A genomics approach reveals the global genetic polymorphism, structure and functional diversity of ten accessions of the marine model diatom *Phaeodactylum tricornutum*

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

[1]Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Université Paris 75005 Paris, France

[2]Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072 China

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

[4] Sorbonne Universités, UPMC Univ Paris 06, CNRS, Biologie Intégrative des Organismes Marins (BIOM), Observatoire Océanologique, F-66650 Banyuls/Mer, France

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

[€]Current affiliation : Université de Nantes, CNRS, UFIP, UMR 6286, F-44000 Nantes, France

***Corresponding authors**

Name: Leila Tirichine
Address: Université de Nantes, CNRS, UFIP, UMR 6286, F-44000 Nantes, France
Phone: +33-276645058
Email: tirichine-l@univ-nantes.fr

Name: Hanhua Hu
Address: Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072 China
Phone: +86-27-68780078
Email: hanhuahu@ihb.ac.cn

Running title: *Phaeodactylum tricornutum* population genomics

Abstract

Diatoms emerged in the Mesozoic period and presently constitute one of the main primary producers in the world's ocean and are of a major economic importance. In the current study, using whole genome sequencing of ten accessions of the model diatom *Phaeodactylum tricornutum*, sampled at broad geospatial and temporal scales, we draw a comprehensive landscape of the genomic diversity within the species. We describe strong genetic subdivisions of the accessions into four genetic clades (A-D) with constituent populations of each clade possessing a conserved genetic and functional makeup, likely a consequence of the limited dispersal of *P. tricornutum* in the open ocean. We further suggest dominance of asexual reproduction across all the populations, as implied by high linkage disequilibrium. Finally, we show limited yet compelling signatures of genetic and functional convergence 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 and impact on aquatic ecosystems where they play a major role. Our work provides valuable resources for functional genomics and for exploiting the biotechnological potential of this model diatom species.

Introduction

Diatoms are unicellular predominantly diploid and photosynthetic eukaryotes. They belong to the large group known as Ochrophytes (plastid-bearing members of the stramenopiles), constituents of CASH (Cryptomonads, Alveolates, Strameopiles, and Haptophytes) lineages, and are believed to have evolved from serial endosymbiosis involving green and red algal symbionts [1-3]. Ehrenberg first discovered diatoms 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 [4] and their closest sister group are the Bolidomonads. In nature, most diatoms likely live in obligate relationships with bacteria [5] but many, like *Phaeodactylum tricornutum*, can be propagated in axenic conditions. In spite of its low abundance in the open ocean [6], *P. tricornutum* is extensively used as a model to study and characterize diatom metabolism, and to understand diatom evolution [1, 7-11].

P. tricornutum is a coastal diatom found under highly unstable environments like estuaries and rock-pools. Although it has never been reported to undergo sexual reproduction, factors such as small cell size, discontinuous sexual phases, and the sensitivity of sexual reproduction to many nonspecific abiotic components in diatoms [12-14] limit our ability to constrain the sexual cycle of these organisms. Since the discovery of *P. tricornutum* by Bohlin in 1897 and the characterization of different morphologies, denoted fusiform, triradiate, oval, round and cruciform, 10 strains from 9 different geographic locations (sea shores, estuaries, rock pools, tidal creeks) around the world, from sub-polar to tropical latitudes, have been accessioned (Fig S1) [well described in [15]]. These accessions have been collected within the time frame of approximately one century, from 1908 (Plymouth strain, Pt2/3) to 2000 (Dalian strain,

Pt10) (Fig S1) [15]. 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 accessions as adaptive responses to various environmental cues [16-19], but very little is known about their genetic diversity. However, based on sequence similarity of the ITS2 region within the 28S rDNA repeat sequence, the accessions 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 [15]. *P. tricornutum* is among the few diatom species with a whole genome sequence available to the community [20], and the only diatom for which extensive state-of-the-art functional and molecular tools have been developed over the past few decades [21-34]. These resources have advanced *P. tricornutum* as a model diatom species and provided a firm platform for future genome-wide structural and functional studies.

The accumulated effects of diverse evolutionary forces such as recombination, mutation, and selection have been found to dictate the structure and diversity of genomes in a wide range of species [35-38]. The existence of genomic diversity within a species reflects its potential to adapt to a changing environment. Exploring the genomic diversity within a species not only provides information about its evolution, it also offers opportunities to understand the role of various biotic and abiotic interactions in structuring a genome [39]. Such studies in diatoms are rare and estimates of genetic diversity within diatom populations are mostly inferred using microsatellite-based genotyping approaches [40-42]. Although these techniques have revealed a wealth of information about diatom evolution, their dispersal and reproductive physiology [39], additional insights can be obtained using state-of-the-

111 art whole genome comparative analysis techniques [42]. Deciphering the standing
112 genomic variation of *P. tricornutum* across different accession populations, sampled
113 at broad geospatial scale, is an important first step to assess the role of various
114 evolutionary forces in regulating the adaptive capacities of diatoms in general
115 (e.g.[43]). To understand the underlying genomic diversity within different accessions
116 of *P. tricornutum* and to establish the functional implications of such diversity, we
117 performed deep whole genome sequencing of the 10 most studied accessions, referred
118 to as Pt1 to Pt10 [15, 18, 44]. We present a genome-wide diversity map of
119 geographically distant *P. tricornutum* accessions, describing a stable genetic structure
120 in the environment. This work further provides the community with whole genome
121 sequences of the accessions, which will be a valuable genetic resource for functional
122 studies of accession-specific ecological traits in the future.
123
124 **Results**
125
126 **Reference-assisted assembly reveals low nucleotide diversity across multiple**
127 **accessions of *P. tricornutum***
128 We sequenced the whole genomes of ten accessions of *P. tricornutum* using Illumina
129 HiSeq 2000, and performed a reference-based assembly using the genome sequence
130 of the reference strain Pt1 8.6 [1]. Across all accessions, the percentage of sequence
131 reads mapped on the reference genome ranged between ~65% to ~80% (Table 1),
132 with an alignment depth ranging between 26X and 162X, covering 92% to 98% of the
133 reference genome (Table 1). Many regions on the reference genome that are observed
134 as being unmapped by reads from individual ecotypes are annotated as being rich in

transposable elements (TEs) (Fig S2). At >90% identity the repeated proportion of unmapped reads varies between ~38% (Pt1) and 75% (Pt4).

Following the assembly, we performed variant calling using Genome Analysis Toolkit [45] and discovered 462,514 (depth >= 4x) single nucleotide polymorphisms (SNPs) including ~25% singleton sites, 573 insertions (of varying lengths from 1 bp to 312 bp) and 1,801 deletions (of lengths from 1 bp to 400 bp) (Fig 1A), across all the accessions. The spectrum of SNPs across all the accessions further reveals a higher rate of transitions (Ts) over transversions (Tv) (Ts/Tv = 1.6). In total, compared to the reference alleles from Pt1.8.6, 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 S3A). Further, most SNPs and INDELs (insertions and deletions) are shared between different accessions, except for Pt4, which possesses the highest proportion of specific SNPs (~35%) and INDELs (~75%) (Fig 1B). Interestingly, we found that most of the SNPs are heterozygous, and the proportion of heterozygous variants across all the accessions varies between ~45% (in Pt5 and Pt10) to ~98% (in Pt1, Pt2 and Pt3) (Fig 1C). Most of the variant alleles in the accessions with high proportions of heterozygous variants were further found to be significantly deviated from Hardy-Weinberg equilibrium (HWE) (chi-square test, P-value < 0.05) (Fig 1C), possibly linked to prolonged asexual reproduction [46]. Surprisingly, despite significant differences in the proportion of heterozygote variant alleles between the accessions, which ranges between 45% to 98%, the average pairwise synonymous nucleotide diversity (π_S) estimated from genes with callable sites across all the accessions is 0.007 per synonymous site. This indicates that any two homologous sequences taken at random across different populations will on average differ by only ~0.7% on synonymous positions. The non-synonymous

pairwise diversity (π_N) over the same genes is 0.003, consistent with an excess of non-synonymous mutations being deleterious. Linkage disequilibrium (LD) analysis using only homozygous SNP sites revealed, on average, high LD (>0.7) over pairs of variations, genome wide (Fig S3B). Further, based on the difference in the allelic frequencies of the SNPs, the pairwise *Fst* between the populations ranges from ~0.005 (between Pt1 and Pt3) to ~0.4 (between Pt4 and Pt10) (Fig 1D). Considering *Fst* as a measure of genetic differentiation or structuring between the populations, the ten *P. tricornutum* accessions can be clustered into 4 genetic groups/Clades with Pt1, Pt2, Pt3 and Pt9 in Clade A; Pt4 in Clade B; Pt5, Pt10 in Clade C; and Pt6, Pt7, Pt8 in Clade D, reflecting low intra-group Fst (~0.02) and high inter-group Fst (0.2 – 0.4) (Fig 1D).

Four genetic Clades of *P. tricornutum*

With the exception of Pt4, where we found the maximum number of variant alleles to be accession-specific, most of the variant alleles are shared between at least two accessions, indicating close genetic relatedness (Fig 1B). Therefore, in order to cluster the accessions based on the genome structure shared among them, we used Bayesian clustering approach by applying *Markov Chain Monte Carlo* (MCMC) estimations, programmed within the ADMIXTURE software [47]. As a result, population clustering of the ten accessions revealed four genetic clusters with Pt1, Pt2, Pt3, and Pt9 in one, Pt4 in a second, Pt5, Pt10 in a third, and Pt6, Pt7, Pt8 in a fourth cluster (Fig. 2A). Each cluster has a distinctive genetic makeup, which is also shared among different accessions in different proportions (represented with different colors in Fig. 2B). These four genetic clusters (Fig 2A) are also in broad agreement with *Fst*-based genetic Clades (Fig 1D), phylogenetic clusters inferred using ribosomal marker genes

(18S (Fig S4A), and ITS2 (Fig S4B)), as also reported previously [15], and at whole genome scale (this study) as inferred by a phylogenetic tree generated using maximum likelihood algorithm based on all (Fig 2C) and only homozygous polymorphic sites (SNVs and INDELS) (Fig S4C). Based on the consensus from different clustering algorithms, the four phylogenetic Clades were defined as Clade A (Pt1, Pt2, Pt3 and Pt9), Clade B (Pt4), Clade C (Pt5 and Pt10), and Clade D (Pt6, Pt7, and Pt8).

Further sequential assessment of the 18S and ITS2 rDNA gene sequences across different clades indicated the presence of multiple variations, including both heterozygous and homozygous variant alleles (Fig S4D and S4E). Because the ribosomal DNA region including 18S and ITS2 is highly repetitive, which is on average ~4 times more than non-ribosomal genes (Fig 3A), these differences can be understood as intra-genomic variations within the genome. However, taxonomists and ecologists use differences within 18S gene sequences as a measure of species assignation and to estimate species delineation [6]. This latter practice has been previously shown to be very conservative as no differences in the 18S gene were found between reproductively isolated species [48]. Alternatively, the possibility of sub-populations or cryptic populations cannot be ignored, as previously reported in planktonic foraminifers [49] and coccolithophores [50].

We examined the possible presence of sub-populations on 18S gene heterozygosity in the Pt8 accession. In particular, 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 (propagated from a single cell) from Pt8 population (constituent of Clade D), referred to as Pt8Tc (Fig S4D), indicating that the cultures are a single population.

Next, concerning the observed polymorphisms within the 18S ribosomal marker gene, we investigated whether the four clades can be considered as different species. We looked for the existence of compensatory base changes (CBCs) within secondary structures of the ITS2 gene between all pairs of accessions. The presence of CBCs within ITS2 has been recently suggested to account for reproductive isolation in multiple plant species [51] and between diatom species [52, 53]. By comparing the ITS2 secondary structure from all the accessions, we did not find any CBCs between any given pair of accessions (Fig S5). As a control, we compared the ITS2 secondary structure of all the *P. tricornutum* accessions with the ITS2 sequences of other diatom species (*Cyclotella meneghiniana*, *Pseudo-nitzschia delicatissima*, *Pseudo-nitzschia multiseries*, *Fragilariopsis cylindrus*) that have significant degrees of evolutionary divergence as depicted previously using multiple molecular marker genes [20, 54], and found multiple CBCs in them (Fig S5).

Close genetic relatedness depicted by large structural genomic variations among accessions

Next, 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, respectively, have been lost or exhibit copy number variation (CNV), across the ten accessions (Fig 3A, 3B) (File S1). Multiple randomly chosen loci were also validated by PCR for their loss from certain accessions compared to the reference strain Pt1 8.6 (Fig S6). Compared to the reference, approximately 70% of the genes that are either lost or show CNV are shared among multiple accessions with an exception of Pt10, which displays the maximum number of lost genes and accession-specific genes exhibiting CNV (Fig 3B). In addition, we detected 207 TEs (~6% of the total

annotated TEs) (File S2) showing CNVs across one or more accessions (Fig 3C, 3D), 80% of which are shared among two or more accessions, with Pt10 again possessing the maximum number of accession-specific TEs exhibiting CNVs (Fig 3C). Not surprisingly, across all the accessions, 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 (Fig 3D, S7) that are transposed by a cut-and-paste mechanism. Euclidean distance estimated between accessions, based on the variation in the number of copies of different genes and TEs displaying CNVs, followed by hierarchical clustering, depicted three genetic clusters: Pt1, Pt2, Pt3, Pt9 in cluster1; Pt5, Pt10 in cluster 2, and Pt4, Pt6, Pt7, Pt8 in cluster 3 (Fig 3A, 3D). These clusters are in broad agreement with the ones described by *Fst*, and indicate the closer genetic makeup between accessions within a cluster than between the clusters. Further, biological processes can only be traced for ~40% of the genes exhibiting accession-specific CNVs. Among all the enriched biological processes (chi-square test, $P<0.01$) (File S1), a gene associated to nitrate assimilation (Phatr3_EG02286) is observed to have higher copy number specifically in Pt4. Likewise, each accession can be characterized by specific genetic features, represented by ~0.3% to ~28% accession-specific CNVs (Fig 3B), possibly linked to the explicit functional behavior of some accessions in response to various environmental cues, as reported previously [16-18].

Functional characterization of the genetic diversity within *P. tricornutum* clades

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 accessions. Localization of the polymorphic sites over genomic features (genes, TEs, and intergenic regions) revealed highest rate of variation within

genes (Fig 2C), specifically on exons, and was consistent across all the studied accessions. An average non-synonymous to synonymous variant ratio (π_N / π_S) was estimated to be ~0.43, which is higher than in the *Ostreococcus tauri*, π_N / π_S = 0.2[55]. We further identified genes within different phylogenetic clades experiencing different selection pressure based on lowest and highest π_N / π_S ratios. Across all the accessions, 241 genes displaying π_N / π_S >1 and a higher frequency of non-synonymous as compared to synonymous polymorphism, as expected under balancing selection [56](File S3). Furthermore, many genes (902) were found to have loss-of-function (LoF hereafter) variant alleles (Fig 4A), including frame-shift mutations and mutations leading to theoretical start/stop codon loss and/or gain.

Based on the presence of functional domains (Pfam domains), all *P. tricornutum* annotated genes [57] were grouped into 3,020 gene families. These families can be as large as the reverse transcriptase gene family, which is highly abundant in marine plankton [58], representing 149 candidate genes having reverse transcriptase domains, or as small as families that constitute single gene candidates. Across all the accessions, 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 [59]. Therefore, to estimate an unbiased effect of any evolutionary pressure (LoF allele mutations) on different gene families, we calculated a ratio, termed the effect ratio (EfR, see Materials and Methods), which normalizes 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 genetic clade displayed a specific set of gene

families as being under selection (Fig 5). Functional characterization of constrained genes revealed genes enriched in two families, (1) AAA family proteins that often perform chaperone like functions that assist in the assembly or disassembly of proteins complexes, protein transport and degradation as well as other functions such as replication, recombination, repair and transcription [60], (2) tetratricopeptide-like repeats known for their role in a variety of biological processes, such as cell cycle regulation, organelle targeting and protein import, vesicle fusion and biomineralization [61]. A redox class of enzymes are common to both groups of genes and a significant proportion of unknown function proteins is found in the group of genes under balancing selection (File S3).

Consistent with the population structure, accessions within individual clades are more closely related than the accessions belonging to other clades (Fig S8A and S8B), suggesting variation in functional relatedness between different proposed phylogenetic clades.

Selection of *MetH* facilitated methionine biosynthesis over *MetE*

Apart from the genetic clade-specific families that are under selection pressure, across all the accessions a group of gene families associated with methionine biosynthesis *(MetH*, Phatr3_J23399) was also observed as experiencing higher π_N / π_S ratio (Fig 5). In *P. tricornutum*, *MetE* (cobalamin-independent methionine synthase) and *MetH* (cobalamin-dependent methionine synthase) are known to catalyze conversion of homocysteine to methionine in the presence of symbiotic bacteria and vitamin B12 in the growth media, respectively. Previous reports have suggested that growing axenic cultures in conditions of high cobalamin (vitamin B12) availability results in repression of MetE, leading to its loss of function and high expression of the MetH gene in *P. tricornutum* and *C. reinhardtii* [62-64]. In accordance with these results,

310 we observed a high expression of *MetH* in axenically grown laboratory cultures (Fig
311 6A) compared to its expression in cells cultured with their natural co-habitant
312 bacteria. However, we were not able to trace any significant signature for the loss of
313 *MetE* gene although its expression is significantly lower in axenic cobalamin-
314 containing cultures (Fig 6B). Similar observations were obtained for *CBA1* and *SHMT*
315 genes (Fig 6C and 6D), which under cobalamin scarcity enhance cobalamin
316 acquisition and manage reduced methionine synthase activity, respectively [63].
317

318 **Discussion**

319 Using whole genome sequence analysis of accessions sampled across multiple
320 geographic locations around the world (Fig S1), the aim of this study was to describe
321 the global genetic and functional diversity of the model diatom *Phaeodactylum*
322 *tricornutum*. By defining a comprehensive landscape of natural variations across
323 multiple accessions, we could investigate genetic structure between *P. tricornutum*
324 populations, and a summary of our results is presented in Fig 7. In order to do so, we
325 first performed reference-based assembly and found consistently high genome
326 coverage (>90%) mapped by sequencing reads from respective accessions, where
327 some accessions have more coverage (>98%, Pt1, Pt2, Pt3, and Pt9) than others
328 (Table 1). This difference is independent of the size of the sequencing library, as it
329 does not correlate with the genome coverage (Table 1), and a portion of unmapped
330 reads is likely a consequence of the incomplete reference genome, which contains
331 several gaps [1]. Additionally, given the redundant nature of unmapped reads together
332 with the fact that the unmapped reference genome is annotated as being rich in TEs
333 (Fig S2), a major portion of unmapped reads likely account for large structural
334 variability within the genomes of individual accessions. This explanation is most clear

in Pt10, which is shown to have the largest number of gene losses (Fig 3B) and the highest number of accession-specific TEs with high copy numbers (Fig 3C), and covers the least (92%) of the reference genome (Table 1). This suggests the role of TEs in creating substantial genetic diversity as also shown in many species of plants and animals [65, 66].

Next, based on patterns of variations discovered over the whole genomes and on the molecular marker genes (18S and ITS2) of all the accessions, and by using various clustering algorithms (see Results), the ten accessions could be grouped into four genetic clades. Clade A clusters Pt1, Pt2, Pt3, Pt9; Clade B includes Pt4; Clade C clusters Pt5, Pt10; and Clade D clusters Pt6, Pt7, Pt8. Most of the structural variants discovered, both small (SNPs and INDELS) and large (CNV and Gene Loss), are shared among populations within a clade rather than between clades. This suggests high intra-clade relatedness over a variety of structural, functional and possibly ecological traits.

P. tricornutum is a coastal species with limited dispersal potential, which is consistent with the reports of its absence in the open ocean from *Tara* Oceans data [6]. Consequently, the Fixation index (*Fst*) between different genetic clades is very high (0.2 – 0.4), (Fig 1D) confirming the existence of strong population subdivisions into four genetic clades. As expected for an organism with limited dispersal potential, most of the populations are structured geographically (Fig 2A), with the exception of Pt5 and Pt9. In addition, the fact that the subdivisions do not correlate with the sampling time (Fig. 7, Fig S1), which spans approximately a century, suggests long and stable genetic populations, which is in line with reports from other diatom species [40, 41]. Although there exist strong genetic structuring within the accessions, the average nucleotide diversity (π), estimated across all the accessions, is remarkably

low (0.2%) compared to the diversity estimates in other unicellular eukaryotes [35, 55, 67-69] but in line with previous estimations in marine phytoplanktonic eukaryotes [70]. The high linkage disequilibrium (>0.7) observed across all the accessions (Fig S3B) can be explained by prolonged asexual reproduction [71], a common behavior among diatoms [72].

Given the observation that there exist a large proportion of heterozygous variant alleles (Fig 1C), the high *Fst* between the clades, and the low nucleotide diversity across the accessions, we propose that allele frequency plays a significant role in the genetic differentiation of the clades. The difference in allele frequencies is possibly linked to adaptive selection. This phenomenon has recently been studied in diatoms where allele-specific expression of numerous loci has been demonstrated to be a significant source of adaptive evolution in the cold-adapted diatom species *Fragilariopsis cylindrus* [73]. Furthermore, high proportions of heterozygous variant alleles in some Clades (Clade A, 98%, Fig 7) compared to others (Clade B, 45%, Fig 7) suggests a high selection pressure in the Clade B accession Pt4, which is further supported by the strong signals of balancing selection in Pt4 (Fig 4A, 7). Conversely, the large number of alleles that are deviated from HWE within Clade A member populations (Fig 7) is most likely due to prolonged asexual reproduction, which is also associated with generating high linkage disequilibrium across all the isolates [71]. Asexual reproduction results in higher proportions of divergent alleles within loci with less genetic variation among individuals and a significant deviation from HWE [71]. Therefore, it is possible that Pt4 undergoes sexual reproduction, reasonably often, as it possesses the least number of heterozygous variant alleles, most of which follow HWE. Besides, recent reports suggest induction of sex in diatoms under low nutrient [13] and low light [14] conditions, which resembles the

385 natural niche of Pt4 [17]. Therefore, it would be interesting to explore Pt4 as a model
386 for investigating sexual phases in *P. tricornutum*.
387 Interestingly, despite high variability in the levels of heterozygosity between different
388 accessions (Fig 7), the mutational spectrum, compared to the reference, and across all
389 the accessions consisted of high G:C -> A:T and A:T -> G:C transitions (Fig S3A).
390 Deamination of cytosines dominantly dictates C to T transitions in both plants and
391 animals [74, 75], and CpG methylation potential of the genome is greatly influenced
392 by heterozygous SNPs in CpG dinucleotides [76]. Previous studies have demonstrated
393 low DNA methylation in *P. tricornutum,* using Pt1 8.6, a monoclonal strain
394 accessioned from a Pt1 single cell as a reference [31, 77]. Because there exist
395 significant differences in the proportion of heterozygote variant alleles between the
396 accessions (45% - 98%), testing for DNA methylation patterns across different
397 accessions may provide an interesting opportunity to dissect cross-talk between loss
398 of heterozygosity and DNA methylation in the selection of certain traits [78].
399 High *Fst*, and yet low nucleotide diversity across all the accessions, suggests some
400 degree of genetic and functional convergence among the accessions. This can be
401 explained as a consequence of maintaining all the accessions in lab cultures. The
402 hypothesis is supported by our observation that the *MetH* gene is under balancing
403 selection in all the accessions (Fig. 5), due to the functional selection of *MetH*-
404 dependent biosynthesis of methionine over *MetE* in the presence of high Vitamin B12
405 in lab-grown cultures [62, 63]. However, such genetic and functional convergence is
406 limited to certain gene families and pathways, as each clade possesses numerous
407 clade-specific gene-families that are under balancing selection (Fig 7), possibly linked
408 to local adaptive traits (Fig 5).

It is also worth considering that genetic homogenization, i.e., low nucleotide diversity, or high linkage-disequilibrium, across the meta-population, can also be a consequence of continuous gene flow between the accessions. However, in the case of *P. tricornutum,* gene flow seems limited as highly differentiated populations structure geographically, except Pt5 of Clade C and Pt9 of Clade A (Fig. 2A). In addition, *P. tricornutum* is not known to reproduce sexually, although various components (genes) of the meiosis pathway are conserved in *P. tricornutum* as well as in other diatom species known to undergo sexual reproduction [79]. Furthermore, the absence of contemporary base changes (CBC) within ITS2 secondary structure between all the accessions compared to the presence of many CBCs between *P. tricornutum* accessions and other diatom species suggests that the accessions may be able to exchange genetic material sexually. However, because *P. tricornutum* is a coastal diatom with only limited dispersal capacity, which is further supported by its apparent absence in the open ocean [6], the possibility of gene flow within different populations is likely to be limited at best.

The four genetic clades are further supported by functional specialization of grouped populations, nicely illustrated with Pt4 in Clade B. Pt4 shows a low non-photochemical quenching capacity (NPQ) [17], which has been proposed to be an adaptive trait to low light conditions. Specifically, this accession has been proposed to establish an up-regulation of a peculiar light harvesting protein LHCX4 in extended dark conditions [17, 19]. In line with these observations, a gene involved in nitrate assimilation (Phatr3_EG02286) in Pt4 shows high copy numbers, suggesting an altered mode of nutrient acquisition. Nitrate assimilation was shown to be regulated extensively under low light or dark conditions to overcome nitrate limitation of growth in *Thalassiosira weissflogii* [80]. Pt4 is likely adapted to the low light and

highly seasonal environment that characterizes the high latitudes where it was found, which may well affect its nitrate assimilation capacity [81, 82]. Additional functions emerging from Clade C (Pt5 and Pt10) include vacuolar sorting and vesicle-mediated transport gene-families to be under balancing selection, which could be an indication of altered intracellular trafficking [83].

In conclusion, the study presents pan-genomic diversity of the model diatom *P. tricornutum*. This is the first study within diatoms that provides a comprehensive landscape of diversity at whole genome sequence level and brings new insights to our understanding of diatom functional ecology and evolution. Given our observation that *P. triconrutum* accessions possess high numbers of heterozygous alleles, it would be interesting to think of possible selective functional preferences of one allele over the other under different environmental conditions or during the life/cell cycle. In the future, such studies could be crucial for deciphering the mechanisms underpinning allele divergence and selection within diatoms. Likewise, more than answers, our study delivers more questions, which should help address the genetic basis of diatom success in diverse ocean ecosystems. Finally, this study provides the community with genomic sequences of *P. tricornutum* accessions that can be useful for functional studies.

Experimental procedures

Sample preparation, sequencing and mapping

Ten different accessions of *P. tricornutum* were obtained from the culture collections of the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP, Pt1=CCMP632, Pt5=CCMP630, Pt6=CCMP631, Pt7=CCMP1327,

Pt9=CCMP633), the Culture Collection of Algae and Protozoa (CCAP, Pt2=CCAP 1052/1A, Pt3= CCAP 1052/1B, Pt4= CCAP 1052/6), the Canadian Center for the Culture of Microorganisms (CCCM, Pt8=NEPCC 640), and the Microalgae Culture Collection of Qingdao University (MACC, Pt10=MACC B228). All of the accessions were grown axenically in batch cultures with a photon fluency rate of 75 μmol photons m-2 s-1 provided by cool-white fluorescent tubes in a 12:12 light: dark (L:D) photoperiod at 20 °C. Exponentially growing cells were harvested and total DNA was extracted with the cetyltrimethylammonium bromide (CTAB) method. At least 6 μg of genomic DNA from each accession was used to construct a sequencing library following the manufacturer's instructions (Illumina Inc.). Paired-end sequencing libraries with a read size of 100 bp and an insert size of approximately 400 bp were sequenced on an Illumina HiSeq 2000 sequencer at Berry Genomics Company (China). The corresponding data can be accessed using bioSample accessions: SAMN08369620, SAMN08369621, SAMN08369622, SAMN08369623, SAMN08369624, SAMN08369625, SAMN08369626, SAMN08369627, SAMN08369628, SAMN08369629. Low quality read-pairs were discarded using FASTQC with a read quality (Phred score) cutoff of 30. Using the genome assembly published in 2008 as reference [1], we performed reference-assisted assembly of all the accessions. 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 S5.

Discovery of small polymorphisms and large structural variants

GATK [45], configured for diploid genomes, was used for variant calling, which included single nucleotide polymorphisms (SNPs), 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). Following this filtration step, the number of sites in the protein coding genes covered for all 10 accessions, and therefore callable to estimate the genome wide synonymous and non-synonymous polymorphism, added up 11.0 Mbp. The average pairwise synonymous and non-synonymous diversity π_S and π_N [84] were estimated for all genes using in R house scripts.

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 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. 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 accession sequencing library failed to map on the reference genome were considered as potentially lost within that accession and reported. Detailed method is provided in File S5. Later, some randomly chosen loci were picked and validated for the loss in the accessions compared to the reference genome by PCR analysis.

Validation of gene loss and quantitative PCR analysis

In order to validate gene loss, DNA was extracted from all the accessions as described previously [21] 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 [22] from accessions grown axenically in Artificial Sea Water (ASW) [85] supplemented with vitamins as well as in the presence of their endemic bacteria in ASW without vitamins. qPCR was performed as described previously [22].

P. tricornutum population structure

***Haplotype analysis*:** First, to cluster the accessions as haplogroups, ITS2 gene (chr13: 42150-43145) and 18S gene (chr13: 43553-45338) were used. Polymorphic sites across all the accessions within ITS2 and 18S genes were called and used to generate their corresponding accession 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 [86] to align and deduce the phylogenetic trees.

***CBC analysis*:** CBC analysis was done by generating the secondary structure of ITS2 sequences, using RNAfold [87], across all *P. tricornutum* accessions 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 secondary structures of ITS2

gene with lowest minimum free energy were used for CBC analysis. We used 4SALE [88] for estimating the presence of CBCs between the secondary 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 accession populations. Within individual accessions, 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 accessions using expectation-maximization (EM) algorithm deployed in the VCFtools. Although no recombination was observed within the accessions, attempts were made to look for recombination signals using LDhat [89] and RAT [90]. Genetic differentiation or variability between the accessions was further assessed using the mathematical function of Fixation index (FST), as described by Weir and Cockerham 1984 [91].

***Genetic clustering*:** Genetic clustering of the accessions was done using Bayesian clustering approach by applying *Markov Chain Monte Carlo* (MCMC) estimation programmed within ADMIXTURE (version linux-1.3.0) [47]. Accessory tools like

PLINK (version 1.07-x86_64) [92] and VCFtools (version 0.1.13) [93] were used to format the VCF files to ADMIXTURE accepted formats. In the absence of data from individuals of each accession/sample, we assumed the behavior of each individual in a sample to be coherent. Conclusively, instead of estimating the genetic structure within an accession, we compared it across all the accessions. We first estimated the possible clusters of genomes, (K), across all the accessions, by using cross-validation error (CV error) function of ADMIXTURE [94]. Finally, we used ADMIXTURE with 200 bootstraps, to estimate the genome clusters within individual accessions by considering the possible number of genomes derived via CV-error function.

Functional characterization of polymorphisms

snpEff [95] and KaKs [96] 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. piN/piS ratios were calculated for 5232 protein coding genes containing more than 10 SNP. 10% of genes with lower piN/piS were considered as under strong purifying selection on amino-acid composition (File S3). Genes with π_N/π_S higher than 1 and average frequency on non-synonymous polymorphism higher than the average frequency of synonymous polymorphism were considered as candidate genes under balancing selection on amino-acid composition (File S3). 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 [97], CYTOSCAPE [98], IGV [99] and R (https://www.r-project.org/about.html). Based on the presence of functional domains all the Phatr3 genes [57] were grouped into 3,020 gene families. Subsequently, the

583 constituents of each gene family were checked for being either affected by loss-of-
584 function mutations or experiencing balancing selection. To estimate an unbiased
585 effect of any evolutionary pressure (LoF allele or balancing selection mutations) on
586 different gene families, induced because of high functional redundancies in the gene
587 families, a normalized ratio named as effect ratio (EfR), was calculated. Precisely, the
588 EfR normalizes the fact that if any gene family have enough candidates to buffer the
589 effect on some genes influencing evolutionary pressures, it will be considered as less
590 affected compared to the situation where all or most of the constituents are under
591 selection pressure. The ratio was estimated as shown below and gene families with
592 EfR larger than 1 were considered as being significantly affected.

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

593

594 Additionally, significantly enriched (chi-square test, P-value < 0.05) biological
595 processes associated within genes experiencing LoF mutations, purifying selection,
596 balancing selection (BS), or showing CNV, or being lost (GnL), were estimated by
597 calculating observed to expected ratio of their percent occurrence within the given
598 functional set (BS, LoF, CNV) and their occurrence in the complete annotated Phatr3
599 (http://protists.ensembl.org/Phaeodactylum_tricornutum /Info/Index) biological
600 process catalog. Later, considering gene family EfR as a function to measure the
601 association rate, we deduced Pearson pairwise correlations between different
602 accessions. The correlation matrix describes that if many equally affected gene
603 families are shared between any given pair of accessions, they will have higher

correlation compared to others. Finally, hierarchical clustering using Pearson pairwise correlation matrix assessed the association between the accessions.

Acknowledgements

HH acknowledges support from National Natural Science Foundation of China (grant No. 91751117). GW acknowledges the Strategic Priority Research Program of the Chinese Academy of Sciences (grant No. XDA17010502). 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. LT acknowledges funds from the CNRS and MEMO LIFE (ANR-10-LABX-54). 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

1. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, et al: **The Phaeodactylum genome reveals the evolutionary history of diatom genomes.** *Nature* 2008, **456:**239-244.
2. Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D: **Genomic footprints of a cryptic plastid endosymbiosis in diatoms.** *Science* 2009, **324:**1724-1726.

3. Dorrell RG, Gile G, McCallum G, Meheust R, Bapteste EP, Klinger CM, et al: **Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome.** *Elife* 2017, **6**.
4. Armbrust EV: **The life of diatoms in the world's oceans.** *Nature* 2009, **459:**185-192.
5. Amin SA, Parker MS, Armbrust EV: **Interactions between diatoms and bacteria.** *Microbiol Mol Biol Rev* 2012, **76:**667-684.
6. Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, Wincker P, Iudicone D, de Vargas C, Bittner L, et al: **Insights into global diatom distribution and diversity in the world's ocean.** *Proc Natl Acad Sci U S A* 2016.
7. Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, McCrow JP, et al: **Evolution and metabolic significance of the urea cycle in photosynthetic diatoms.** *Nature* 2011, **473:**203-207.
8. Huysman MJ, Fortunato AE, Matthijs M, Costa BS, Vanderhaeghen R, Van den Daele H, et al: **AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (Phaeodactylum tricornutum).** *Plant Cell* 2013, **25:**215-228.
9. Morrissey J, Sutak R, Paz-Yepes J, Tanaka A, Moustafa A, Veluchamy A, et al: **A novel protein, ubiquitous in marine phytoplankton, concentrates iron at the cell surface and facilitates uptake.** *Curr Biol* 2015, **25:**364-371.
10. Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Marechal E, et al: **Oil accumulation by the oleaginous diatom Fistulifera solaris as revealed by the genome and transcriptome.** *Plant Cell* 2015, **27:**162-176.
11. Fortunato AE, Jaubert M, Enomoto G, Bouly JP, Raniello R, Thaler et al: **Diatom phytochromes reveal the existence of far-red-light-based sensing in the ocean.** *Plant Cell* 2016, **28:**616-628.
12. Godhe A, Kremp A, Montresor M: **Genetic and microscopic evidence for sexual reproduction in the centric diatom Skeletonema marinoi.** *Protist* 2014, **165:**401-416.
13. Moore ER, Bullington BS, Weisberg AJ, Jiang Y, Chang J, Halsey KH: **Morphological and transcriptomic evidence for ammonium induction of sexual reproduction in Thalassiosira pseudonana and other centric diatoms.** *PLoS One* 2017, **12:**e0181098.
14. Mouget JL, Gastineau R, Davidovich O, Gaudin P, Davidovich NA: **Light is a key factor in triggering sexual reproduction in the pennate diatom Haslea ostrearia.** *FEMS Microbiol Ecol* 2009, **69:**194-201.
15. De Martino AM, A. Juan Shi, K.P. Bowler, C.: **Genetic and phenotypic characterization of Phaeodactylum tricornutum (Bacillariophyceae) accessions.** *J Phycol* 2007, **43:**992–1009.
16. Stanley MS, Callow JA: **Whole cell adhesion strength of morphotypes and isolates of Phaeodactylum tricornutum (Bacillariophyceae).** *European Journal of Phycology* 2007, **42:**191-197.
17. Bailleul B, Rogato A, de Martino A, Coesel S, Cardol P, Bowler C, et al: **An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light.** *Proc Natl Acad Sci U S A* 2010, **107:**18214-18219.

18. Abida H, Dolch LJ, Mei C, Villanova V, Conte M, Block MA, et al: **Membrane glycerolipid remodeling triggered by nitrogen and phosphorus starvation in Phaeodactylum tricornutum.** *Plant Physiol* 2015, **167:**118-136.
19. Taddei L, Stella GR, Rogato A, Bailleul B, Fortunato AE, Annunziata R, et al: **Multisignal control of expression of the LHCX protein family in the marine diatom Phaeodactylum tricornutum.** *J Exp Bot* 2016, **67:**3939-3951.
20. Tirichine L, Rastogi A, Bowler C: **Recent progress in diatom genomics and epigenomics.** *Curr Opin Plant Biol* 2017, **36:**46-55.
21. Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C: **Transformation of nonselectable reporter genes in marine diatoms.** *Mar Biotechnol (NY)* 1999, **1:**239-251.
22. Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, et al: **Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum.** *Gene* 2007, **406:**23-35.
23. De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A: **Gene silencing in the marine diatom Phaeodactylum tricornutum.** *Nucleic Acids Res* 2009, **37:**e96.
24. Huysman MJ, Martens C, Vandepoele K, Gillard J, Rayko E, Heijde M, et al: **Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling.** *Genome Biol* 2010, **11:**R17.
25. Maheswari U, Jabbari K, Petit JL, Porcel BM, Allen AE, Cadoret JP, et al: **Digital expression profiling of novel diatom transcripts provides insight into their biological functions.** *Genome Biol* 2010, **11:**R85.
26. Maheswari U, Mock T, Armbrust EV, Bowler C: **Update of the Diatom EST Database: a new tool for digital transcriptomics.** *Nucleic Acids Res* 2009, **37:**D1001-1005.
27. Kaur S, Spillane C: **Reduction in carotenoid levels in the marine diatom Phaeodactylum tricornutum by artificial microRNAs targeted against the endogenous phytoene synthase gene.** *Mar Biotechnol (NY)* 2015, **17:**1-7.
28. Diner RE, Bielinski VA, Dupont CL, Allen AE, Weyman PD: **Refinement of the Diatom Episome Maintenance Sequence and Improvement of Conjugation-Based DNA Delivery Methods.** *Front Bioeng Biotechnol* 2016, **4:**65.
29. Nymark M, Sharma AK, Sparstad T, Bones AM, Winge P: **A CRISPR/Cas9 system adapted for gene editing in marine algae.** *Sci Rep* 2016, **6:**24951.
30. Rastogi A, Murik O, Bowler C, Tirichine L: **PhytoCRISP-Ex: a web-based and stand-alone application to find specific target sequences for CRISPR/CAS editing.** *BMC Bioinformatics* 2016, **17:**261.
31. Veluchamy A, Lin X, Maumus F, Rivarola M, Bhavsar J, Creasy T, et al: **Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum.** *Nat Commun* 2013, **4**.
32. Veluchamy A, Rastogi A, Lin X, Lombard B, Murik O, Thomas Y, et al: **An integrative analysis of post-translational histone modifications in the marine diatom Phaeodactylum tricornutum.** *Genome Biol* 2015, **16:**102.
33. Daboussi F, Leduc S, Marechal A, Dubois G, Guyot V, Perez-Michaut C, et al: **Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology.** *Nat Commun* 2014, **5:**3831.

34. Serif M, Dubois G, Finoux AL, Teste MA, Jallet D, Daboussi F: **One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing.** *Nat Commun* 2018, **9:**3924.
35. Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, et al: **Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga Chlamydomonas reinhardtii.** *Plant Cell* 2015, **27:**2353-2369.
36. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, et al: **Whole-genome sequencing of multiple Arabidopsis thaliana populations.** *Nat Genet* 2011, **43:**956-963.
37. Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, et al: **Population genomics of domestic and wild yeasts.** *Nature* 2009, **458:**337-341.
38. Lachance J, Tishkoff SA: **Population Genomics of Human Adaptation.** *Annu Rev Ecol Evol Syst* 2013, **44:**123-143.
39. Godhe A, Rynearson T: **The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments.** *Philos Trans R Soc Lond B Biol Sci* 2017, **372**.
40. Harnstrom K, Ellegaard M, Andersen TJ, Godhe A: **Hundred years of genetic structure in a sediment revived diatom population.** *Proc Natl Acad Sci U S A* 2011, **108:**4252-4257.
41. Whittaker KA, Rynearson TA: **Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow.** *Proc Natl Acad Sci U S A* 2017, **114:**2651-2656.
42. Rengefors K, Kremp A, Thorsten BH, Reusch A, Wood M: **Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies.** *Journal of plankton research* 2017, **39:**165-179.
43. Matuszewski S, Hermisson J, Kopp M: **Catch Me if You Can: Adaptation from Standing Genetic Variation to a Moving Phenotypic Optimum.** *Genetics* 2015, **200:**1255-1274.
44. Bailleul B, Berne N, Murik O, Petroutsos D, Prihoda J, Tanaka A, et al: **Energetic coupling between plastids and mitochondria drives CO2 assimilation in diatoms.** *Nature* 2015, **524:**366-369.
45. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
46. Chen G, Rynearson TA: **Genetically distinct populations of a diatom co-exist during the North Atlantic spring bloom.** *LIMNOLOGY and OCEANOGRAPHY* 2016**:**2165-2179.
47. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
48. Piganeau G, Eyre-Walker A, Jancek S, Grimsley N, Moreau H: **How and why DNA barcodes underestimate the diversity of microbial eukaryotes.** *PLoS One* 2011, **6:**e16342.
49. de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J: **Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces.** *Proc Natl Acad Sci U S A* 1999, **96:**2864-2868.

50. Saez AG, Probert I, Geisen M, Quinn P, Young JR, Medlin LK: **Pseudo-cryptic speciation in coccolithophores.** *Proc Natl Acad Sci U S A* 2003, **100:**7163-7168.
51. Wolf M, Chen S, Song J, Ankenbrand M, Muller T: **Compensatory base changes in ITS2 secondary structures correlate with the biological species concept despite intragenomic variability in ITS2 sequences--a proof of concept.** *PLoS One* 2013, **8:**e66726.
52. Kaczmarska I, Mather L, Luddington I, Muise F, Ehrman J: **Cryptic diversity in a cosmopolitan diatom known as Asterionellopsis glacialis (Fragilariaceae): Implications for ecology, biogeography, and taxonomy.** *American Journal of Botany* 2014.
53. Amato A, Kooistra WH, Ghiron JH, Mann DG, Proschold T, Montresor M: **Reproductive isolation among sympatric cryptic species in marine diatoms.** *Protist* 2007, **158:**193-207.
54. Medlin LK: **A timescale for diatom evolution based on four molecular markers: reassessment of ghost lineages and major steps defining diatom evolution.** *Vie Milieu / Life & Environment* 2015.
55. Blanc-Mathieu R, Krasovec M, Hebrard M, Yau S, Desgranges E, Martin J, et al: **Population genomics of picophytoplankton unveils novel chromosome hypervariability.** *Sci Adv* 2017, **3:**e1700239.
56. Bitarello BD, de Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D, et al: **Signatures of Long-Term Balancing Selection in Human Genomes.** *Genome Biol Evol* 2018, **10:**939-955.
57. Rastogi A, Maheswari U, Dorrell RG, Vieira FRJ, Maumus F, Kustka A, et al: **Integrative analysis of large scale transcriptome data draws a comprehensive landscape of Phaeodactylum tricornutum genome and evolutionary origin of diatoms.** *Sci Rep* 2018, **8:**4834.
58. Lescot M, Hingamp P, Kojima KK, Villar E, Romac S, Veluchamy A, et al: **Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages.** *ISME J* 2016, **10:**1134-1146.
59. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: **Role of duplicate genes in genetic robustness against null mutations.** *Nature* 2003, **421:**63-66.
60. Ogura T, Wilkinson AJ: **AAA+ superfamily ATPases: common structure--diverse function.** *Genes Cells* 2001, **6:**575-597.
61. Zeytuni N, Zarivach R: **Structural and functional discussion of the tetra-trico-peptide repeat, a protein interaction module.** *Structure* 2012, **20:**397-405.
62. Helliwell KE, Wheeler GL, Leptos KC, Goldstein RE, Smith AG: **Insights into the evolution of vitamin B12 auxotrophy from sequenced algal genomes.** *Mol Biol Evol* 2011, **28:**2921-2933.
63. Bertrand EM, Allen AE, Dupont CL, Norden-Krichmar TM, Bai J, Valas RE, et al: **Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein.** *Proc Natl Acad Sci U S A* 2012, **109:**E1762-1771.
64. Helliwell KE, Collins S, Kazamia E, Purton S, Wheeler GL, Smith AG: **Fundamental shift in vitamin B12 eco-physiology of a model alga demonstrated by experimental evolution.** *ISME J* 2015, **9:**1446-1455.

65. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, et al: **The Arabidopsis thaliana mobilome and its impact at the species level.** *Elife* 2016, **5**.
66. Bonchev G, Parisod C: **Transposable elements and microevolutionary changes in natural populations.** *Mol Ecol Resour* 2013, **13:**765-775.
67. Liti G: **The fascinating and secret wild life of the budding yeast S. cerevisiae.** *Elife* 2015, **4**.
68. Blanc-Mathieu R, Verhelst B, Derelle E, Rombauts S, Bouget FY, Carre I, et al: **An improved genome of the model marine alga Ostreococcus tauri unfolds by assessing Illumina de novo assemblies.** *BMC Genomics* 2014, **15:**1103.
69. Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S, et al: **Genetic and phenotypic intra-species variation in Candida albicans.** *Genome Res* 2015, **25:**413-425.
70. Filatov DA: **Extreme Lewontin's Paradox in Ubiquitous Marine Phytoplankton Species.** *Mol Biol Evol* 2019, **36:**4-14.
71. Allen DE, Lynch M: **The effect of variable frequency of sexual reproduction on the genetic structure of natural populations of a cyclical parthenogen.** *Evolution* 2012, **66:**919-926.
72. Koester JA, Berthiaume CT, Hiranuma N, Parker MS, Iverson V, Morales R, et al: **Sexual ancestors generated an obligate asexual and globally dispersed clone within the model diatom species Thalassiosira pseudonana.** *Sci Rep* 2018, **8:**10492.
73. Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, et al: **Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus.** *Nature* 2017, **541:**536-540.
74. Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, et al: **Spontaneous epigenetic variation in the Arabidopsis thaliana methylome.** *Nature* 2011, **480:**245-249.
75. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, et al: **Timing, rates and spectra of human germline mutation.** *Nat Genet* 2016, **48:**126-133.
76. Shoemaker R, Deng J, Wang W, Zhang K: **Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome.** *Genome Res* 2010, **20:**883-889.
77. Huff JT, Zilberman D: **Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes.** *Cell* 2014, **156:**1286-1297.
78. Kanai Y, Ushijima S, Tsuda H, Sakamoto M, Hirohashi S: **Aberrant DNA methylation precedes loss of heterozygosity on chromosome 16 in chronic hepatitis and liver cirrhosis.** *Cancer Lett* 2000, **148:**73-80.
79. Patil S, Moeys S, von Dassow P, Huysman MJ, Mapleson D, De Veylder L, et al: **Identification of the meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51 homologs in the sexual species Pseudo-nitzschia multistriata and Seminavis robusta.** *BMC Genomics* 2015, **16:**930.
80. Clark DP, Flynn KJ, Ownes NJ: **The large capacity for dark nitrate-assimilation in diatoms may overcome nitrate limitation of growth.** *New Phytologist* 2002.

81. Ivanikova NV, McKay R, Bullerjahn GS: **Construction and characterization of a cyanobacterial bioreportercapable of assessing nitrate assimilatory capacity in freshwaters.** *Limnology and Oceanography* 2005, **3:**86-93.
82. Weiguo Li JW: **Influence of light and nitrate assimilation on the growth strategy in clonal weed Eichhornia crassipes.** *Aquatic Ecology* 2011.
83. Pickett-Heaps JD, Forer A: **Pac-Man does not resolve the enduring problem of anaphase chromosome movement.** *Protoplasma* 2001, **215:**16-20.
84. Nei M, Li WH: **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proc Natl Acad Sci U S A* 1979, **76:**5269-5273.
85. Vartanian M, Descles J, Quinet M, Douady S, Lopez PJ: **Plasticity and robustness of pattern formation in the model diatom Phaeodactylum tricornutum.** *The New phytologist* 2009, **182:**429-442.
86. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.** *Mol Biol Evol* 2016, **33:**1870-1874.
87. Lorenz R, Bernhart SH, Honer Zu, Siederdissen C, Tafer H, Flamm C, et al: **ViennaRNA Package 2.0.** *Algorithms Mol Biol* 2011, **6:**26.
88. Seibel PN, Muller T, Dandekar T, Schultz J, Wolf M: **4SALE--a tool for synchronous RNA sequence and secondary structure alignment and editing.** *BMC Bioinformatics* 2006, **7:**498.
89. Auton A, McVean G: **Recombination rate estimation in the presence of hotspots.** *Genome Res* 2007, **17:**1219-1227.
90. Etherington GJ, Dicks J, Roberts IN: **Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination.** *Bioinformatics* 2005, **21:**278-281.
91. Weir BS, Cockerham CC: **Estimating F-Statistics for the Analysis of Population Structure.** *Evolution* 1984, **38:**1358-1370.
92. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559-575.
93. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27:**2156-2158.
94. Alexander DH, Lange K: **Enhancements to the ADMIXTURE algorithm for individual ancestry estimation.** *BMC Bioinformatics* 2011, **12:**246.
95. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.** *Fly (Austin)* 2012, **6:**80-92.
96. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J: **KaKs_Calculator: calculating Ka and Ks through model selection and model averaging.** *Genomics Proteomics Bioinformatics* 2006, **4:**259-263.
97. Cheong WH, Tan YC, Yap SJ, Ng KP: **ClicO FS: an interactive web-based service of Circos.** *Bioinformatics* 2015, **31:**3685-3687.
98. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
99. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29:**24-26.

910
911
912

Table Legends

Table 1. Reference-assisted mapping statistics. The table summarizes the origin and year of sampling of each accession 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 Legends

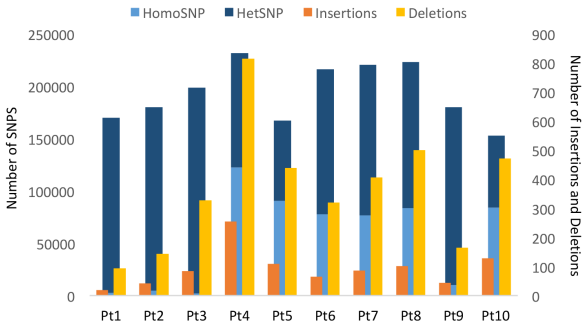
Figure 1. Genetic diversity between *P. tricornutum* accessions. (A) The bar plot represents total number of discovered SNPs, with the proportion of heterozygous SNPs (dark blue) and homozygous SNPs (light blue), INSERTIONS (orange) and DELETIONS (yellow) in each accession compared to the reference genome. (B) 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 accessions. (C) The world map indicates proportion of heterozygous (dark violet) and homozygous SNPs (violet) in each accession, represented as pie charts. The outer ring represents the proportion of variant alleles being significantly deviated from HWE (deep red). (D) The heat-map shows the genetic differentiation or association between all possible pairs of accessions. The colors indicate FST values, which range from 0.02 to 0.4, with a color gradient from yellow to green, respectively. Values closer to 0 signify close genetic makeup and values closer to 1 indicate strong genetic structuring between the populations.

Figure 2: Clustering of *P. tricornutum* accessions. (A) Principal component analysis (PCA) showing the distribution of the ten accessions based on their shared genome structure, revealing four genetic clusters referred to as Clades A, B, C, and D. (B) Pie charts showing the genetic make-up of the genomes of each accession. Each accession have a distinct genetic makeup, which is shared at both inter and intra Clade level, and represented with different colors. (C) Phylogenetic association of the accessions 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 correspond to the proportion of SNPs and INDELs over all functional features of the genome; GENEs (blue), TEs (yellow), IGRs (Intergenic Regions, represented in grey).

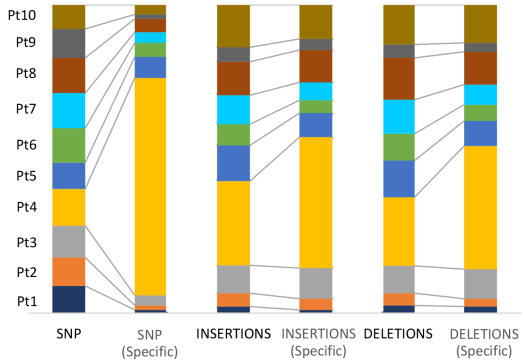
Figure 3. Large structural variations within accessions. (A) The heat-map displays the fold-change (FC) of read depth between each reference gene and median of read depth of all the reference genes, within each accession. Using Z-score as a measure of normalized read depth, log2 fold change (FC) is calculated as a ratio of Z-score per gene to the average normalized read depth of all the genes per accession. A blue to red color gradient in the heat-map represents low to high log2FC. From all the accessions only those genes are plotted where log2FC is more than 2 in at least one of the accessions and are considered to exhibit copy number variation (CNV). (B) The bar plots represent the total and specific numbers of genes, denoted on Y-axis, that

962 exhibit a loss or multiple copies (CNV) within one or more accessions. (C) The bar
963 plots represent the number of total- and accession-specific TEs exhibiting CNVs
964 across one or more accessions. (D) With similar principle aesthetics as panel (A) of
965 this figure, the heat-map shows the patterns of log2FC only across all the accessions
966 of those TEs exhibiting CNV in at least one of the ten accessions studied.
967
968 **Figure 4. Evolutionary and functional consequences of polymorphisms.** (A) The
969 bar plot represents total and specific numbers of genes that are subject to balancing
970 selection, or experiencing loss-of-function (LoF) mutations. For each category, the
971 accessions are plotted as stack plots with total and specific numbers of genes.
972 Numbers of genes in each category are indicated. (B) The box plot represents the
973 number of gene families affected by loss-of-function (LOF) mutations and suggests a
974 bias of such mutations on the genes belonging to large gene families. Y-axis
975 represents, as log scale, the number of genes in the gene families vs those that are not
976 affected by LoF mutations.
977
978 **Figure 5. Balancing selection within each genetic Clade.** Based on the EfR metric,
979 the network displays highly affected gene families experiencing balancing selection.
980 Gene families associated with *MetH* genes under balancing selection in all the
981 accessions are indicated within the blue circles. The red circles group individual
982 accessions as Clades.
983
984 **Figure 6. Selection of MetH-facilitated methionine biosynthesis over MetE.** The
985 bar plots represent relative expression of (A) *MetH*, (B) *MetE*, (C) *CBA1* and (D)
986 *SHMT* genes in four (Pt2, Pt3, Pt4 and Pt8) of the ten accessions with the presence of
987 vitamin B12 in axenic cultures (light gray), and with natural bacteria and no vitamin
988 B12 in the growing media (black).
989
990 **Figure 7. Population structure of the ecotypes.** The color gradient from yellow to
991 blue indicates low to high numerical values across each ecotype (indicated on top X-
992 axis of panel A) within different functional categories indicated on Y-axis. These
993 functional categories includes (from top to bottom), Year of sampling = Year in
994 which the respective ecotype was sampled, Total SNP = Absolute number of SNPs
995 found in each ecotype, Specific/Total SNP (%) = percentage of ecotype specific
996 SNPs, Heterozygosity = Number of heterozygous SNPs from a set of total SNPs
997 within each ecotype, Homozygosity = Number of homozygous SNPs from a set of
998 total SNPs within respective ecotype, Sites deviated from HWE = Number of SNP
999 sites predicted to be deviated from Hardy-Weinberg equilibrium (HWE), BS =
1000 Number of genes under Balancing Selection, LoF = Number of genes localizing Loss
1001 of Function variant sites.
1002
1003

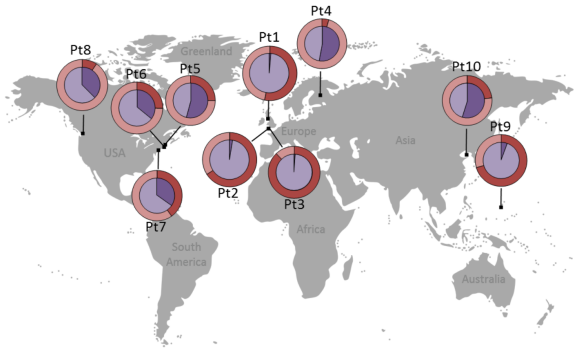
A



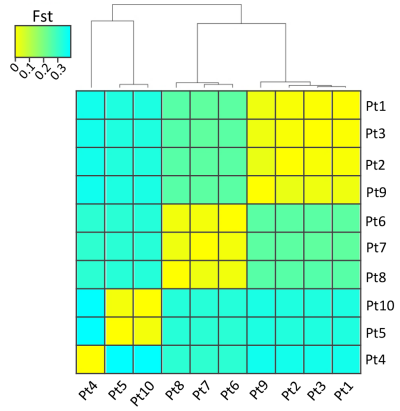
B



C

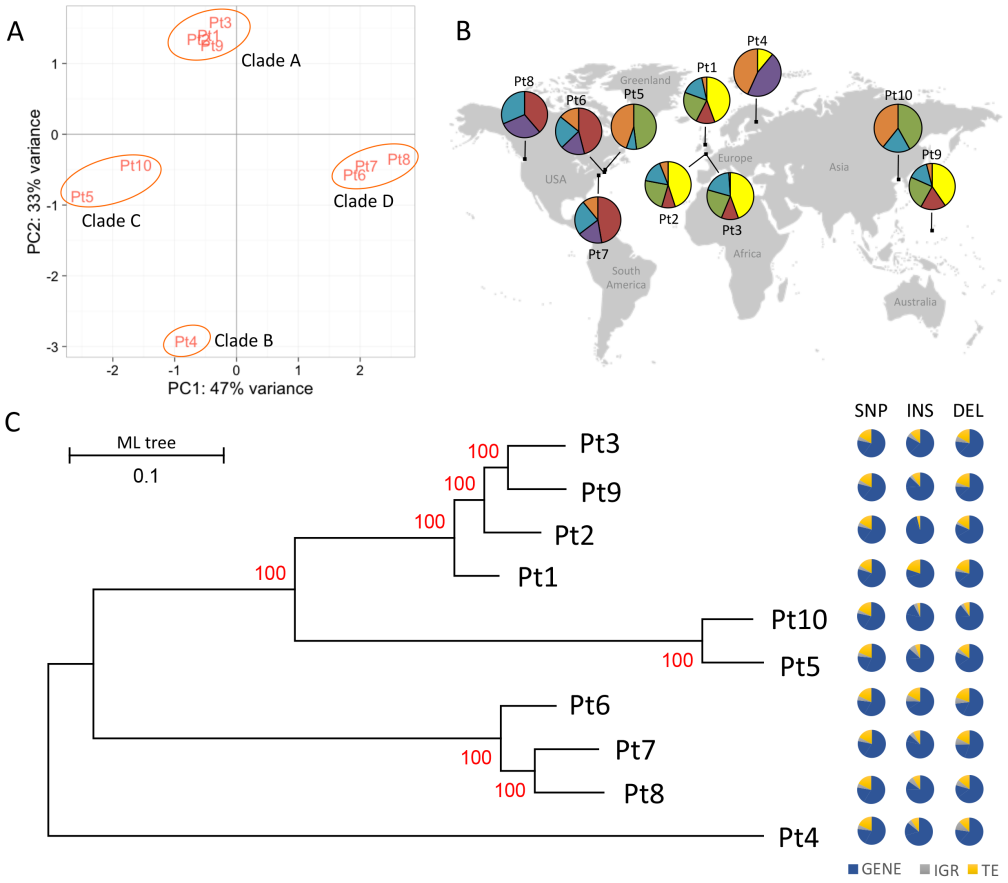


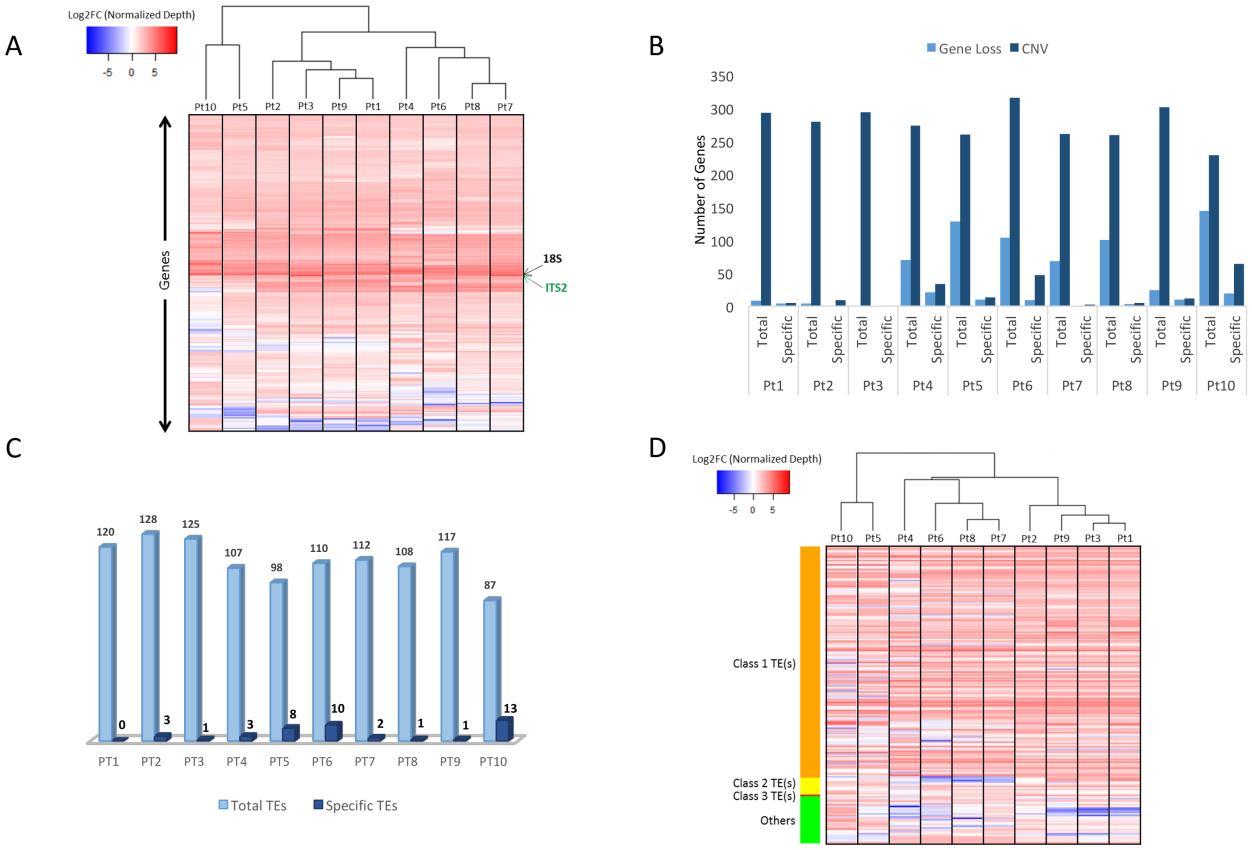
D



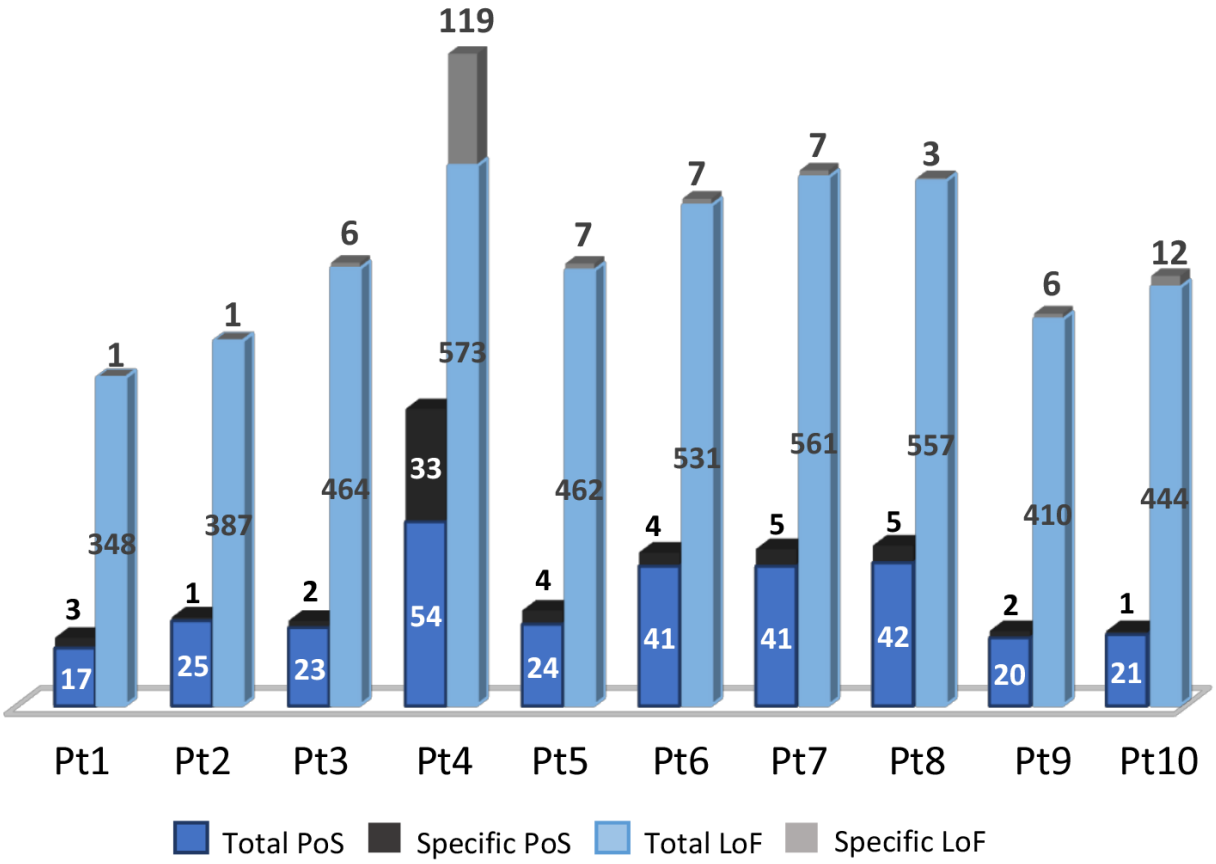
Library Name	Origin	Year of Isolation	Mapped Read-Pairs	% Mapped Read-Pairs	Alignment Depth (X)	Genome Coverage (%)
Pt1	Blackpool, UK	1956	3, 642,044	79.41	26.5	98.0
Pt2	Plymouth, UK	Prior to 1910	6, 016,241	78.23	43.8	98.0
Pt3	Plymouth, UK	1930s	6, 373,591	65.62	46.4	98.3
Pt4	Finland	1951	15, 583,665	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

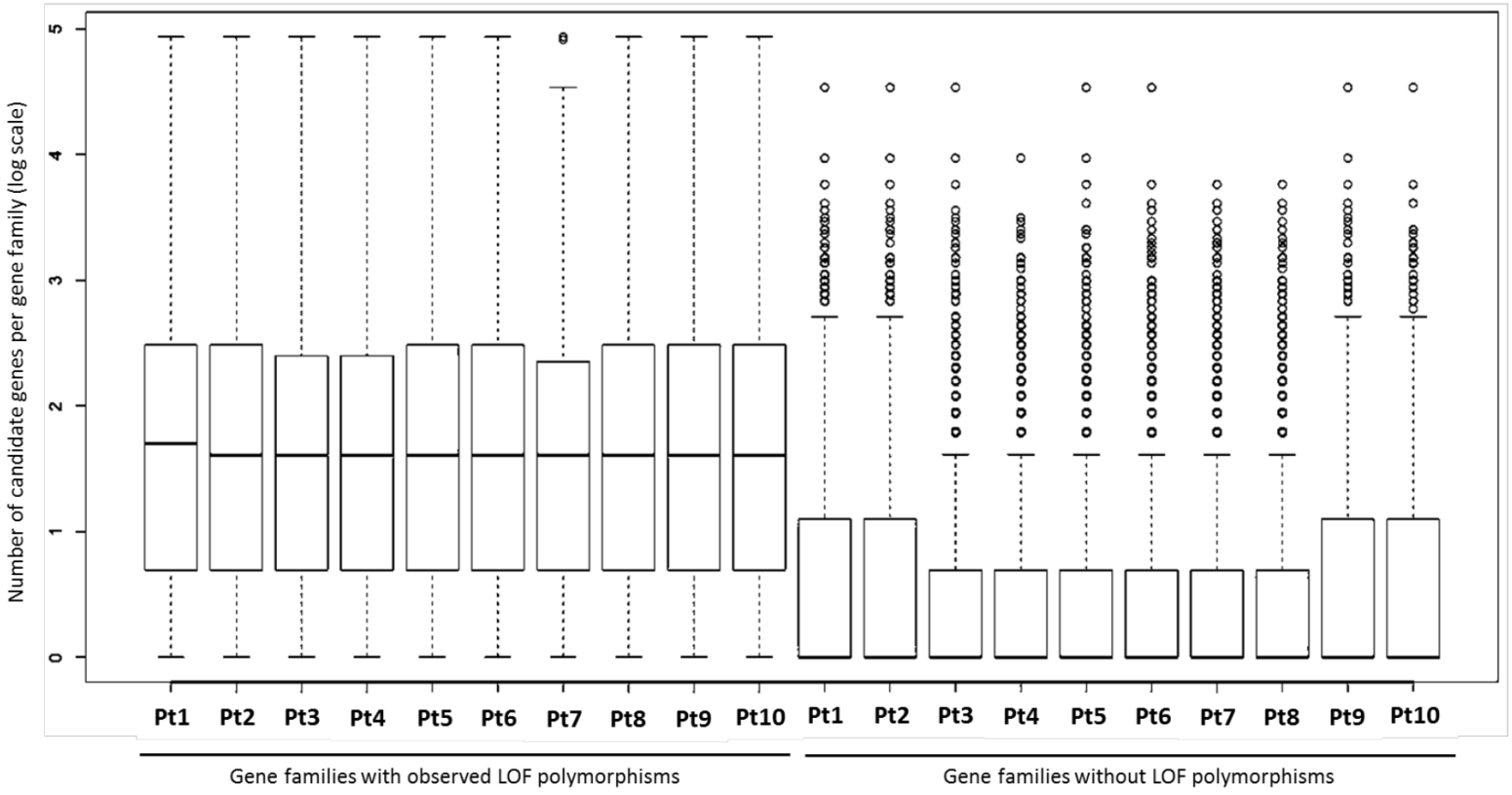


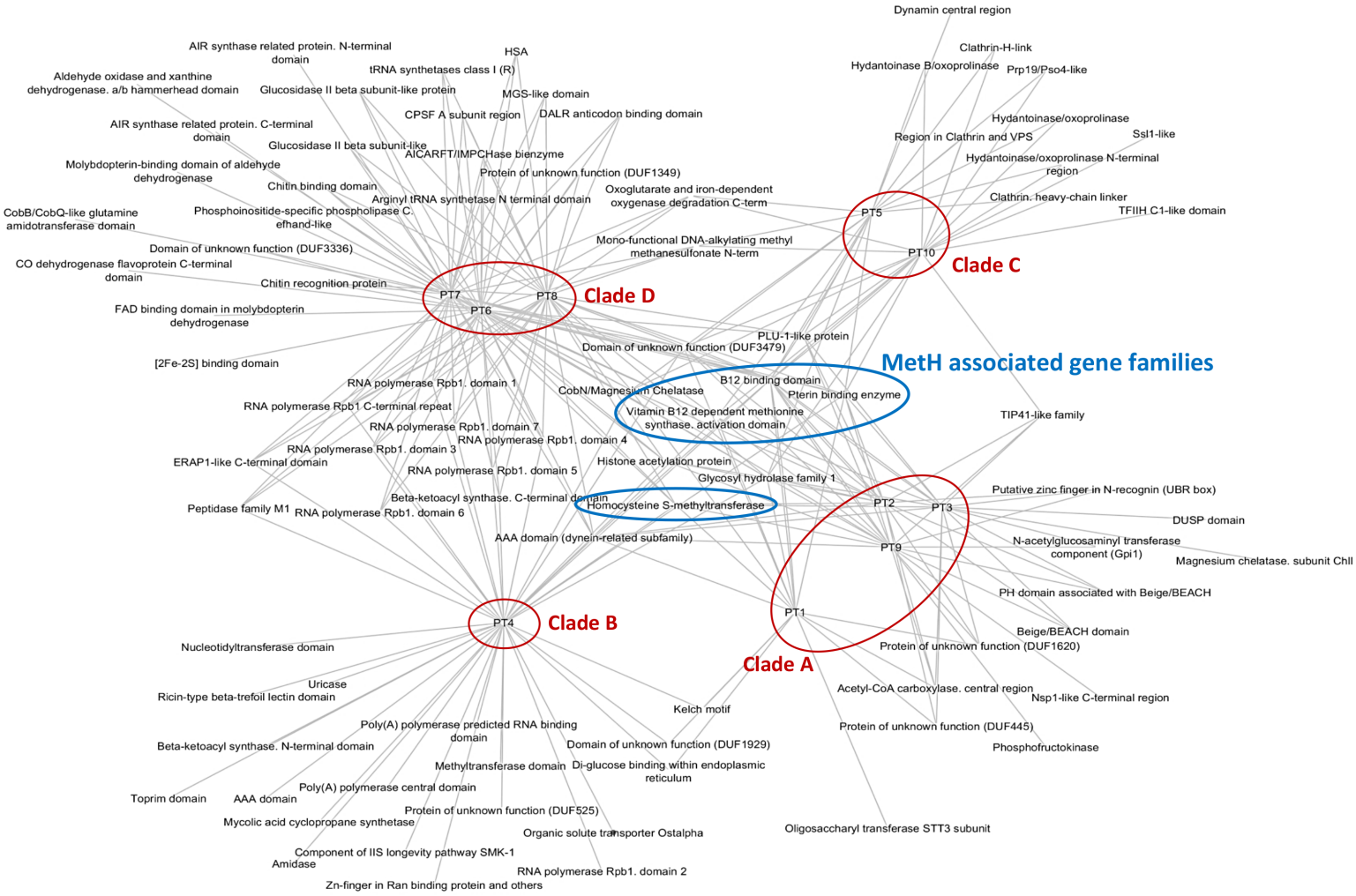


A

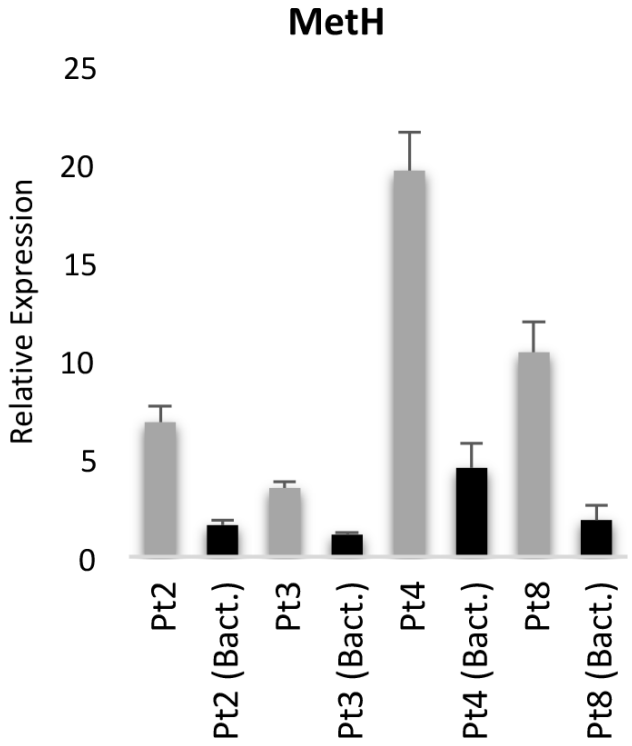


B

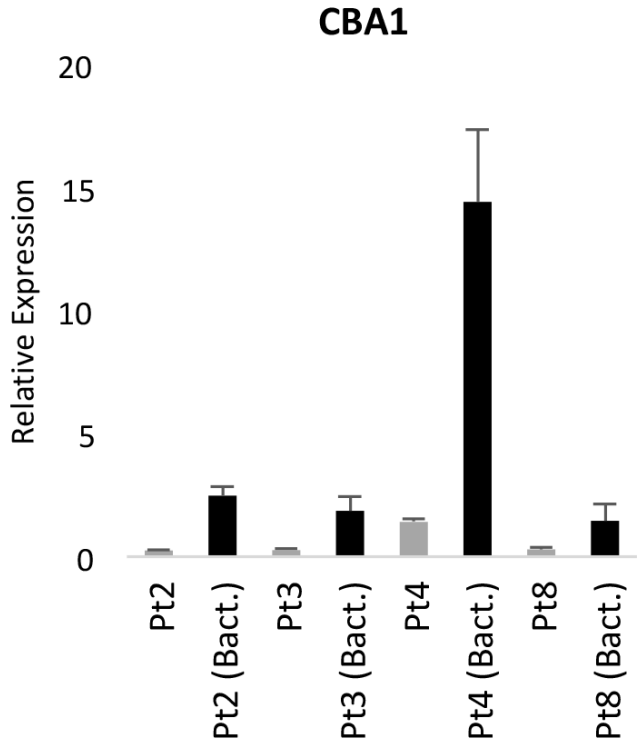




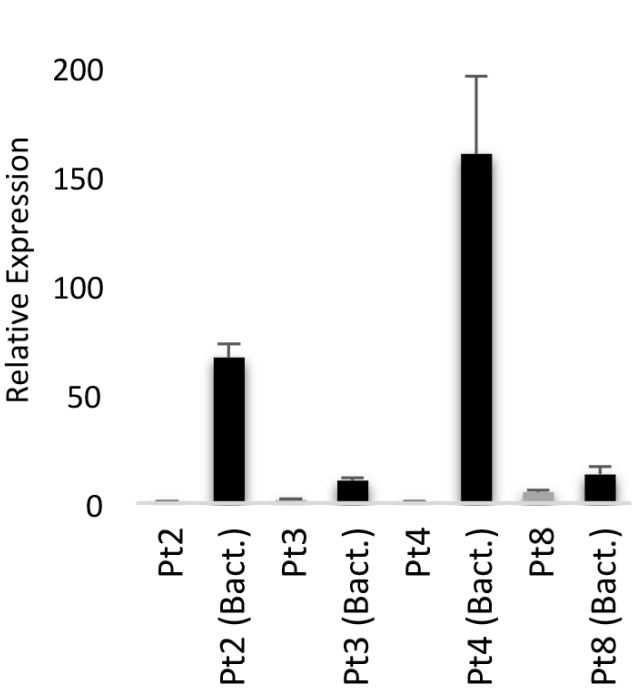
A



C



B



D