%. Alleles were considered heterozygous if the proportion of ref/alt allele is between 20-80%. The proportion of ref/alt allele was calculated by dividing the number of reads supporting ref/alt base change by total number of reads mapped at the position. We evaluated average R^2 as a function to measure the linkage disequilibrium with increasing distance (1 kb, 5 kb, 10 kb, 20 kb, 30 kb,

40 kb and 50 kb) between any given pair of mutant alleles across all the ecotypes using expectation-maximization (EM) algorithm deployed in the VCFtools. Although no recombination was observed within the ecotypes, attempts were made to look for recombination signals using LDhat (Auton and McVean 2007) and RAT (Etherington et al. 2005). Again using VCFtools, Nucleotide diversity (π) was estimated in a 1 kb non-overlapping window along the whole genome across all the ecotypes, using the method described by (Nei and Li 1979). Genetic differentiation or variability between the ecotypes was further assessed using the mathematical function of Fixation index (FST), as described by Wright in 1931, as also stated in (Whitlock and McCauley 1999; Rottenstreich et al. 2007). We estimated FST as a function to measure, mathematically, the similarity between different pairs of ecotypes sharing multiple SNV positions using the following formula, $FST = \frac{Hp-He}{Hp}$, where *Hp* and *He* represent the total number of polymorphic positions between any given pair of ecotypes and number of total polymorphic sites within an individual ecotype, respectively.

***Admixture analysis*:** Ancestral admixture within ecotypes was estimated using ADMIXTURE (version linux-1.3.0) (Alexander et al. 2009), PLINK (version 1.07-x86_64) (Purcell et al. 2007) and VCFtools (version 0.1.13) (Danecek et al. 2011). In the absence of data from individuals of each ecotype/sample, we assumed the behavior of each individual in a sample to be coherent. Conclusively, instead of estimating the genetic structure within an ecotype, we compared it across all the ecotypes. Using VCFtools and PLINK to format the VCF file, containing variant information of all the ecotypes, to ADMIXTURE accepted format, we first estimated the possible number of ancestral populations (K) by using cross-validation error (CV error) function of ADMIXTURE (Alexander and Lange 2011). Finally, we used ADMIXTURE with 200 bootstraps, to estimate the admixing within individual ecotypes by considering the number of ancestral populations derived via CV-error function.

***Distribution of* P. tricornutum *sequences in* Tara *Oceans meta-genomics (MetaG) and meta-transcriptomics (MetaT) datasets*:** Distribution was estimated using the assembled *Tara* Oceans unigenes from meta-genomics (metaG) and meta-transcriptomics (metaT) read data (Carradec et al. Under revision). We used the Lowest Common Ancestor (LCA) algorithm (Garcia-Etxebarria et al. 2014) to classify unigenes according to the highest score BLAST (Altschul et al. 1990) hits (at least 50% of coverage and 3 orders of magnitude under the best hit). The BLAST reference database was composed of PhyloDB release 1.076 (Dupont et al. 2015). The abundance of each *P. tricornutum*-assigned unigene is relative to the abundance of all the Bacillariophyta assigned unigenes. PhyloDB contains only one reference for genus *Phaeodactylum* compared to 12 reference species of *Skeletonema* and 11 reference species of *Pseudo-nitzschia*. Therefore, the relative abundances of *Skeletonema* and *Pseudo-nitzschia* assigned unigenes were further normalized with the total number of references from them in PhyloDB. From all the size fractions sampled during *Tara* Oceans expeditions (Karsenti et al. 2011), we considered analyzing data corresponding to size fraction 0.8 - 5 micrometers for *P. tricornutum*, 5 – 20 micrometers for *Skeletonema*, and 20 – 180 micrometers for *Pseudo-nitzschia*. Detailed methods are provided in File S6.

Functional characterization of polymorphisms

snpEff (Cingolani et al. 2012) and KaKs (Zhang et al. 2006) calculator were used to annotate the functional nature of the polymorphisms. Along with the non-synonymous, synonymous, loss-of-function (LOF) alleles, transition to transversion ratio and mutational spectrum of the single nucleotide polymorphisms were also measured. Genes with Ka/Ks also known as dN/dS ratio more than 1 with a p-value less than 0.05 are considered as undergoing natural or Darwinian selection. Various in-house scripts were also used at different levels for analysis

and for plotting graphs. Data visualization and graphical analysis were performed principally using ClicO (Cheong et al. 2015), CYTOSCAPE (Shannon et al. 2003), IGV (Robinson et al. 2011) and R (https://www.r-project.org/about.html). Based on the presence of functional domains all the Phatr3 genes (http://protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index) were grouped into 3,020 gene families. Subsequently, the constituents of each gene family was checked for being either affected by loss-of-function mutations or experiencing natural selection. To estimate an unbiased effect of any evolutionary pressure (LoF allele or Positive selection mutations) on different gene families, induced because of high functional redundancies in the gene families, a normalized ratio named as effect ratio (EfR), was calculated. Precisely, the EfR normalizes the fact that if any gene family have enough candidates to buffer the effect on some genes influencing evolutionary pressures, it will be considered as less affected compared to the situation where all or most of the constituents are under selection pressure. The ratio was estimated as shown below and gene families with EfR larger than 1 were considered as being significantly affected.

$$Effect\ Ratio\ (EfR) = \frac{\frac{Number\ of\ genes\ affected\ within\ the\ given\ gene\ family}{Total\ number\ of\ genes\ in\ the\ given\ gene\ family}}{\frac{Total\ number\ of\ genes\ affected\ in\ all\ the\ gene\ families}{Total\ number\ of\ genes\ in\ all\ the\ gene\ families}}$$

Additionally, significantly enriched (chi-square test, P-value < 0.05) biological processes associated within genes experiencing LoF mutations, natural selection (PoS), or showing CNV, or being lost (GnL), were estimated by calculating observed to expected ratio of their percent occurrence within the given functional set (PoS, LoF, CNV, LoF) and their occurrence in the complete annotated Phatr3 (http://protists.ensembl.org/Phaeodactylum_tricornutum/Info/Index) (Rastogi et al. Submitted) biological process catalog. Later, considering gene family EfR as a function to measure the association rate, we deduced Pearson pairwise correlations between different ecotypes. The correlation matrix describes that if many equally

affected gene families are shared between any given pair of ecotypes, they will have higher correlation compared to others. Finally, hierarchical clustering using Pearson pairwise correlation matrix assessed the association between the ecotypes.

Acknowledgements

CB acknowledges funding from the ERC Advanced Award 'Diatomite', the LouisD Foundation of the Institut de France, the Gordon and Betty Moore Foundation, and the French Government 'Investissements d'Avenir' programmes MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-1253 11-IDEX-0001-02), and OCEANOMICS (ANR-11-BTBR-0008). CB also thanks the Radcliffe Institute of Advanced Study at Harvard University for a scholar's fellowship during the 2016-2017 academic year. AR was supported by an International PhD fellowship from MEMO LIFE (ANR-10-LABX-54).

Conflict of interest

The authors declare no conflicts of interest.

References

Abida H, Dolch LJ, Mei C, Villanova V, Conte M, Block MA, Finazzi G, Bastien O, Tirichine L, Bowler C et al. 2015. Membrane glycerolipid remodeling triggered by nitrogen and phosphorus starvation in Phaeodactylum tricornutum. *Plant physiology* **167**: 118-136.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC bioinformatics* **12**: 246.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655-1664.

Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR et al. 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**: 203-207.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Armbrust EV. 2009. The life of diatoms in the world's oceans. *Nature* **459**: 185-192.

905 Auton A, McVean G. 2007. Recombination rate estimation in the presence of
906 hotspots. *Genome Res* **17**: 1219-1227.
907 Bailleul B, Rogato A, de Martino A, Coesel S, Cardol P, Bowler C, Falciatore A,
908 Finazzi G. 2010. An atypical member of the light-harvesting complex stress-
909 related protein family modulates diatom responses to light. *Proceedings of*
910 *the National Academy of Sciences of the United States of America* **107**:
911 18214-18219.
912 Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D. 2011.
913 Spontaneous epigenetic variation in the Arabidopsis thaliana methylome.
914 *Nature* **480**: 245-249.
915 Bertrand EM, Allen AE, Dupont CL, Norden-Krichmar TM, Bai J, Valas RE, Saito MA.
916 2012. Influence of cobalamin scarcity on diatom molecular physiology and
917 identification of a cobalamin acquisition protein. *Proceedings of the*
918 *National Academy of Sciences of the United States of America* **109**: E1762-
919 1771.
920 Blanc-Mathieu R, Verhelst B, Derelle E, Rombauts S, Bouget FY, Carre I, Chateau A,
921 Eyre-Walker A, Grimsley N, Moreau H et al. 2014. An improved genome of
922 the model marine alga Ostreococcus tauri unfolds by assessing Illumina de
923 novo assemblies. *BMC genomics* **15**: 1103.
924 Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U,
925 Martens C, Maumus F, Otillar RP et al. 2008. The Phaeodactylum genome
926 reveals the evolutionary history of diatom genomes. *Nature* **456**: 239-244.
927 Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C,
928 Stegle O, Lippert C et al. 2011. Whole-genome sequencing of multiple
929 Arabidopsis thaliana populations. *Nature genetics* **43**: 956-963.
930 Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-
931 Mendez G, Rocha F, Tirichine L, Labadie K et al. Under revision. A global
932 ocean atlas of eukaryotic genes. *Nature*.
933 Casteleyn G, Leliaert F, Backeljau T, Debeer AE, Kotaki Y, Rhodes L, Lundholm N,
934 Sabbe K, Vyverman W. 2010. Limits to gene flow in a cosmopolitan marine
935 planktonic diatom. *Proceedings of the National Academy of Sciences of the*
936 *United States of America* **107**: 12952-12957.
937 Cermeno P, Falkowski PG. 2009. Controls on diatom biogeography in the ocean.
938 *Science* **325**: 1539-1541.
939 Cheong WH, Tan YC, Yap SJ, Ng KP. 2015. ClicO FS: an interactive web-based
940 service of Circos. *Bioinformatics* **31**: 3685-3687.
941 Chepurnov VA, Mann DG, Sabbe K, Vyverman W. 2004. Experimental studies on
942 sexual reproduction in diatoms. *Int Rev Cytol* **237**: 91-154.
943 Chepurnov VA, Mann DG, Vyverman W, Sabbe K, Danielidis DB. 2002. Sexual
944 reproduction, mating system, and protoplast dynamics of Seminavis
945 (Bacillariophyceae). *Journal of Phycology* **38**: 1004-1019.
946 Chu JH, Wegmann D, Yeh CF, Lin RC, Yang XJ, Lei FM, Yao CT, Zou FS, Li SH. 2013.
947 Inferring the geographic mode of speciation by contrasting autosomal and
948 sex-linked genetic diversity. *Mol Biol Evol* **30**: 2519-2530.
949 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden
950 DM. 2012. A program for annotating and predicting the effects of single
951 nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila
952 melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80-92.

953 Clark DP, Flynn KJ, Ownes NJ. 2002. The large capacity for dark nitrate-
954 assimilation in diatoms may overcome nitrate limitation of growth. *New*
955 *Phytologist* doi:10.1046/j.1469-8137.2002.00435.x.
956 Crow JF. 2005. Timeline: Hermann Joseph Muller, evolutionist. *Nature reviews*
957 *Genetics* **6**: 941-945.
958 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE,
959 Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and
960 VCFtools. *Bioinformatics* **27**: 2156-2158.
961 Davidovich NA, Bates SS. 1998. Sexual reproduction in the pennate diatoms
962 Pseudo-nitzschia multiseries and P-pseudodelicatissima
963 (Bacillariophyceae). *Journal of Phycology* **34**: 126-137.
964 Davidovich NA, Kaczmarska I, Karpov SA, Davidovich OI, MacGillivary ML, Mather
965 L. 2012. Mechanism of male gamete motility in araphid pennate diatoms
966 from the genus Tabularia (Bacillariophyta). *Protist* **163**: 480-494.
967 De Martino AM, A. Juan Shi, K.P. Bowler, C. 2007. Genetic and phenotypic
968 characterization of Phaeodactylum tricornutum (Bacillariophyceae)
969 accessions. *J Phycol* **43**: 992–1009.
970 De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A. 2009. Gene
971 silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic acids*
972 *research* **37**: e96.
973 de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J. 1999. Molecular evidence
974 of cryptic speciation in planktonic foraminifers and their relation to
975 oceanic provinces. *Proceedings of the National Academy of Sciences of the*
976 *United States of America* **96**: 2864-2868.
977 Diner RE, Bielinski VA, Dupont CL, Allen AE, Weyman PD. 2016. Refinement of the
978 Diatom Episome Maintenance Sequence and Improvement of Conjugation-
979 Based DNA Delivery Methods. *Front Bioeng Biotechnol* **4**: 65.
980 Dorak M. 2014. Basic population genetics.
981 Dorrell RG, Gile G, McCallum G, Meheust R, Bapteste EP, Klinger CM, Brillet-
982 Gueguen L, Freeman KD, Richter DJ, Bowler C. 2017. Chimeric origins of
983 ochrophytes and haptophytes revealed through an ancient plastid
984 proteome. *Elife* **6**.
985 Dupont CL, McCrow JP, Valas R, Moustafa A, Walworth N, Goodenough U, Roth R,
986 Hogle SL, Bai J, Johnson ZI et al. 2015. Genomes and gene expression across
987 light and productivity gradients in eastern subtropical Pacific microbial
988 communities. *ISME J* **9**: 1076-1092.
989 Etherington GJ, Dicks J, Roberts IN. 2005. Recombination Analysis Tool (RAT): a
990 program for the high-throughput detection of recombination.
991 *Bioinformatics* **21**: 278-281.
992 Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C. 1999. Transformation of
993 Nonselectable Reporter Genes in Marine Diatoms. *Mar Biotechnol (NY)* **1**:
994 239-251.
995 Ferreira A, Marguti I, Bechmann I, Jeney V, Chora A, Palha NR, Rebelo S, Henri A,
996 Beuzard Y, Soares MP. 2011. Sickle hemoglobin confers tolerance to
997 Plasmodium infection. *Cell* **145**: 398-409.
998 Finlay BJ. 2002. Global dispersal of free-living microbial eukaryote species. *Science*
999 **296**: 1061-1063.
1000 Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, Nelson DR,
1001 Jijakli K, Abdrabu R, Harris EH et al. 2015. Whole-Genome Resequencing

Reveals Extensive Natural Variation in the Model Green Alga Chlamydomonas reinhardtii. *The Plant cell* **27**: 2353-2369.

Foissner W. 2006. Biogeography and Dispersal of Micro-organisms: A Review Emphasizing Protists. *Acta Protozool* **45**: 111-136.

Fortunato AE, Jaubert M, Enomoto G, Bouly JP, Raniello R, Thaler M, Malviya S, Bernardes JS, Rappaport F, Gentili B et al. 2016. Diatom Phytochromes Reveal the Existence of Far-Red-Light-Based Sensing in the Ocean. *The Plant cell* **28**: 616-628.

Garcia-Etxebarria K, Garcia-Garcera M, Calafell F. 2014. Consistency of metagenomic assignment programs in simulated and real data. *BMC bioinformatics* **15**: 90.

Godhe A, Kremp A, Montresor M. 2014. Genetic and microscopic evidence for sexual reproduction in the centric diatom Skeletonema marinoi. *Protist* **165**: 401-416.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**: 63-66.

Harnstrom K, Ellegaard M, Andersen TJ, Godhe A. 2011. Hundred years of genetic structure in a sediment revived diatom population. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 4252-4257.

Helliwell KE, Collins S, Kazamia E, Purton S, Wheeler GL, Smith AG. 2015. Fundamental shift in vitamin B12 eco-physiology of a model alga demonstrated by experimental evolution. *ISME J* **9**: 1446-1455.

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

Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S, Zeng Q, Zisson E, Wang JM, Greenberg JM et al. 2015. Genetic and phenotypic intra-species variation in Candida albicans. *Genome Res* **25**: 413-425.

Huysman MJ, Fortunato AE, Matthijs M, Costa BS, Vanderhaeghen R, Van den Daele H, Sachse M, Inze D, Bowler C, Kroth PG et al. 2013. AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (Phaeodactylum tricornutum). *The Plant cell* **25**: 215-228.

Huysman MJ, Martens C, Vandepoele K, Gillard J, Rayko E, Heijde M, Bowler C, Inze D, Van de Peer Y, De Veylder L et al. 2010. Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome biology* **11**: R17.

Ivanikova NV, McKay R, Bullerjahn GS. 2005. Construction and characterization of a cyanobacterial bioreporter
capable of assessing nitrate assimilatory capacity in freshwaters. *Limnology and Oceanography* **3**: 86-93.

Kaczmarska I, Mather L, Luddington I, Muise F, Ehrman J. 2014. Cryptic diversity in a cosmopolitan diatom known as Asterionellopsis glacialis (Fragilariaceae): Implications for ecology, biogeography, and taxonomy. *American Journal of Botany*.

Kanai Y, Ushijima S, Tsuda H, Sakamoto M, Hirohashi S. 2000. Aberrant DNA methylation precedes loss of heterozygosity on chromosome 16 in chronic hepatitis and liver cirrhosis. *Cancer Lett* **148**: 73-80.

1051 Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D,
1052 Benzoni F, Claverie JM et al. 2011. A holistic approach to marine eco-
1053 systems biology. *PLoS Biol* **9**: e1001177.
1054 Kaur S, Spillane C. 2015. Reduction in carotenoid levels in the marine diatom
1055 Phaeodactylum tricornutum by artificial microRNAs targeted against the
1056 endogenous phytoene synthase gene. *Mar Biotechnol (NY)* **17**: 1-7.
1057 Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge University
1058 Press.
1059 Kooistra WH, Sarno D, Balzano S, Gu H, Andersen RA, Zingone A. 2008. Global
1060 diversity and biogeography of Skeletonema species (bacillariophyta).
1061 *Protist* **159**: 177-193.
1062 Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics
1063 Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**: 1870-1874.
1064 Lescot M, Hingamp P, Kojima KK, Villar E, Romac S, Veluchamy A, Boccara M,
1065 Jaillon O, Iudicone D, Bowler C et al. 2016. Reverse transcriptase genes are
1066 highly abundant and transcriptionally active in marine plankton
1067 assemblages. *ISME J* **10**: 1134-1146.
1068 Liti G. 2015. The fascinating and secret wild life of the budding yeast S. cerevisiae.
1069 *Elife* **4**.
1070 Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN,
1071 Burt A, Koufopanou V et al. 2009. Population genomics of domestic and
1072 wild yeasts. *Nature* **458**: 337-341.
1073 Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF,
1074 Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.
1075 Maheswari U, Jabbari K, Petit JL, Porcel BM, Allen AE, Cadoret JP, De Martino A,
1076 Heijde M, Kaas R, La Roche J et al. 2010. Digital expression profiling of novel
1077 diatom transcripts provides insight into their biological functions. *Genome*
1078 *biology* **11**: R85.
1079 Maheswari U, Mock T, Armbrust EV, Bowler C. 2009. Update of the Diatom EST
1080 Database: a new tool for digital transcriptomics. *Nucleic acids research* **37**:
1081 D1001-1005.
1082 Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, Wincker P, Iudicone
1083 D, de Vargas C, Bittner L et al. 2016. Insights into global diatom distribution
1084 and diversity in the world's ocean. *Proceedings of the National Academy of*
1085 *Sciences of the United States of America* doi:10.1073/pnas.1509523113.
1086 Matuszewski S, Hermisson J, Kopp M. 2015. Catch Me if You Can: Adaptation from
1087 Standing Genetic Variation to a Moving Phenotypic Optimum. *Genetics* **200**:
1088 1255-1274.
1089 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella
1090 K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit:
1091 a MapReduce framework for analyzing next-generation DNA sequencing
1092 data. *Genome Res* **20**: 1297-1303.
1093 Medlin LK. 2015. A timescale for diatom evolution based on four molecular
1094 markers: reassessment of ghost lineages and major steps defining diatom
1095 evolution. *Vie Milieu / Life & Environment*.
1096 Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, Salamov A,
1097 Sanges R, Toseland A, Ward BJ et al. 2017. Evolutionary genomics of the
1098 cold-adapted diatom Fragilariopsis cylindrus. *Nature* **541**: 536-540.

1099 Morrissey J, Sutak R, Paz-Yepes J, Tanaka A, Moustafa A, Veluchamy A, Thomas Y,
1100 Botebol H, Bouget FY, McQuaid JB et al. 2015. A novel protein, ubiquitous
1101 in marine phytoplankton, concentrates iron at the cell surface and
1102 facilitates uptake. *Current biology : CB* **25**: 364-371.
1103 Mouget J-L, RGastineau R, Davidovich O, Gaudin P, Davidovich NA. 2009. Light is a
1104 key factor in triggering sexual reproduction in the pennate diatom Haslea
1105 ostrearia. *FEMS Microbial Ecology*.
1106 Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D. 2009.
1107 Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science*
1108 **324**: 1724-1726.
1109 Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms
1110 of restriction endonucleases. *Proceedings of the National Academy of*
1111 *Sciences of the United States of America* **76**: 5269-5273.
1112 Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino
1113 acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929-
1114 936.
1115 Nikula R, Spencer HG, Waters JM. 2013. Passive rafting is a powerful driver of
1116 transoceanic gene flow. *Biol Lett* **9**: 20120821.
1117 Nymark M, Sharma AK, Sparstad T, Bones AM, Winge P. 2016. A CRISPR/Cas9
1118 system adapted for gene editing in marine algae. *Sci Rep* **6**: 24951.
1119 Patil S, Moeys S, von Dassow P, Huysman MJ, Mapleson D, De Veylder L, Sanges R,
1120 Vyverman W, Montresor M, Ferrante MI. 2015. Identification of the meiotic
1121 toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51
1122 homologs in the sexual species Pseudo-nitzschia multistriata and
1123 Seminavis robusta. *BMC genomics* **16**: 930.
1124 Pickett-Heaps JD, Forer A. 2001. Pac-Man does not resolve the enduring problem
1125 of anaphase chromosome movement. *Protoplasma* **215**: 16-20.
1126 Piganeau G, Eyre-Walker A, Jancek S, Grimsley N, Moreau H. 2011. How and why
1127 DNA barcodes underestimate the diversity of microbial eukaryotes. *PloS*
1128 *one* **6**: e16342.
1129 Proctor VW. 1966. Dispersal of Desmids by birds. *Phycologia* **5**: 227-232.
1130 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar
1131 P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome
1132 association and population-based linkage analyses. *American journal of*
1133 *human genetics* **81**: 559-575.
1134 Rabosky DL, Sorhannus U. 2009. Diversity dynamics of marine planktonic diatoms
1135 across the Cenozoic. *Nature* **457**: 183-186.
1136 Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S,
1137 Dominiczak A, Morris A, Porteous D, Smith B et al. 2016. Timing, rates and
1138 spectra of human germline mutation. *Nature genetics* **48**: 126-133.
1139 Rastogi A, Maheswari U, Dorrell RG, Maumus F, Kustka A, McCarthy J, Allen AE,
1140 Kersey P, Bowler C, Tirichine L. Submitted. Integrative analysis of large
1141 scale transcriptome data draws a comprehensive landscape of
1142 Phaeodactylum tricornutum genome and evolutionary origin of diatoms.
1143 Rastogi A, Murik O, Bowler C, Tirichine L. 2016. PhytoCRISP-Ex: a web-based and
1144 stand-alone application to find specific target sequences for CRISPR/CAS
1145 editing. *BMC bioinformatics* **17**: 261.

1146 Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G,
1147 Mesirov JP. 2011. Integrative genomics viewer. *Nature biotechnology* **29**:
1148 24-26.
1149 Rottenstreich S, Hamilton MB, Miller JR. 2007. Dynamics of Fst for the island
1150 model. *Theor Popul Biol* **72**: 485-503.
1151 Saez AG, Probert I, Geisen M, Quinn P, Young JR, Medlin LK. 2003. Pseudo-cryptic
1152 speciation in coccolithophores. *Proceedings of the National Academy of*
1153 *Sciences of the United States of America* **100**: 7163-7168.
1154 Schlichting HE. 1960. The rôle of waterfowl in the dispersal of algae. *Trans Am*
1155 *Microsc*
1156 *Soc* **79**: 160-166.
1157 Seibel PN, Muller T, Dandekar T, Schultz J, Wolf M. 2006. 4SALE--a tool for
1158 synchronous RNA sequence and secondary structure alignment and
1159 editing. *BMC bioinformatics* **7**: 498.
1160 Sellis D, Callahan BJ, Petrov DA, Messer PW. 2011. Heterozygote advantage as a
1161 natural consequence of adaptation in diploids. *Proceedings of the National*
1162 *Academy of Sciences of the United States of America* **108**: 20666-20671.
1163 Sellis D, Kvitek DJ, Dunn B, Sherlock G, Petrov DA. 2016. Heterozygote Advantage
1164 Is a Common Outcome of Adaptation in Saccharomyces cerevisiae. *Genetics*
1165 **203**: 1401-1413.
1166 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N,
1167 Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for
1168 integrated models of biomolecular interaction networks. *Genome Res* **13**:
1169 2498-2504.
1170 Shoemaker R, Deng J, Wang W, Zhang K. 2010. Allele-specific methylation is
1171 prevalent and is contributed by CpG-SNPs in the human genome. *Genome*
1172 *Res* **20**: 883-889.
1173 Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A,
1174 Falciatore A, Bowler C. 2007. Molecular toolbox for studying diatom
1175 biology in Phaeodactylum tricornutum. *Gene* **406**: 23-35.
1176 Stanley JAC. 2007. Whole cell adhesion strength of morphotypes and isolates of
1177 Phaeodactylum tricornutum (Bacillariophyceae). *European Journal of*
1178 *Phycology* doi:10.1080/09670260701240863.
1179 Taddei L, Stella GR, Rogato A, Bailleul B, Fortunato AE, Annunziata R, Sanges R,
1180 Thaler M, Lepetit B, Lavaud J et al. 2016. Multisignal control of expression
1181 of the LHCX protein family in the marine diatom Phaeodactylum
1182 tricornutum. *J Exp Bot* **67**: 3939-3951.
1183 Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Marechal E, Bowler C, Muto
1184 M, Sunaga Y, Tanaka M et al. 2015. Oil Accumulation by the Oleaginous
1185 Diatom Fistulifera solaris as Revealed by the Genome and Transcriptome.
1186 *The Plant cell* **27**: 162-176.
1187 Thiel MaG, L. 2005. The ecology of rafting in the marine environment. II. The
1188 rafting organisms and community. *Oceanography and Marine Biology: An*
1189 *Annual Review* **43**: 279-418.
1190 Tirichine L, Rastogi A, Bowler C. 2017. Recent progress in diatom genomics and
1191 epigenomics. *Curr Opin Plant Biol* **36**: 46-55.
1192 Vartanian M, Descles J, Quinet M, Douady S, Lopez PJ. 2009. Plasticity and
1193 robustness of pattern formation in the model diatom Phaeodactylum
1194 tricornutum. *The New phytologist* **182**: 429-442.

1195 Veluchamy A, Lin X, Maumus F, Rivarola M, Bhavsar J, Creasy T, O'Brien K,
1196 Sengamalay NA, Tallon LJ, Smith AD et al. 2013. Insights into the role of
1197 DNA methylation in diatoms by genome-wide profiling in Phaeodactylum
1198 tricornutum. *Nat Commun* **4**.
1199 Veluchamy A, Rastogi A, Lin X, Lombard B, Murik O, Thomas Y, Dingli F, Rivarola
1200 M, Ott S, Liu X et al. 2015. An integrative analysis of post-translational
1201 histone modifications in the marine diatom Phaeodactylum tricornutum.
1202 *Genome biology* **16**: 102.
1203 Wecek K, Hartmann S, Paijmans JL, Taron U, Xenikoudakis G, Cahill JA, Heintzman
1204 PD, Shapiro B, Baryshnikov G, Bunevich AN et al. 2016. Complex Admixture
1205 Preceded and Followed the Extinction of Wisent in the Wild. *Mol Biol Evol*
1206 doi:10.1093/molbev/msw254.
1207 Weiguo Li JW. 2011. Influence of light and nitrate assimilation on the growth
1208 strategy in clonal weed Eichhornia crassipes. *Aquatic Ecology*
1209 doi:10.1007/s10452-010-9318-8.
1210 Whitlock MC, McCauley DE. 1999. Indirect measures of gene flow and migration:
1211 FST not equal to 1/(4Nm + 1). *Heredity (Edinb)* **82 (Pt 2)**: 117-125.
1212 Whittaker KA, Rynearson TA. 2017. Evidence for environmental and ecological
1213 selection in a microbe with no geographic limits to gene flow. *Proceedings*
1214 *of the National Academy of Sciences of the United States of America* **114**:
1215 2651-2656.
1216 Wolf M, Chen S, Song J, Ankenbrand M, Muller T. 2013. Compensatory base
1217 changes in ITS2 secondary structures correlate with the biological species
1218 concept despite intragenomic variability in ITS2 sequences--a proof of
1219 concept. *PloS one* **8**: e66726.
1220 Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for
1221 heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431-
1222 449.
1223 Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. 2006. KaKs_Calculator: calculating
1224 Ka and Ks through model selection and model averaging. *Genomics*
1225 *Proteomics Bioinformatics* **4**: 259-263.
1226