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: Corteva Agriscience™, The V Ascendas, Atria Block, 12th Floor, Madhapur, Hyderabad 500081, India

[$]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 a big group of heterokonts, constituent of chromalveolate (SAR group) which are believed to be derived from serial endosymbiosis combining genes from green and red algae predecessors and further diversified via horizontal gene transfer from a wide range of prokaryotes [1-3]. Ehrenberg [4] 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 [5]. In nature, most diatoms likely live in obligate relationships with bacteria [6] but many, like *Phaeodactylum tricornutum*, can be propagated in axenic conditions. In spite of its low abundance in the open ocean [7], *P. tricornutum* is extensively used as a model to study and characterize diatom metabolism, and to understand diatom evolution [1, 8-12].

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 sensitivity to many nonspecific abiotic components and the general lack of knowledge on sexual reproduction in this species [13-15] 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 isolates 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 [16]. These accessions have been collected within the time frame of approximately one century, from 1908 (Plymouth isolate, Pt2/3) to 2000 (Dalian isolate, Pt10) (Fig. S1) [16]. All the isolates 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 [17-20], 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 [16]. *P. tricornutum* is among the few diatom species with a whole genome sequence available to the community [21], and the only diatom for which extensive state-of-the-art functional and molecular tools have been developed over the past few decades [22-35]. 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 [36-39]. 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 [40]. Such studies in diatoms are rare and estimates of genetic diversity within diatom populations are mostly inferred using microsatellite-based genotyping approaches [41-43]. Although these techniques have revealed a wealth of information about diatom evolution, their dispersal and reproductive physiology [40], additional insights can be obtained using state-of-the-art whole genome comparative analysis techniques [43]. Deciphering the standing

genomic variation of *P. tricornutum* across different accession populations, sampled at broad geospatial scale, is an important first step to assess the role of various evolutionary forces in regulating the adaptive capacities of diatoms in general (e.g.[44]). To understand the underlying genomic diversity within different accessions of *P. tricornutum* and to establish the functional implications of such diversity, we performed deep whole genome sequencing of the 10 most studied accessions, referred to as Pt1 to Pt10 [16, 19, 45]. We present a genome-wide diversity map of geographically distant *P. tricornutum* accessions, describing a stable genetic structure in the environment. This work further provides the community with whole genome sequences of the accessions, which will be a valuable genetic resource for functional studies of accession-specific ecological traits in the future.

Results

Reference-assisted assembly reveals low nucleotide diversity across multiple accessions of *P. tricornutum*

We sequenced the whole genomes of ten accessions of *P. tricornutum* using Illumina HiSeq 2000, and performed a reference-based assembly using the genome sequence of the reference strain Pt1 8.6 [1]. Across all accessions, the percentage of sequence reads mapped on the reference genome ranged between ~65% to ~80% (Table 1), with an alignment depth ranging between 26X and 162X, covering 92% to 98% of the reference genome (Table 1). Many regions on the reference genome that are observed as being unmapped by reads from individual ecotypes are annotated as rich in transposable elements (TEs) (Fig. S2). At >90% identity, the repeated proportion of unmapped reads varies between ~38% (Pt1) and 75% (Pt4).

138 Following the assembly, we performed variant calling using Genome Analysis
139 Toolkit [46] and discovered 462,514 (depth >= 4x) single nucleotide polymorphisms
140 (SNPs) including ~25% singleton sites, 573 insertions (of varying lengths from 1 bp
141 to 312 bp) and 1,801 deletions (of lengths from 1 bp to 400 bp) (Fig. 1A), across all
142 the accessions. The spectrum of SNPs across all the accessions further reveals a
143 higher rate of transitions (Ts) over transversions (Tv) (Ts/Tv = 1.6). In total,
144 compared to the reference alleles from Pt1.8.6, six possible types of single nucleotide
145 changes could be distinguished, among which G:C -> A:T and A:T -> G:C accounted
146 for more than ~60% of the observed mutations (Fig. S3A). Further, most SNPs and
147 INDELs (insertions and deletions) are shared between different accessions, except for
148 Pt4, which possesses the highest proportion of specific SNPs (~35%) and INDELs
149 (~75%) (Fig. 1B). Interestingly, we found that most of the SNPs are heterozygous,
150 and the proportion of heterozygous variants across all the accessions varies between
151 ~45% (in Pt5 and Pt10) to ~98% (in Pt1, Pt2 and Pt3) (Fig. 1C). Most of the variant
152 alleles in the accessions with high proportions of heterozygous variants were further
153 found to be significantly deviated from Hardy-Weinberg equilibrium (HWE) (chi-
154 square test, P-value < 0.05) (Fig. 1C), possibly linked to prolonged asexual
155 reproduction [47]. Surprisingly, despite significant differences in the proportion of
156 heterozygote variant alleles between the accessions, which ranges between 45% to
157 98%, the average pairwise synonymous nucleotide diversity (π_S) estimated from
158 genes with callable sites across all the accessions is 0.007 per synonymous site. This
159 indicates that any two homologous sequences taken at random across different
160 populations will on average differ by only ~0.7% on synonymous positions. The non-
161 synonymous pairwise diversity (π_N) over the same genes is 0.003, consistent with an
162 excess of non-synonymous mutations being deleterious. An average non-synonymous

(N) to synonymous (S) 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 [48]. Since π_N / π_S is negatively correlated with the effective population size, which is the number of distinct clones in unicellular organisms[49], this could reflect either a lower census population size in *P. tricornutum*, and/or a lower rate of sexual reproduction. 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 [50]. Based on the allelic composition of the ten genomes, six genomic clusters (K=6) can be formed, which is distributed within each accession genome in different proportions (represented by pie-chart colors) (Fig. 2A). Further, depending on the pattern (both

188 qualitative and quantitative) of distributed genomic clusters across different accession
189 genomes, the ten accessions revealed four genetic clusters with Pt1, Pt2, Pt3, and Pt9
190 in one, Pt4 in a second, Pt5, Pt10 in a third, and Pt6, Pt7, Pt8 in a fourth cluster (Fig.
191 2B). These clusters (Fig. 2B) are in broad agreement with *Fst*-based genetic clades
192 (Fig. 1D), phylogenetic clusters inferred using ribosomal marker genes (18S (Fig.
193 S4A), and ITS2 (Fig S4B)), as also reported previously [16], and at whole genome
194 scale (this study) as inferred by a phylogenetic tree generated using maximum
195 likelihood algorithm based on all (Fig. 2C) and only homozygous polymorphic sites
196 (SNVs and INDELS) (Fig. S4C).
197 Further sequential assessment of the 18S and ITS2 rDNA gene sequences across
198 different clades indicated the presence of multiple variations, including both
199 heterozygous and homozygous variant alleles (Fig. S4D and S4E). Because the
200 ribosomal DNA region including 18S and ITS2 is highly repetitive, which is on
201 average ~4 times more than non-ribosomal genes (Fig. 3A), these differences can be
202 understood as intra-genomic variations within the genome. However, taxonomists and
203 ecologists use differences within 18S gene sequences as a measure of species
204 assignation and to estimate species delineation [7]. This latter practice has been
205 previously shown to be very conservative as no differences in the 18S gene were
206 found between reproductively isolated species [51]. Alternatively, the possibility of
207 sub-populations or cryptic populations cannot be ignored, as previously reported in
208 planktonic foraminifers [52] and coccolithophores [53].
209 We examined the possible presence of sub-populations on 18S gene heterozygosity in
210 some of the accessions. We confirmed the expression of all the heterozygous alleles
211 within the 18S rDNA gene using whole genome and total-RNA sequencing of a
212 monoclonal culture (propagated from a single cell) from Pt8 (constituent of Clade D)

and Pt3 (constituent of Clade A) population, referred to as Pt8Tc and Pt3Ov (Fig. S4D), respectively. This experiment indicates that the cultures (Pt1-Pt10) are a single population with no or undetectable heterogeneity.

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 [54] and between diatom species [55, 56]. 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 [21, 57], 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 [17-19].

Evolutionary adaptation in *P. tricornutum* clades

262 Species are under continuous pressure to adapt to a changing environment over time.
263 We therefore wanted to understand the functional consequences of the genetic
264 diversity between the accessions. Localization of the polymorphic sites over genomic
265 features (genes, TEs, and intergenic regions) revealed highest rate of variation within
266 genes (Fig. 2C), specifically on exons, and was consistent across all the studied
267 accessions. We further identified genes within different phylogenetic clades
268 experiencing different selection pressure based on lowest and highest π_N / π_S ratios.
269 Across all the accessions, 241 genes displaying π_N / π_S >1 and a higher frequency of
270 non-synonymous to synonymous polymorphism, as expected under balancing
271 selection (BS) [58] (File S3). Furthermore, across all the accessions, 128 genes
272 exhibit a signature of relaxed selection (RS) with accumulation of NS
273 polymorphisms, among which 47% are specific to one or other clades (Fig. 4A). In
274 addition, many genes (902) were found to have loss-of-function (LoF hereafter)
275 variant alleles (Fig. 4A), including frame-shift mutations and mutations leading to
276 theoretical start/stop codon loss and/or gain.
277 Based on the presence of functional domains (Pfam domains), all *P. tricornutum*
278 annotated genes [63] were grouped into 3,020 gene families. These families can be as
279 large as the reverse transcriptase gene family, which is highly abundant in marine
280 plankton [64], representing 149 candidate genes having reverse transcriptase domains,
281 or as small as families that constitute single gene candidates. Across all the
282 accessions, we observed that most genes experiencing LoF mutations belong to large
283 gene families (Fig. 4B). This is consistent with a previous observation of the existence
284 of functional redundancy in gene families as a balancing mechanism for null
285 mutations in yeast [65]. Therefore, to estimate an unbiased effect of any evolutionary
286 pressure (eg. 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 enrichment of constrained genes revealed enrichment of (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 [66], (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 [67]. 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).

Finally, considering all pairwise correlated gene families exhibiting similar selection signals, measured using EfR among the 10 accessions, we used hierarchical clustering to examine the functional closeness of accessions with one another (See methods). 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 clade-specific genes that are under high selection pressure as depicted by high rate of N/S divergence when compared to the reference, a group of gene

families associated with methionine biosynthesis *(MetH*, Phatr3_J23399) was also observed to accumulate non synonymous polymorphisms in all the accessions. 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* [68-70]. In accordance with these results, we observed a high expression of *MetH* in axenically grown laboratory cultures (Fig. 6A) compared to its expression in cells cultured with their natural co-habitant bacteria. However, we were not able to trace any significant signature for the loss of *MetE* gene although its expression is significantly lower in axenic cobalamin-containing cultures (Fig. 6B). Similar observations were obtained for *CBA1* and *SHMT* genes (Fig. 6C and 6D), which under cobalamin scarcity enhance cobalamin acquisition and manage reduced methionine synthase activity, respectively [69].

Discussion

Using whole genome sequence analysis of accessions sampled across multiple geographic locations around the world (Fig. S1), the aim of this study was to describe the global genetic and functional diversity of the model diatom *Phaeodactylum tricornutum*. By defining a comprehensive landscape of natural variations across multiple accessions, we could investigate genetic structure between *P. tricornutum* populations, and a summary of our results is presented in Fig. 7. To do so, we first performed reference-based assembly and found consistently high genome coverage

337 (>90%) mapped by sequencing reads from respective accessions, where some
338 accessions have more coverage (>98%, Pt1, Pt2, Pt3, and Pt9) than others (Table 1).
339 This difference is independent of the size of the sequencing library, as it does not
340 correlate with the genome coverage (Table 1), and a portion of unmapped reads is
341 likely a consequence of the incomplete reference genome, which contains several
342 gaps [1]. Additionally, given the redundant nature of unmapped reads together with
343 the fact that the unmapped reference genome is annotated as being rich in TEs (Fig.
344 S2), a major portion of unmapped reads likely account for large structural variability
345 within the genomes of individual accessions. This explanation is most clear in Pt10,
346 which is shown to have the largest number of gene losses (Fig. 3B) and the highest
347 number of accession-specific TEs with high copy numbers (Fig. 3C) and covers at
348 least (92%) of the reference genome (Table 1). This suggests the role of TEs in
349 creating substantial genetic diversity as also shown in many species of plants and
350 animals [71, 72].
351 Next, based on patterns of variations discovered over the whole genomes and on the
352 molecular marker genes (18S and ITS2) of all the accessions, and by using various
353 clustering algorithms (see Results), the ten accessions could be grouped into four
354 genetic clades. Clade A clusters Pt1, Pt2, Pt3, Pt9; clade B includes Pt4; clade C
355 clusters Pt5, Pt10; and clade D clusters Pt6, Pt7, Pt8. Most of the structural variants
356 discovered, both small (SNPs and INDELS) and large (CNV and Gene Loss), are
357 shared among populations within a clade rather than between clades. This suggests
358 high intra-clade relatedness over a variety of structural, functional and possibly
359 ecological traits.
360 *P. tricornutum* is a coastal species with limited dispersal potential, which is consistent
361 with the reports of its absence in the open ocean from *Tara* Oceans data [7].

Consequently, the Fixation index (*Fst*) between different genetic clades is very high (0.2 – 0.4) (Fig. 1D), confirming the existence of accessions subdivisions into four genetic clades. As also expected for an organism with limited dispersal potential, the accessions show partial geographical structuring (Fig. 2A, 2B), as Pt5, Pt9 and Pt10 clusters with accessions not sampled from proximal locations. These dispersals to different localities may be fostered by ocean currents [42], human activities like rafting, ballasting [73, 74], and migration of birds [75-77]. 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 [41, 42]. This suggests as reported in these studies that environmental conditions have a more important impact in structuring the populations than dispersal potential and generation time.

Although there exist partial 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 [36, 48, 78-80] but in line with previous estimations in marine phytoplanktonic eukaryotes [81]. 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* [82]. 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). High *Fst*, and yet low nucleotide diversity across all the accessions, suggests some degree of genetic and functional convergence among the accessions. This can be explained as a consequence of laboratory culture mediated domestication where some genes accumulate nonsynonymous mutations which might result from the relaxation of selective constraints in an artificial environment.

It is also worth considering that genetic homogenization or functional convergence 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 show partial geographical structure, except Pt9 of clade A, Pt5 of clade C and Pt8 of clade D (Fig. 2A, 2B). This is consistent with earlier findings in *Emilinia huxleyi* and *Thalassiosira pseudonana* in which genomes of isolates from different geographical locations clustered together in the phylogeny [83, 84]. 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 [85]. 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 [7], the possibility of gene flow within different populations is likely to be limited at best.

Next, high linkage disequilibrium (>0.7) observed across all the accessions (Fig. S3B) can be explained by prolonged asexual reproduction [86], a common behavior among

412 diatoms [84]. Asexual reproduction results in higher proportions of divergent alleles
413 within loci with less genetic variation among individuals, and a significant deviation
414 from HWE [86]. Therefore, it is likely that clade B with one isolate, Pt4 also
415 undergoes sexual reproduction reasonably often compared to clade A accessions, as
416 Pt4 possesses the smallest number of heterozygous variant alleles, most of which
417 follow HWE (Fig. 1C, 7). To our surprise, despite high variability in the levels of
418 heterozygosity between different accessions (Fig. 7), the mutational spectrum,
419 compared to the reference, and across all the accessions consisted of high G:C -> A:T
420 and A:T -> G:C transitions (Fig. S3A). Deamination of cytosines dominantly dictates
421 C to T transitions in both plants and animals [87, 88], and CpG methylation potential
422 of the genome is greatly influenced by heterozygous SNPs in CpG dinucleotides [89].
423 Previous studies have demonstrated low DNA methylation in *P. tricornutum,* using
424 Pt1 8.6, a monoclonal isolate accessioned from a Pt1 single cell as a reference [32,
425 90]. Because there exist significant differences in the proportion of heterozygote
426 variant alleles between the accessions (45% - 98%), testing for DNA methylation
427 patterns across different accessions may provide an interesting opportunity to dissect
428 cross-talk between loss of heterozygosity and DNA methylation in the selection of
429 certain traits [91].
430 The four genetic clades are further supported by functional specialization of grouped
431 populations (Fig. 5), nicely illustrated with Pt4 in clade B. Pt4 shows a low non-
432 photochemical quenching capacity (NPQ) [18], which was proposed to be an adaptive
433 trait to low light conditions. Specifically, this accession has been proposed to
434 establish an up-regulation of a peculiar light harvesting protein LHCX4 in extended
435 dark conditions [18, 20]. In line with these observations, a gene involved in nitrate
436 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* [92]. 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 [93, 94]. Additional functions emerging from clade C (Pt5 and Pt10) include vacuolar sorting and vesicle-mediated transport gene-families, which could be an indication of altered intracellular trafficking [95].

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[96]. 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, SAMN12551644 (Pt3Ov genomic), SAMN12551645 (Pt3Ov Transcriptomic), and SAMN12551646 (Pt8Tc genomic). 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 S4.

Discovery of small polymorphisms and large structural variants

GATK [46], configured for diploid genomes, was used for variant calling, which included single nucleotide polymorphisms (SNPs), small insertions (of varying lengths from 1 bp to 312 bp) and deletions (of lengths from 1 bp to 400 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 [97] were estimated for all genes using in-house R script from [97] equation 22 for each gene and the complete callable coding sequences (available from the authors upon request).

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 S4. 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 [22] 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 [23] from three biological replicates of accessions grown axenically in Artificial Sea Water (ASW) [98] supplemented with vitamins as well as in the presence of their endemic bacteria in ASW without vitamins. qPCR was performed as described previously [23]. Briefly, cDNA was synthetized from 1 µg RNA using High-Capacity cDNA Reverse Transcription Kit (catalogue number 4368813) from Fischer scientific and according to manufacturer instructions. 1µl of cDNA was used in the QPCR reaction with the LightCycler® 480 SYBR Green I Master (catalogue number 04707516001) from Roche and according to manufacturer instructions. Two reference genes were used, Tata Box binding Protein and Ribosomal Protein Small subunit 30S for normalization [23].

***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 [99] to align and deduce the phylogenetic trees.

***CBC analysis*:** CBC analysis was done by generating the secondary structure of ITS2 sequences, using RNAfold [100], 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 [101] 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 [102] and RAT [103]. 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 [104].

***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) [50]. Accessory tools like PLINK (version 1.07-x86_64) [105] and VCFtools (version 0.1.13) [106] 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 [107]. We chose the value of K with lowest cross-validation error (see extended methods, File S4). 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 [108] and KaKs [109] 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. π_N/π_S ratios were calculated for 5232 protein coding genes containing more than 10 SNP. 10% of genes with lower π_N/π_S were considered as under strong purifying selection on amino-acid composition (File S3). Genes with π_N/π_S >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 [110], CYTOSCAPE [111], IGV [112] and R (https://www.r-project.org/about.html). Based on the presence of functional domains all the Phatr3 genes [63] were grouped into 3,020 gene families. Subsequently, the constituents of each gene family were checked for being either affected by loss-of-function mutations or under relaxed selective constraints. To estimate an unbiased effect of any evolutionary pressure (LoF allele or balancing 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, purifying selection, balancing selection (BS), or showing CNV, or being lost (GnL), were estimated by calculating observed to expected ratio of their percent occurrence within the given functional set (BS, LoF, CNV) and their occurrence in the complete annotated Phatr3 (http://protists.ensembl.org/Phaeodactylum_tricornutum /Info/Index) biological process catalog. Later, considering gene family EfR as a function to measure the association rate, we deduced Pearson pairwise correlations between different accessions. The correlation matrix describes that if many equally affected gene 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, MEMO LIFE (ANR-10-LABX-54) and the region of Pays de la Loire (ConnecTalent EPIALG project). AR was supported by an International PhD fellowship from MEMO LIFE (ANR-10-LABX-54).

Contributions

LT, HH and CB conceived the study. LT, AR and GP designed the study. GW, PV, AFDC, CC and LT did the experiments. AR, FRJV, AV and GP developed and performed the bioinformatics analysis. AR, GP, FRJV and LT interpreted the results. AR and LT wrote the manuscript with input from all the authors. LT coordinated and supervised the study.

Conflict of interest

The authors declare no conflicts of interest.

References

1. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, 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, Brillet-Gueguen L, Freeman KD, Richter DJ, Bowler C: **Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome.** *Elife* 2017, **6**.
4. C.G. E: **Die Infusionsthierchen als vollkommene Organismen. Ein Blick in das tiefere organische Leben der Natur.** In *Leipzig,* . Leipzig: Leopold Voss.; 1838.
5. Armbrust EV: **The life of diatoms in the world's oceans.** *Nature* 2009, **459:**185-192.
6. Amin SA, Parker MS, Armbrust EV: **Interactions between diatoms and bacteria.** *Microbiol Mol Biol Rev* 2012, **76:**667-684.
7. 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**

665 **distribution and diversity in the world's ocean.** *Proc Natl Acad Sci U S A*
666 2016.
667 8. Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, McCrow JP, Zheng H,
668 Johnson DA, Hu H, Fernie AR, Bowler C: **Evolution and metabolic significance**
669 **of the urea cycle in photosynthetic diatoms.** *Nature* 2011, **473:**203-207.
670 9. Huysman MJ, Fortunato AE, Matthijs M, Costa BS, Vanderhaeghen R, Van den
671 Daele H, Sachse M, Inze D, Bowler C, Kroth PG, et al: **AUREOCHROME1a-**
672 **mediated induction of the diatom-specific cyclin dsCYC2 controls the onset**
673 **of cell division in diatoms (Phaeodactylum tricornutum).** *Plant Cell* 2013,
674 **25:**215-228.
675 10. Morrissey J, Sutak R, Paz-Yepes J, Tanaka A, Moustafa A, Veluchamy A,
676 Thomas Y, Botebol H, Bouget FY, McQuaid JB, et al: **A novel protein,**
677 **ubiquitous in marine phytoplankton, concentrates iron at the cell surface**
678 **and facilitates uptake.** *Curr Biol* 2015, **25:**364-371.
679 11. Tanaka T, Maeda Y, Veluchamy A, Tanaka M, Abida H, Marechal E, Bowler C,
680 Muto M, Sunaga Y, Tanaka M, et al: **Oil accumulation by the oleaginous**
681 **diatom Fistulifera solaris as revealed by the genome and transcriptome.**
682 *Plant Cell* 2015, **27:**162-176.
683 12. Fortunato AE, Jaubert M, Enomoto G, Bouly JP, Raniello R, Thaler M, Malviya
684 S, Bernardes JS, Rappaport F, Gentili B, et al: **Diatom phytochromes reveal**
685 **the existence of far-red-light-based sensing in the ocean.** *Plant Cell* 2016,
686 **28:**616-628.
687 13. Godhe A, Kremp A, Montresor M: **Genetic and microscopic evidence for**
688 **sexual reproduction in the centric diatom Skeletonema marinoi.** *Protist*
689 2014, **165:**401-416.
690 14. Moore ER, Bullington BS, Weisberg AJ, Jiang Y, Chang J, Halsey KH:
691 **Morphological and transcriptomic evidence for ammonium induction of**
692 **sexual reproduction in Thalassiosira pseudonana and other centric diatoms.**
693 *PLoS One* 2017, **12:**e0181098.
694 15. Mouget JL, Gastineau R, Davidovich O, Gaudin P, Davidovich NA: **Light is a**
695 **key factor in triggering sexual reproduction in the pennate diatom Haslea**
696 **ostrearia.** *FEMS Microbiol Ecol* 2009, **69:**194-201.
697 16. De Martino AM, A. Juan Shi, K.P. Bowler, C.: **Genetic and phenotypic**
698 **characterization of Phaeodactylum tricornutum (Bacillariophyceae)**
699 **accessions.** *J Phycol* 2007, **43:**992–1009.
700 17. Stanley MS, Callow JA: **Whole cell adhesion strength of morphotypes and**
701 **isolates of Phaeodactylum tricornutum (Bacillariophyceae).** *European*
702 *Journal of Phycology* 2007, **42:**191-197.
703 18. Bailleul B, Rogato A, de Martino A, Coesel S, Cardol P, Bowler C, Falciatore A,
704 Finazzi G: **An atypical member of the light-harvesting complex stress-related**
705 **protein family modulates diatom responses to light.** *Proc Natl Acad Sci U S A*
706 2010, **107:**18214-18219.
707 19. Abida H, Dolch LJ, Mei C, Villanova V, Conte M, Block MA, Finazzi G, Bastien
708 O, Tirichine L, Bowler C, et al: **Membrane glycerolipid remodeling triggered**
709 **by nitrogen and phosphorus starvation in Phaeodactylum tricornutum.**
710 *Plant Physiol* 2015, **167:**118-136.

711 20. Taddei L, Stella GR, Rogato A, Bailleul B, Fortunato AE, Annunziata R, Sanges
712 R, Thaler M, Lepetit B, Lavaud J, et al: **Multisignal control of expression of**
713 **the LHCX protein family in the marine diatom Phaeodactylum tricornutum.**
714 *J Exp Bot* 2016, **67:**3939-3951.
715 21. Tirichine L, Rastogi A, Bowler C: **Recent progress in diatom genomics and**
716 **epigenomics.** *Curr Opin Plant Biol* 2017, **36:**46-55.
717 22. Falciatore A, Casotti R, Leblanc C, Abrescia C, Bowler C: **Transformation of**
718 **nonselectable reporter genes in marine diatoms.** *Mar Biotechnol (NY)* 1999,
719 **1:**239-251.
720 23. Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia
721 A, Falciatore A, Bowler C: **Molecular toolbox for studying diatom biology in**
722 **Phaeodactylum tricornutum.** *Gene* 2007, **406:**23-35.
723 24. De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A: **Gene**
724 **silencing in the marine diatom Phaeodactylum tricornutum.** *Nucleic Acids*
725 *Res* 2009, **37:**e96.
726 25. Huysman MJ, Martens C, Vandepoele K, Gillard J, Rayko E, Heijde M, Bowler
727 C, Inze D, Van de Peer Y, De Veylder L, Vyverman W: **Genome-wide analysis**
728 **of the diatom cell cycle unveils a novel type of cyclins involved in**
729 **environmental signaling.** *Genome Biol* 2010, **11:**R17.
730 26. Maheswari U, Jabbari K, Petit JL, Porcel BM, Allen AE, Cadoret JP, De Martino
731 A, Heijde M, Kaas R, La Roche J, et al: **Digital expression profiling of novel**
732 **diatom transcripts provides insight into their biological functions.** *Genome*
733 *Biol* 2010, **11:**R85.
734 27. Maheswari U, Mock T, Armbrust EV, Bowler C: **Update of the Diatom EST**
735 **Database: a new tool for digital transcriptomics.** *Nucleic Acids Res* 2009,
736 **37:**D1001-1005.
737 28. Kaur S, Spillane C: **Reduction in carotenoid levels in the marine diatom**
738 **Phaeodactylum tricornutum by artificial microRNAs targeted against the**
739 **endogenous phytoene synthase gene.** *Mar Biotechnol (NY)* 2015, **17:**1-7.
740 29. Diner RE, Bielinski VA, Dupont CL, Allen AE, Weyman PD: **Refinement of the**
741 **Diatom Episome Maintenance Sequence and Improvement of Conjugation-**
742 **Based DNA Delivery Methods.** *Front Bioeng Biotechnol* 2016, **4:**65.
743 30. Nymark M, Sharma AK, Sparstad T, Bones AM, Winge P: **A CRISPR/Cas9**
744 **system adapted for gene editing in marine algae.** *Sci Rep* 2016, **6:**24951.
745 31. Rastogi A, Murik O, Bowler C, Tirichine L: **PhytoCRISP-Ex: a web-based and**
746 **stand-alone application to find specific target sequences for CRISPR/CAS**
747 **editing.** *BMC Bioinformatics* 2016, **17:**261.
748 32. Veluchamy A, Lin X, Maumus F, Rivarola M, Bhavsar J, Creasy T, O'Brien K,
749 Sengamalay NA, Tallon LJ, Smith AD, et al: **Insights into the role of DNA**
750 **methylation in diatoms by genome-wide profiling in Phaeodactylum**
751 **tricornutum.** *Nat Commun* 2013, **4**.
752 33. Veluchamy A, Rastogi A, Lin X, Lombard B, Murik O, Thomas Y, Dingli F,
753 Rivarola M, Ott S, Liu X, et al: **An integrative analysis of post-translational**
754 **histone modifications in the marine diatom Phaeodactylum tricornutum.**
755 *Genome Biol* 2015, **16:**102.
756 34. Daboussi F, Leduc S, Marechal A, Dubois G, Guyot V, Perez-Michaut C, Amato
757 A, Falciatore A, Juillerat A, Beurdeley M, et al: **Genome engineering**

758 **empowers the diatom Phaeodactylum tricornutum for biotechnology.** *Nat*
759 *Commun* 2014, **5:**3831.
760 35. Serif M, Dubois G, Finoux AL, Teste MA, Jallet D, Daboussi F: **One-step**
761 **generation of multiple gene knock-outs in the diatom Phaeodactylum**
762 **tricornutum by DNA-free genome editing.** *Nat Commun* 2018, **9:**3924.
763 36. Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B,
764 Nelson DR, Jijakli K, Abdrabu R, Harris EH, et al: **Whole-Genome**
765 **Resequencing Reveals Extensive Natural Variation in the Model Green Alga**
766 **Chlamydomonas reinhardtii.** *Plant Cell* 2015, **27:**2353-2369.
767 37. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz
768 C, Stegle O, Lippert C, et al: **Whole-genome sequencing of multiple**
769 **Arabidopsis thaliana populations.** *Nat Genet* 2011, **43:**956-963.
770 38. Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP,
771 Roberts IN, Burt A, Koufopanou V, et al: **Population genomics of domestic**
772 **and wild yeasts.** *Nature* 2009, **458:**337-341.
773 39. Lachance J, Tishkoff SA: **Population Genomics of Human Adaptation.** *Annu*
774 *Rev Ecol Evol Syst* 2013, **44:**123-143.
775 40. Godhe A, Rynearson T: **The role of intraspecific variation in the ecological**
776 **and evolutionary success of diatoms in changing environments.** *Philos Trans*
777 *R Soc Lond B Biol Sci* 2017, **372**.
778 41. Harnstrom K, Ellegaard M, Andersen TJ, Godhe A: **Hundred years of genetic**
779 **structure in a sediment revived diatom population.** *Proc Natl Acad Sci U S A*
780 2011, **108:**4252-4257.
781 42. Whittaker KA, Rynearson TA: **Evidence for environmental and ecological**
782 **selection in a microbe with no geographic limits to gene flow.** *Proc Natl*
783 *Acad Sci U S A* 2017, **114:**2651-2656.
784 43. Rengefors K, Kremp A, Thorsten BH, Reusch A, Wood M: **Genetic diversity**
785 **and evolution in eukaryotic phytoplankton: revelations from population**
786 **genetic studies.** *Journal of plankton research* 2017, **39:**165-179.
787 44. Matuszewski S, Hermisson J, Kopp M: **Catch Me if You Can: Adaptation from**
788 **Standing Genetic Variation to a Moving Phenotypic Optimum.** *Genetics*
789 2015, **200:**1255-1274.
790 45. Bailleul B, Berne N, Murik O, Petroutsos D, Prihoda J, Tanaka A, Villanova V,
791 Bligny R, Flori S, Falconet D, et al: **Energetic coupling between plastids and**
792 **mitochondria drives CO2 assimilation in diatoms.** *Nature* 2015, **524:**366-369.
793 46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A,
794 Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome**
795 **Analysis Toolkit: a MapReduce framework for analyzing next-generation**
796 **DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.
797 47. Chen G, Rynearson TA: **Genetically distinct populations of a diatom co-exist**
798 **during the North Atlantic spring bloom.** *LIMNOLOGY and OCEANOGRAPHY*
799 2016**:**2165-2179.
800 48. Blanc-Mathieu R, Krasovec M, Hebrard M, Yau S, Desgranges E, Martin J,
801 Schackwitz W, Kuo A, Salin G, Donnadieu C, et al: **Population genomics of**
802 **picophytoplankton unveils novel chromosome hypervariability.** *Sci Adv*
803 2017, **3:**e1700239.

804 49. Tsai IJ, Bensasson D, Burt A, Koufopanou V: **Population genomics of the wild**
805 **yeast Saccharomyces paradoxus: Quantifying the life cycle.** *Proc Natl Acad*
806 *Sci U S A* 2008, **105:**4957-4962.
807 50. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of**
808 **ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
809 51. Piganeau G, Eyre-Walker A, Jancek S, Grimsley N, Moreau H: **How and why**
810 **DNA barcodes underestimate the diversity of microbial eukaryotes.** *PLoS*
811 *One* 2011, **6:**e16342.
812 52. de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J: **Molecular evidence**
813 **of cryptic speciation in planktonic foraminifers and their relation to oceanic**
814 **provinces.** *Proc Natl Acad Sci U S A* 1999, **96:**2864-2868.
815 53. Saez AG, Probert I, Geisen M, Quinn P, Young JR, Medlin LK: **Pseudo-cryptic**
816 **speciation in coccolithophores.** *Proc Natl Acad Sci U S A* 2003, **100:**7163-
817 7168.
818 54. Wolf M, Chen S, Song J, Ankenbrand M, Muller T: **Compensatory base**
819 **changes in ITS2 secondary structures correlate with the biological species**
820 **concept despite intragenomic variability in ITS2 sequences--a proof of**
821 **concept.** *PLoS One* 2013, **8:**e66726.
822 55. Kaczmarska I, Mather L, Luddington I, Muise F, Ehrman J: **Cryptic diversity in**
823 **a cosmopolitan diatom known as Asterionellopsis glacialis (Fragilariaceae):**
824 **Implications for ecology, biogeography, and taxonomy.** *American Journal of*
825 *Botany* 2014.
826 56. Amato A, Kooistra WH, Ghiron JH, Mann DG, Proschold T, Montresor M:
827 **Reproductive isolation among sympatric cryptic species in marine diatoms.**
828 *Protist* 2007, **158:**193-207.
829 57. Medlin LK: **A timescale for diatom evolution based on four molecular**
830 **markers: reassessment of ghost lineages and major steps defining diatom**
831 **evolution.** *Vie Milieu / Life & Environment* 2015.
832 58. Bitarello BD, de Filippo C, Teixeira JC, Schmidt JM, Kleinert P, Meyer D,
833 Andres AM: **Signatures of Long-Term Balancing Selection in Human**
834 **Genomes.** *Genome Biol Evol* 2018, **10:**939-955.
835 59. Booker TR, Jackson BC, Keightley PD: **Detecting positive selection in the**
836 **genome.** *BMC Biol* 2017, **15:**98.
837 60. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ,
838 Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al: **A scan for positively**
839 **selected genes in the genomes of humans and chimpanzees.** *PLoS Biol* 2005,
840 **3:**e170.
841 61. Frost SD, Gunthard HF, Wong JK, Havlir D, Richman DD, Leigh Brown AJ:
842 **Evidence for positive selection driving the evolution of HIV-1 env under**
843 **potent antiviral therapy.** *Virology* 2001, **284:**250-258.
844 62. Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta
845 CJ, Farrar J, Dougan G, Achtman M: **Evolutionary history of Salmonella typhi.**
846 *Science* 2006, **314:**1301-1304.
847 63. Rastogi A, Maheswari U, Dorrell RG, Vieira FRJ, Maumus F, Kustka A,
848 McCarthy J, Allen AE, Kersey P, Bowler C, Tirichine L: **Integrative analysis of**
849 **large scale transcriptome data draws a comprehensive landscape of**

850 **Phaeodactylum tricornutum genome and evolutionary origin of diatoms.** *Sci*
851 *Rep* 2018, **8:**4834.
852 64. Lescot M, Hingamp P, Kojima KK, Villar E, Romac S, Veluchamy A, Boccara M,
853 Jaillon O, Iudicone D, Bowler C, et al: **Reverse transcriptase genes are highly**
854 **abundant and transcriptionally active in marine plankton assemblages.**
855 *ISME J* 2016, **10:**1134-1146.
856 65. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH: **Role of duplicate**
857 **genes in genetic robustness against null mutations.** *Nature* 2003, **421:**63-66.
858 66. Ogura T, Wilkinson AJ: **AAA+ superfamily ATPases: common structure--**
859 **diverse function.** *Genes Cells* 2001, **6:**575-597.
860 67. Zeytuni N, Zarivach R: **Structural and functional discussion of the tetra-trico-**
861 **peptide repeat, a protein interaction module.** *Structure* 2012, **20:**397-405.
862 68. Helliwell KE, Wheeler GL, Leptos KC, Goldstein RE, Smith AG: **Insights into the**
863 **evolution of vitamin B12 auxotrophy from sequenced algal genomes.** *Mol*
864 *Biol Evol* 2011, **28:**2921-2933.
865 69. Bertrand EM, Allen AE, Dupont CL, Norden-Krichmar TM, Bai J, Valas RE, Saito
866 MA: **Influence of cobalamin scarcity on diatom molecular physiology and**
867 **identification of a cobalamin acquisition protein.** *Proc Natl Acad Sci U S A*
868 2012, **109:**E1762-1771.
869 70. Helliwell KE, Collins S, Kazamia E, Purton S, Wheeler GL, Smith AG:
870 **Fundamental shift in vitamin B12 eco-physiology of a model alga**
871 **demonstrated by experimental evolution.** *ISME J* 2015, **9:**1446-1455.
872 71. Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA,
873 Jeddeloh JA, Colot V: **The Arabidopsis thaliana mobilome and its impact at**
874 **the species level.** *Elife* 2016, **5**.
875 72. Bonchev G, Parisod C: **Transposable elements and microevolutionary**
876 **changes in natural populations.** *Mol Ecol Resour* 2013, **13:**765-775.
877 73. Thiel MaG, L.: **The ecology of rafting in the marine environment. II. The**
878 **rafting organisms and community.** *Oceanography and Marine Biology: An*
879 *Annual Review* 2005, **43:**279-418.
880 74. Nikula R, Spencer HG, Waters JM: **Passive rafting is a powerful driver of**
881 **transoceanic gene flow.** *Biol Lett* 2013, **9:**20120821.
882 75. Schlichting HE: **The rôle of waterfowl in the dispersal of algae.** *Trans Am*
883 *Microsc Soc* 1960, **79:**160-166.
884 76. Proctor VW: **Dispersal of Desmids by birds.** *Phycologia* 1966, **5:**227-232.
885 77. Foissner W: **Biogeography and Dispersal of Micro-organisms: A Review**
886 **Emphasizing Protists.** *Acta Protozool* 2006, **45:**111-136.
887 78. Liti G: **The fascinating and secret wild life of the budding yeast S. cerevisiae.**
888 *Elife* 2015, **4**.
889 79. Blanc-Mathieu R, Verhelst B, Derelle E, Rombauts S, Bouget FY, Carre I,
890 Chateau A, Eyre-Walker A, Grimsley N, Moreau H, et al: **An improved**
891 **genome of the model marine alga Ostreococcus tauri unfolds by assessing**
892 **Illumina de novo assemblies.** *BMC Genomics* 2014, **15:**1103.
893 80. Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S,
894 Zeng Q, Zisson E, Wang JM, Greenberg JM, et al: **Genetic and phenotypic**
895 **intra-species variation in Candida albicans.** *Genome Res* 2015, **25:**413-425.

896 81. Filatov DA: **Extreme Lewontin's Paradox in Ubiquitous Marine**
897 **Phytoplankton Species.** *Mol Biol Evol* 2019, **36:**4-14.
898 82. Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, Salamov A,
899 Sanges R, Toseland A, Ward BJ, et al: **Evolutionary genomics of the cold-**
900 **adapted diatom Fragilariopsis cylindrus.** *Nature* 2017, **541:**536-540.
901 83. Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J,
902 Monier A, Salamov A, et al: **Pan genome of the phytoplankton Emiliania**
903 **underpins its global distribution.** *Nature* 2013, **499:**209-213.
904 84. Koester JA, Berthiaume CT, Hiranuma N, Parker MS, Iverson V, Morales R,
905 Ruzzo WL, Armbrust EV: **Sexual ancestors generated an obligate asexual and**
906 **globally dispersed clone within the model diatom species Thalassiosira**
907 **pseudonana.** *Sci Rep* 2018, **8:**10492.
908 85. Patil S, Moeys S, von Dassow P, Huysman MJ, Mapleson D, De Veylder L,
909 Sanges R, Vyverman W, Montresor M, Ferrante MI: **Identification of the**
910 **meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and**
911 **RAD51 homologs in the sexual species Pseudo-nitzschia multistriata and**
912 **Seminavis robusta.** *BMC Genomics* 2015, **16:**930.
913 86. Allen DE, Lynch M: **The effect of variable frequency of sexual reproduction**
914 **on the genetic structure of natural populations of a cyclical parthenogen.**
915 *Evolution* 2012, **66:**919-926.
916 87. Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D:
917 **Spontaneous epigenetic variation in the Arabidopsis thaliana methylome.**
918 *Nature* 2011, **480:**245-249.
919 88. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S,
920 Dominiczak A, Morris A, Porteous D, Smith B, et al: **Timing, rates and spectra**
921 **of human germline mutation.** *Nat Genet* 2016, **48:**126-133.
922 89. Shoemaker R, Deng J, Wang W, Zhang K: **Allele-specific methylation is**
923 **prevalent and is contributed by CpG-SNPs in the human genome.** *Genome*
924 *Res* 2010, **20:**883-889.
925 90. Huff JT, Zilberman D: **Dnmt1-independent CG methylation contributes to**
926 **nucleosome positioning in diverse eukaryotes.** *Cell* 2014, **156:**1286-1297.
927 91. Kanai Y, Ushijima S, Tsuda H, Sakamoto M, Hirohashi S: **Aberrant DNA**
928 **methylation precedes loss of heterozygosity on chromosome 16 in chronic**
929 **hepatitis and liver cirrhosis.** *Cancer Lett* 2000, **148:**73-80.
930 92. Clark DP, Flynn KJ, Ownes NJ: **The large capacity for dark nitrate-assimilation**
931 **in diatoms may overcome nitrate limitation of growth.** *New Phytologist*
932 2002.
933 93. Ivanikova NV, McKay R, Bullerjahn GS: **Construction and characterization of a**
934 **cyanobacterial bioreporter capable of assessing nitrate assimilatory**
935 **capacity in freshwaters.** *Limnology and Oceanography* 2005, **3:**86-93.
936 94. Weiguo Li JW: **Influence of light and nitrate assimilation on the growth**
937 **strategy in clonal weed Eichhornia crassipes.** *Aquatic Ecology* 2011.
938 95. Pickett-Heaps JD, Forer A: **Pac-Man does not resolve the enduring problem**
939 **of anaphase chromosome movement.** *Protoplasma* 2001, **215:**16-20.
940 96. Doyle JJaJLD: **A rapid DNA isolation procedure for small quantities of fresh**
941 **leaf tissue.** *Phytoch Bull* 1987, **19** 11-15.

97. 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.
98. 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.
99. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.** *Mol Biol Evol* 2016, **33:**1870-1874.
100. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL: **ViennaRNA Package 2.0.** *Algorithms Mol Biol* 2011, **6:**26.
101. 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.
102. Auton A, McVean G: **Recombination rate estimation in the presence of hotspots.** *Genome Res* 2007, **17:**1219-1227.
103. Etherington GJ, Dicks J, Roberts IN: **Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination.** *Bioinformatics* 2005, **21:**278-281.
104. Weir BS, Cockerham CC: **Estimating F-Statistics for the Analysis of Population Structure.** *Evolution* 1984, **38:**1358-1370.
105. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559-575.
106. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27:**2156-2158.
107. Alexander DH, Lange K: **Enhancements to the ADMIXTURE algorithm for individual ancestry estimation.** *BMC Bioinformatics* 2011, **12:**246.
108. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **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.
109. 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.
110. Cheong WH, Tan YC, Yap SJ, Ng KP: **ClicO FS: an interactive web-based service of Circos.** *Bioinformatics* 2015, **31:**3685-3687.
111. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
112. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29:**24-26.

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 makeup of the genomes of each accession. Based on the allelic composition of the ten genomes, six genomic clusters (K=6) are formed, which are distributed across individual accession genome in different proportions (represented by 6 different colors in the pie-chart). (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 exhibit a loss or multiple copies (CNV) within one or more accessions. (C) The bar plots represent the number of total- and accession-specific TEs exhibiting CNVs across one or more accessions. (D) With similar principle aesthetics as panel (A) of this figure, the heat-map shows the patterns of log2FC only across all the accessions of those TEs exhibiting CNV in at least one of the ten accessions studied.

Figure 4. Evolutionary and functional consequences of polymorphisms. (A) The bar plot represents total and specific numbers of genes that are subject to positive selection, or experiencing loss-of-function (LoF) mutations. For each category, the accessions are plotted as stack plots with total and specific numbers of genes. Numbers of genes in each category are indicated. (B) The box plot represents the number of gene families 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.

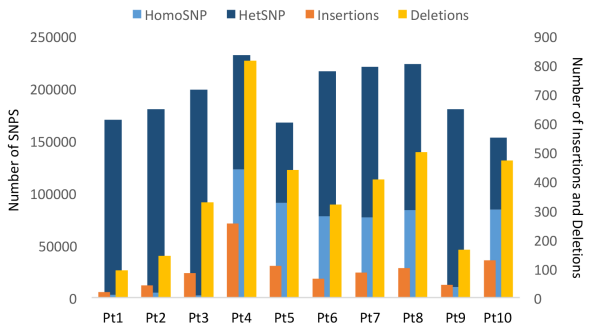
Figure 5. Relaxed selection within each genetic clade. Based on the EfR metric, the network displays highly affected gene families experiencing balancing selection. Gene families associated with *MetH* gene in all the accessions are indicated within the blue circles. The red circles group individual accessions as clades.

Figure 6. Selection of MetH-facilitated methionine biosynthesis over MetE. The bar plots represent relative expression of (A) *MetH*, (B) *MetE*, (C) *CBA1* and (D) *SHMT* genes in four (Pt2, Pt3, Pt4 and Pt8) of the ten accessions with the presence of vitamin B12 in axenic cultures (light gray), and with natural bacteria and no vitamin B12 in the growing media (black).

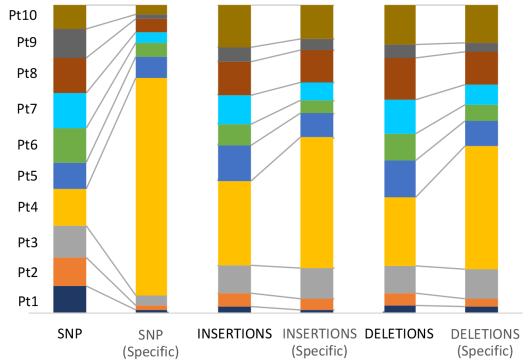
Figure 7. Population structure of the ecotypes. The color gradient from yellow to blue indicates low to high numerical values across each ecotype (indicated on top X-axis of panel A) within different functional categories indicated on Y-axis. These functional categories include (from top to bottom), Year of sampling = Year 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

Library Name	Accession number/axenicity	Origin	Year of Isolation	Mapped Read-Pairs	% Mapped Read-Pairs	Alignment Depth (X)	Genome Coverage (%)
Pt1	CCMP2561 (axenic)	Blackpool, UK	1956	3, 642,044	79.41	26.5	98.0
Pt2	CCMP2557 (xenic)	Plymouth, UK	Prior to 1910	6, 016,241	78.23	43.8	98.0
Pt3	CCMP2558 (axenic)	Plymouth, UK	1930s	6, 373,591	65.62	46.4	98.3
Pt4	CCMP2559 (xenic)	Finland	1951	15, 583,665	67.31	113.5	94.0
Pt5	CCMP630 (axenic)	West Dennis, MA, USA	1972	5, 346,009	75.50	38.9	93.2
Pt6	CCMP631 (xenic)	MA, USA	1956	3, 922,830	64.50	28.5	94.1
Pt7	CCMP1327 (axenic)	Long Island, NY, USA	1952	4, 937,516	67.30	35.9	94.9
Pt8	CCMP2560 (xenic)	Vancouver, Canada	1987	22, 235,170	78.36	162.1	94.4
Pt9	CCMP633 (axenic)	Guam, Micronesia	1981	7, 551,099	74.68	55.2	97.5
Pt10	CCMP2928 (axenic)	Dalian, China	2000	5, 436,057	72.59	39.6	92.1

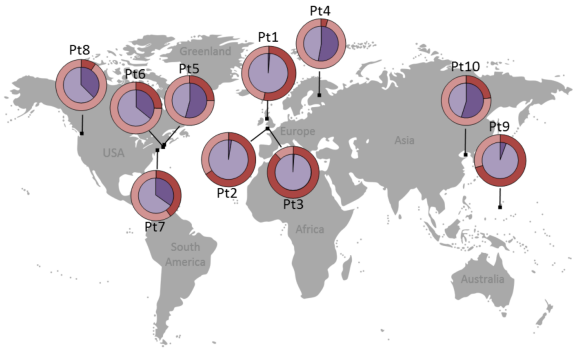
A


HomoSNP
HetSNP
Insertions
Deletions
Number of SNPS
Number of Insertions and Deletions
250000
200000
150000
100000
50000
0
900
800
700
600
500
400
300
200
100
0
Pt1
Pt2
Pt3
Pt4
Pt5
Pt6
Pt7
Pt8
Pt9
Pt10

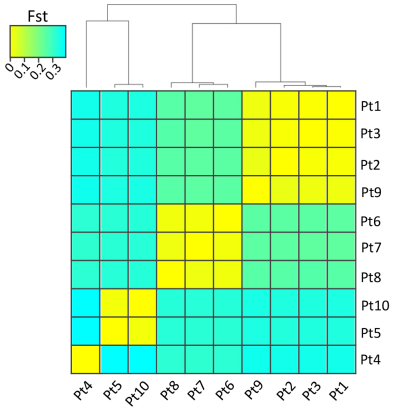
B

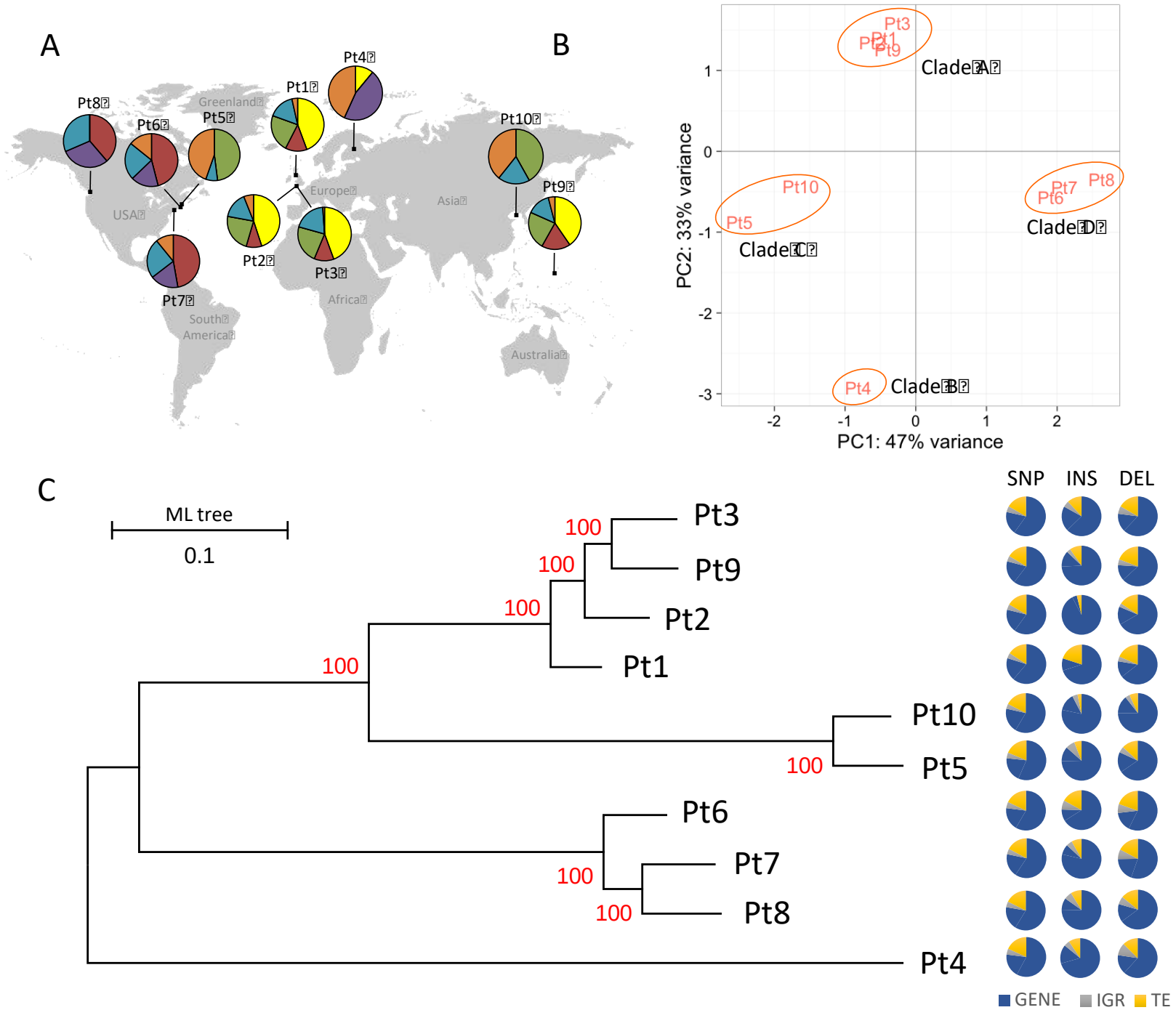

Pt10
Pt9
Pt8
Pt7
Pt6
Pt5
Pt4
Pt3
Pt2
Pt1
SNP
SNP (Specific)
INSERTIONS
INSERTIONS (Specific)
DELETIONS
DELETIONS (Specific)

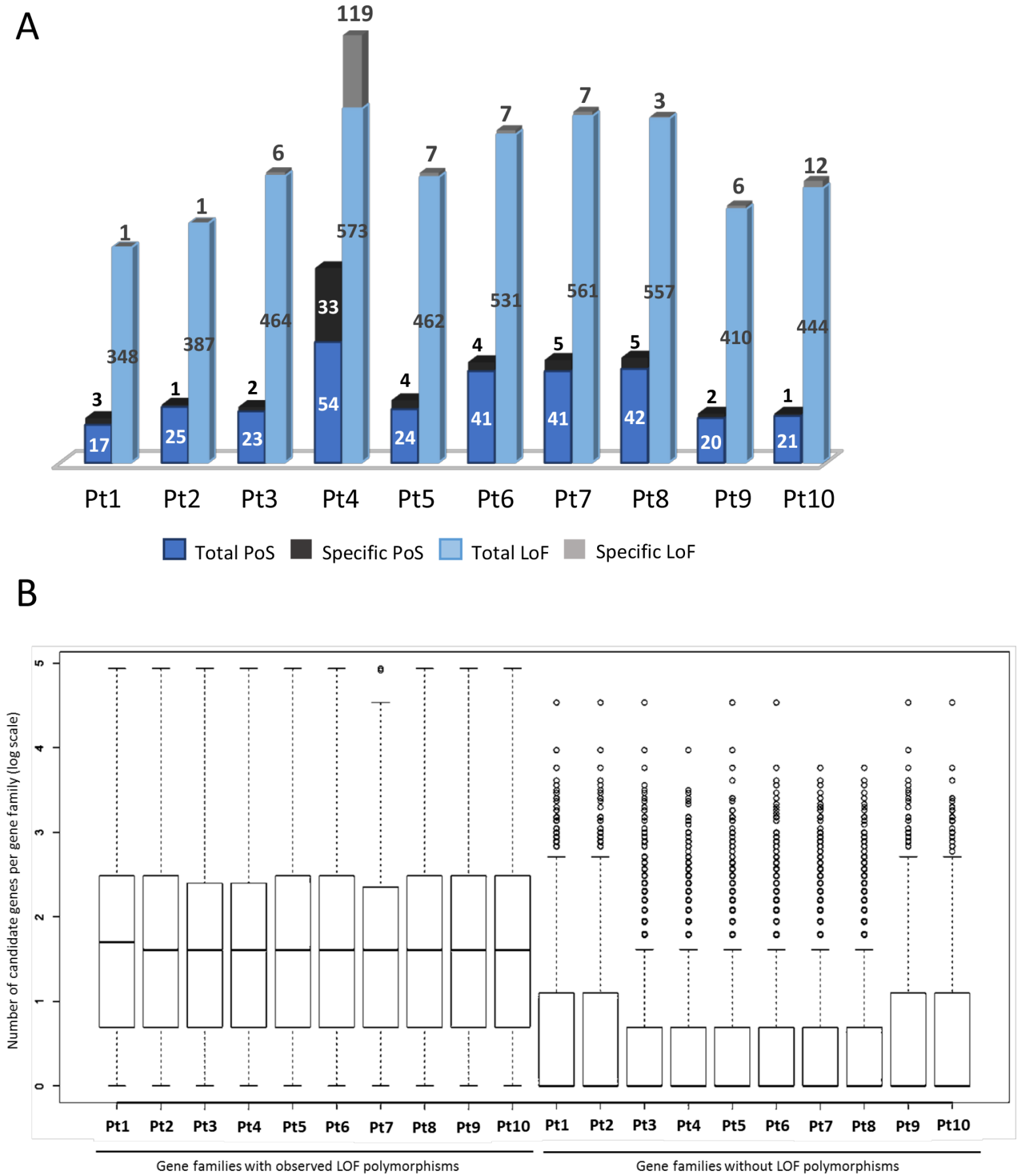
C

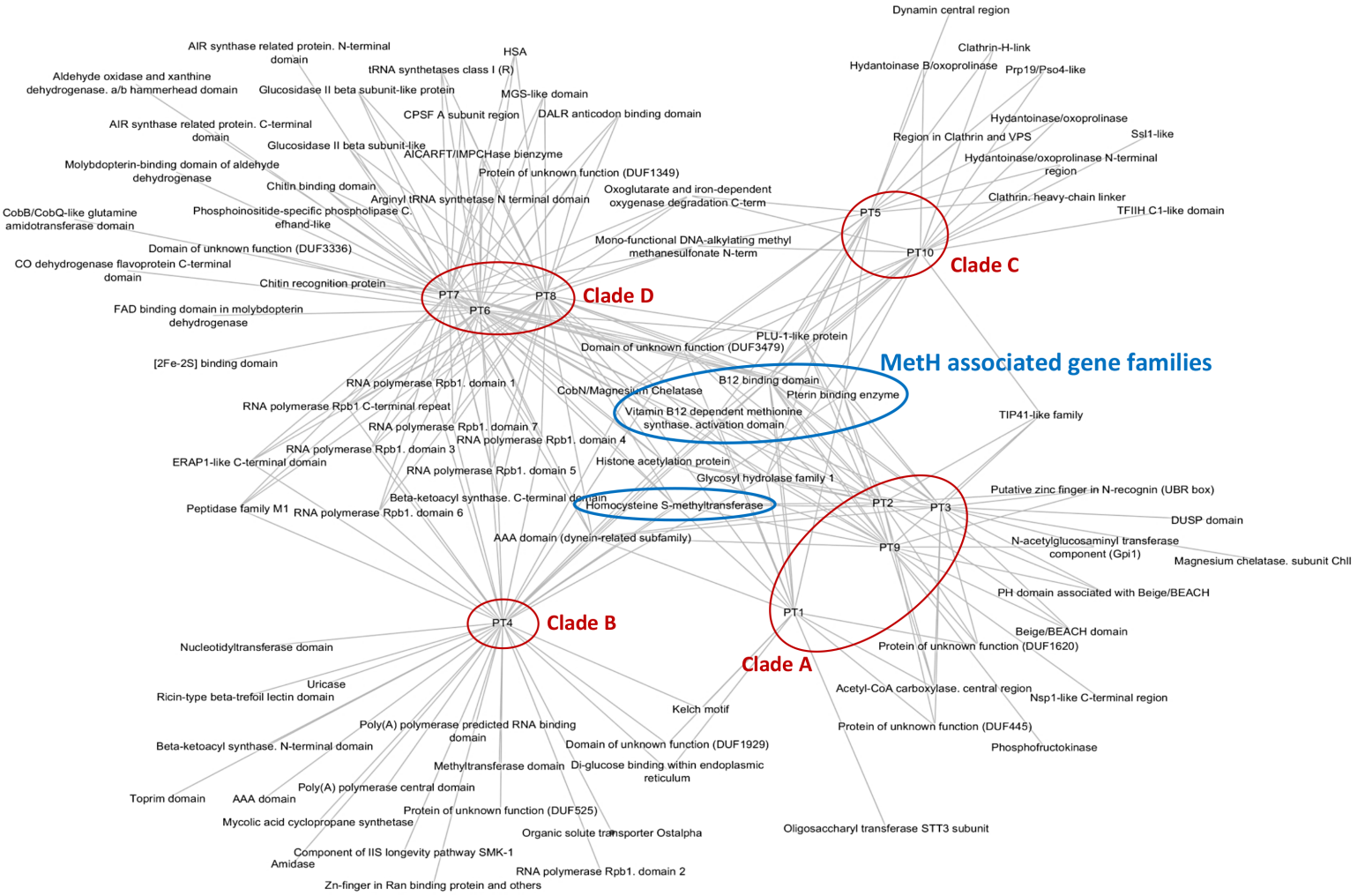

Pt8
Pt6
Pt5
Pt4
Pt1
Pt10
Pt9
Pt2
Pt3
Pt7
Greenland
USA
Europe
Asia
Africa
South America
Australia

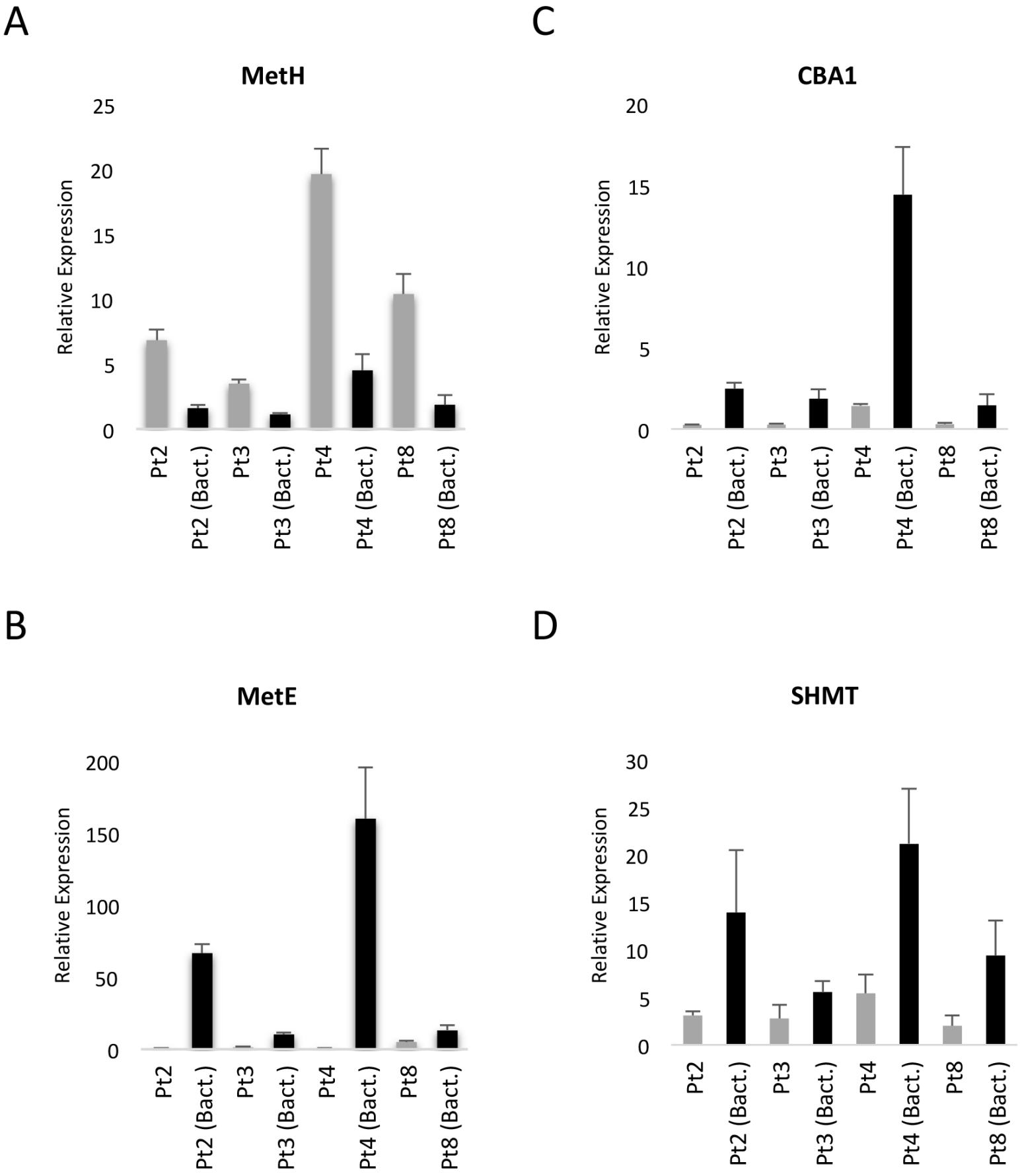
D

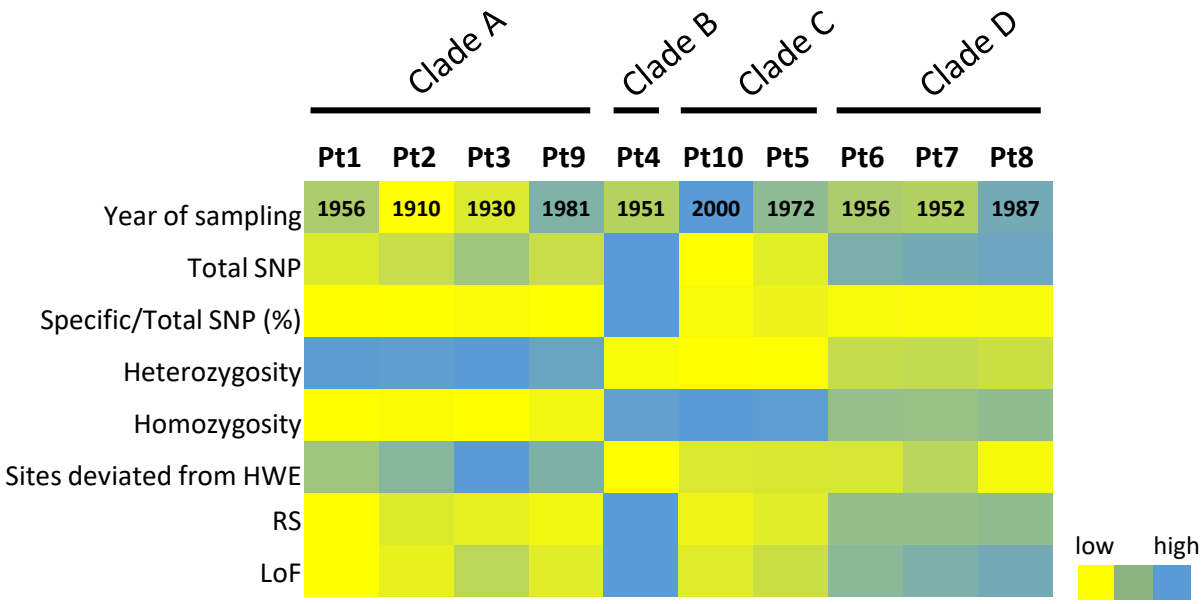

Fst
0
0.1
0.2
0.3
Pt1
Pt3
Pt2
Pt9
Pt6
Pt7
Pt8
Pt10
Pt5
Pt4
Pt4
Pt5
Pt10
Pt8
Pt7
Pt6
Pt9
Pt2
Pt3
Pt1


A
Pt8
Pt6
Pt5
Pt4
Pt1
Pt10
Pt9
Pt2
Pt3
Pt7
Greenland
USA
Europe
Asia
Africa
South America
Australia
B
Pt3
Pt1
Pt9
Clade A
Pt10
Pt5
Clade C
Pt7
Pt8
Pt6
Clade D
Pt4
Clade B
PC2: 33% variance
PC1: 47% variance
C
ML tree
0.1
100
Pt3
Pt9
Pt2
Pt1
Pt10
Pt5
Pt6
Pt7
Pt8
Pt4
SNP
INS
DEL
GENE
IGR
TE


A
119
1
1
6
7
7
7
3
6
12
573
348
387
464
33
462
531
561
557
410
444
3
1
2
4
4
5
5
2
1
17
25
23
54
24
41
41
42
20
21
Pt1
Pt2
Pt3
Pt4
Pt5
Pt6
Pt7
Pt8
Pt9
Pt10
Total PoS
Specific PoS
Total LoF
Specific LoF
B
Number of candidate genes per gene family (log scale)
0
1
2
3
4
5
Pt1
Pt2
Pt3
Pt4
Pt5
Pt6
Pt7
Pt8
Pt9
Pt10
Pt1
Pt2
Pt3
Pt4
Pt5
Pt6
Pt7
Pt8
Pt9
Pt10
Gene families with observed LOF polymorphisms
Gene families without LOF polymorphisms


Dynamin central region
Clathrin-H-link
Hydantoinase B/oxoprolinase
Prp19/Pso4-like
AIR synthase related protein. N-terminal domain
HSA
tRNA synthetases class I (R)
Aldehyde oxidase and xanthine dehydrogenase. a/b hammerhead domain
Glucosidase II beta subunit-like protein
MGS-like domain
CPSF A subunit region
DALR anticodon binding domain
AIR synthase related protein. C-terminal domain
Glucosidase II beta subunit-like
AICARFT/IMPCHase bienzyme
Hydantoinase/oxoprolinase
Region in Clathrin and VPS
Ssl1-like
Hydantoinase/oxoprolinase N-terminal region
Molybdopterin-binding domain of aldehyde dehydrogenase
Protein of unknown function (DUF1349)
Chitin binding domain
Arginyl tRNA synthetase N terminal domain
Oxoglutarate and iron-dependent oxygenase degradation C-term
Clathrin. heavy-chain linker
TFIIH C1-like domain
CobB/CobQ-like glutamine amidotransferase domain
Phosphoinositide-specific phospholipase C. efhand-like
PT5
PT10
Clade C
Domain of unknown function (DUF3336)
Mono-functional DNA-alkylating methyl methanesulfonate N-term
CO dehydrogenase flavoprotein C-terminal domain
Chitin recognition protein
PT7
PT8
PT6
Clade D
FAD binding domain in molybdopterin dehydrogenase
PLU-1-like protein
Domain of unknown function (DUF3479)
MetH associated gene families
[2Fe-2S] binding domain
RNA polymerase Rpb1. domain 1
B12 binding domain
CobN/Magnesium Chelatase
Pterin binding enzyme
RNA polymerase Rpb1 C-terminal repeat
Vitamin B12 dependent methionine synthase. activation domain
TIP41-like family
RNA polymerase Rpb1. domain 7
RNA polymerase Rpb1. domain 4
RNA polymerase Rpb1. domain 3
ERAP1-like C-terminal domain
Histone acetylation protein
RNA polymerase Rpb1. domain 5
Glycosyl hydrolase family 1
Putative zinc finger in N-recognin (UBR box)
Beta-ketoacyl synthase. C-terminal domain
Homocysteine S-methyltransferase
PT2
PT3
Peptidase family M1
RNA polymerase Rpb1. domain 6
DUSP domain
AAA domain (dynein-related subfamily)
PT9
N-acetylglucosaminyl transferase component (Gpi1)
Magnesium chelatase. subunit ChlI
PH domain associated with Beige/BEACH
PT1
PT4
Clade B
Beige/BEACH domain
Nucleotidyltransferase domain
Protein of unknown function (DUF1620)
Clade A
Uricase
Acetyl-CoA carboxylase. central region
Ricin-type beta-trefoil lectin domain
Nsp1-like C-terminal region
Kelch motif
Poly(A) polymerase predicted RNA binding domain
Protein of unknown function (DUF445)
Beta-ketoacyl synthase. N-terminal domain
Domain of unknown function (DUF1929)
Phosphofructokinase
Methyltransferase domain
Di-glucose binding within endoplasmic reticulum
Poly(A) polymerase central domain
Toprim domain
AAA domain
Protein of unknown function (DUF525)
Mycolic acid cyclopropane synthetase
Organic solute transporter Ostalpha
Oligosaccharyl transferase STT3 subunit
Component of IIS longevity pathway SMK-1
Amidase
RNA polymerase Rpb1. domain 2
Zn-finger in Ran binding protein and others


A
MetH
Relative Expression
25
20
15
10
5
0
Pt2
Pt2 (Bact.)
Pt3
Pt3 (Bact.)
Pt4
Pt4 (Bact.)
Pt8
Pt8 (Bact.)
C
CBA1
Relative Expression
20
15
10
5
0
Pt2
Pt2 (Bact.)
Pt3
Pt3 (Bact.)
Pt4
Pt4 (Bact.)
Pt8
Pt8 (Bact.)
B
MetE
Relative Expression
200
150
100
50
0
Pt2
Pt2 (Bact.)
Pt3
Pt3 (Bact.)
Pt4
Pt4 (Bact.)
Pt8
Pt8 (Bact.)
D
SHMT
Relative Expression
30
25
20
15
10
5
0
Pt2
Pt2 (Bact.)
Pt3
Pt3 (Bact.)
Pt4
Pt4 (Bact.)
Pt8
Pt8 (Bact.)


Clade A
Clade B
Clade C
Clade D
Pt1
Pt2
Pt3
Pt9
Pt4
Pt10
Pt5
Pt6
Pt7
Pt8
Year of sampling
1956
1910
1930
1981
1951
2000
1972
1956
1952
1987
Total SNP
Specific/Total SNP (%)
Heterozygosity
Homozygosity
Sites deviated from HWE
RS
LoF
low
high