Strong population genomic structure of the toxic dinoflagellate Alexandrium minutum inferred from meta-transcriptome samples

Mickael Le Gac, Lou Mary, Gabriel Metegnier, Julien Quéré, Raffaele Siano, Francisco Rodríguez, Christophe Destombe, Marc Sourisseau

MOLECULAR ECOLOGY

Inferring population genetic structure of the toxic dinoflagellate Alexandrium minutum from meta-transcriptomic samples

Journal:	*Molecular Ecology*
Manuscript ID	MEC-22-0590
Manuscript Type:	Original Article
Date Submitted by the Author:	01-Jun-2022
Complete List of Authors:	Le Gac, Mickael; Ifremer, ODE/Dyneco/Pelagos Mary, Lou; Ifremer, ODE/Dyneco/Pelagos Metegnier, Gabriel; Ifremer, ODE/Dyneco/Pelagos Quéré, Julien; Ifremer, ODE/Dyneco/Pelagos Siano, Raffaele; Ifremer, ODE/Dyneco/Pelagos Rodríguez Hernández , Francisco; Instituto Español de Oceanografía Centro Oceanográfico de Vigo Destombe, Christophe Sourisseau, Marc; Ifremer, ODE/Dyneco/Pelagos
Keywords:	Adaptation, Population Genetics - Empirical, Protists, Metagenomics, Transcriptomics

Inferring population genetic structure of the toxic dinoflagellate *Alexandrium minutum* from meta-transcriptomic samples

Mickael le Gac[1*], Lou Mary[1], Gabriel Metegnier[1], Julien Quéré[1], Raffaele Siano[1], Francisco Rodríguez[2], Christophe Destombe[3] and Marc Sourisseau[1]

[1] Ifremer, Dyneco, Plouzané, France

[2] Instituto Español de Oceanografía (IEO, CSIC). Vigo, Spain

[3] Station Biologique de Roscoff, IRL 3614, CNRS, Sorbonne Université, Roscoff, France

Corresponding author: Mickael.Le.Gac@ifremer.fr

Running title: A. minutum metaT population genomics

Abstract

Marine microeukaryote population structures are difficult to study. These organisms display huge census population sizes, yet genotyping usually requires clonal strains originating from single cells, hindering proper population sampling. Estimating allelic frequency directly from population wide samples, without any isolation step, offers an interesting alternative. Here we investigated the use of metatranscriptomic (metaT) environmental samples to determine the population genetic structure of the dinoflagellate *Alexandrium minutum*. Using a resampling approach and comparing strain and environmental metaT datasets, we showed that metaT samples enabled an unbiased and precise estimation of allelic frequencies when coverage was higher than 1e+06 reads for this species. Strain and metaT based results both indicated a strong genetic structure for *A. minutum* in Western Europe. The presence of numerous private alleles, and even fixed polymorphism, would indicate ancient divergence and absence of gene flow between populations. Single Nucleotide Polymorphisms (SNPs) displaying strong allele frequency differences were distributed throughout the genome, which might indicate the importance of selection from standing genetic variation (soft selective sweeps). However, a few genomic regions displayed extremely low diversity that could result from the fixation of adaptive *de novo* mutations (hard selective sweeps) within the populations. Altogether, this work demonstrated that metaT samples may be used to elucidate the evolutionary processes underlying the complex population structure of marine microeukaryotes.

Keywords: Dinoflagellate, Population Genomics, Harmful Algal Blooms, meta-transcriptomic

Introduction

Microbial census population sizes are often huge and difficult to apprehend. This is especially the case for pelagic marine microbes. For instance, the smallest and most abundant photosynthetic organism on the planet, the cyanobacteria *Prochlorococcus* has a census population size estimated to be around 10^{27} cells (Biller et al., 2015). Even without going to such extreme values, numerous marine protists (a polyphyletic group of organisms encompassing all unicellular eukaryotes) may easily reach cell densities of thousands of cells per liter, representing several hundreds of billions of cells in a small bay. With such population sizes, adequate sampling to investigate population genetic diversity and structure from individual cells is a non-trivial problem. A problem amplified by the need to isolate cells from natural samples and initiate clonal cultures to obtain sufficient genetic material for genotyping. With such constraints, even for species relatively easy to cultivate, genotyping tens of clones per natural population quickly becomes a daunting task and working on difficult to cultivate species is impossible. Strategies based on single cell genotyping might be an alternative but there are technical and financial hurdles preventing their widespread and high throughput development.

However, community wide samples, containing the genetic material of several thousands and even millions of protist cells may easily be obtained by filtering a few liters of water. Such environmental DNA (eDNA) or RNA (eRNA) samples are classically used to infer community composition following targeted sequencing (metabarcoding) or to investigate functional genomic aspects of natural communities using non-targeted sequencing (metagenomic or metatranscriptomic). Here, we explore the use of metatranscriptomic (metaT) datasets to infer protist population genetics for species of interest. The metaT datasets are composed of mRNA sequences extracted from natural communities and give access to a wide range of Single Nucleotide Polymorphism (SNPs) markers. *A priori*, the main advantages of such approaches are the relative simplicity of the sample processing (filtering water instead of isolation and

cultivation steps) combined with the ability to infer population allelic frequencies from samples containing thousands of cells. Another non-negligible advantage is the absence of targeted sequencing, enabling reuse of previously obtained datasets. This may be especially interesting as metaT datasets obtained worldwide are quickly accumulating (Alexander, Jenkins, et al., 2015; Alexander, Rouco, et al., 2015; Carradec et al., 2018; Cohen et al., 2017; Gong et al., 2018; Hu et al., 2018; Ji et al., 2018; Lambert et al., 2021; Lampe et al., 2018; Marchetti et al., 2012; Metegnier et al., 2020; Wurch et al., 2019; Zhang et al., 2019). The main drawback of such approach, but also of all potential approaches based on eDNA/RNA, is that it relies on the direct estimation of allelic frequencies in a population, without obtaining individual genotypes, limiting the type of possible analyses. Moreover, population genetic analyses from metaT datasets may only be considered if the mapping of the environmental reads is specific and the estimated allelic frequencies are precise and unbiased. Mapping environmental reads to a metareference composed of the reference transcriptomes of several hundreds of species has previously been shown to be a good way to ensure specific mapping (Metegnier et al., 2020). For marine protists, this reference corresponds to the concatenation of the species specific reference transcriptomes obtained from the Marine Micro-eukaryote Transcriptome Project (Keeling et al., 2014). Quantifying the precision of allelic frequency estimates in relation to sequencing depth is especially critical. Indeed, in a metaT sample, the sequencing depth of a given species strongly depends on the relative abundance of this species in a community at the time of sampling, and the sequencing depth of a given gene strongly depends on the relative expression of this gene. Using a resampling approach, we explore how allele frequency estimates are impacted by sequencing depth. Using metaT datasets, allelic frequencies are based on mRNA, not DNA. This may also be a critical point if differential gene expression tends to bias the estimated allelic frequencies, for instance in case of allele specific gene expression. This aspect is explored by comparing allelic frequencies inferred after either genotyping individual cells or pooling RNAseq reads obtained from these same strains. The former enabled to determine the true allelic frequency of the population of

cultivated strains, while the latter included any potential bias due to differential gene expression.

The dinoflagellate *Alexandrium minutum* is responsible for harmful algal blooms that may transiently dominate local micro-eukaryote coastal communities (Lewis et al., 2018). During these events cell densities may reach several million of cells per liter (Garcés et al., 2004). This species produces Paralytic Shellfish Toxins (PSTs) that tend to bioaccumulate in the trophic network and more precisely in shellfish with potential sanitary and socio-economic impacts (Ben-Gigirey et al., 2020; Nogueira et al., 2022). Using 18S and ITS markers, two main clades were identified, one cosmopolite clade and the other restricted to the Southern Pacific (Lilly et al., 2005; McCauley et al., 2009). Within the cosmopolite clade, a global geographic structure was suggested by microsatellites (McCauley et al., 2009). Still using microsatellites but at a more local scale, a strong genetic structure was identified in the Mediterranean Sea. This structuration pattern was at least partly compatible with hydrodynamics (Casabianca et al., 2012). In the South-West of the English Channel a moderate spatio-temporal structure was identified across a couple of years among two estuaries less than 200 km apart (Dia et al., 2014). Finally, two highly divergent populations, that may be considered as cryptic species, were identified in Noth-Western Europe using a SNP based approach (Le Gac et al., 2016c). Here taking as a model system *A. minutum* populations from Western Europe, we: 1. Pooled and resampled RNAseq reads obtained from strains to quantify how differential gene expression may bias observed allelic frequencies and to estimate the relationship between coverage and allelic frequency precision. 2. Compared *A. minutum* genetic structure inferred from strains and metaT datasets. 3. Determined coding genome wide genetic diversity and divergence among *A. minutum* populations using metaT datasets.

Material and Methods

1. Strains

Each strain corresponded to a clonal culture initiated after micropipetting a single cell into fresh culture medium under an inverted microscope. These clonal strains are haploid. Origin of strains is indicated in Supplementary Table 1 and Figure 1. Out of the 37 strains used in the present study, 25 were isolated from water samples. The twelve others were isolated after germination of resting cysts from dated sediment cores, as indicated in Delebecq et al. (2020). Strains were grown to late exponential phase in K medium. Cultures were centrifuged at 4500 g for 8 min. RNA extraction occurred either directly after centrifugation, or cell pellets were frozen into liquid nitrogen with RNA Later and stored at -80°C until RNA extraction

2. Metatranscriptomic samples

MetaT samples were obtained during *A. minutum* blooms in Western Europe from 2013 to 2018 by filtering water on 20 µm polycarbonate filters using a peristaltic pump. The filters were frozen into liquid nitrogen with RNA Later and stored at -80°C until RNA extraction. Information regarding the 77 metaT samples are indicated in Supplementary Table 2 and Figure 1.

3. RNA extraction and library preparation

Samples were ultra-sonicated on ice in extraction buffers. RNA was extracted using either the Qiagen Rneasy Plus Mini kit or NucleoSpin® RNA Plus kit (Macherey-Nagel) Kit (Supplementary Table 1 and 2). Library prepared with either the Illumina Truseq mRNA V2 kit or Illumina mRNA TruSeq stranded kit. Samples were sequenced at Get-PlaGe France Genomics sequencing platform (Toulouse, France) on Illumina HiSeq 2000/2500 2*100 pb or HiSeq 3000 2*150 pb. Raw sequencing reads are available in public databases (Metegnier et al. 2015, Le Gac et al. 2016a, ENAXXX)

4. Bioinformatic analyses

4.1. Trimming

Trimmomatic (V. 0.33) (Bolger et al., 2014) was used to trim ambiguous, low quality reads and sequencing adapters with parameters ILLUMINACLIP: Adapt.fasta:2:30:10:8:TRUE LEADING:3 TRAILING:3 MAXINFO:40:0.5 MINLEN:80 for 2*150 or MINLEN:60 for 2*100 reads.

4.2. Generating simulated population datasets from strains

Using bash scripts, 5e+06 reads were subsampled for each trimmed strain fastq files. They were concatenated to obtain a forward and a reverse fastq files simulating a population of strains. These files were subsampled in order to obtain 10 replicate files for seven coverage levels corresponding to a total of 1e+04, 5e+04, 1e+05, 5e+05, 1e+06, 5e+06 and 1e+07 reads. This was done separately for the 18 strains sequenced using 2*100bp reads and the 19 strains sequenced using 2*150bp reads (Supplementary Table 1).

4.3. Aligning reads

The strain as well as the simulated population samples were aligned to the *A. minutum* reference transcriptome previously developed and corresponding to 153,222 contigs (Le Gac et al., 2016b,c). The metaT datasets were aligned to a metareference corresponding to the concatenation of 313 species specific reference transcriptomes, representing 213 unique genus. It corresponded to the resources developed during the Marine Micro-eukaryote Transcriptome Project (Keeling et al., 2014) and also included the *A. minutum* reference transcriptome (Metegnier et al., 2020). Alignments were performed using BWA-MEM (Li, 2013). Only reads with a mapping score >10 were retained. Pairs for which the two reads did not map concordantly on the same transcript were removed. Samtools (Li et al., 2009) was used to sort and index bam files.

4.4. Identifying SNPs

A single nucleotide polymorphism (SNP) database was developed using the strain data following the approach proposed previously (Le Gac et al., 2016c). Briefly, FreeBayes (Garrison & Marth, 2012) was run twice using the 37 strains by enforcing haploidy and diploidy. As in culture conditions, *A. minutum* cells are in a vegetative, haploid stage, only SNPs detected using the haploidy enforced run and identified as haploid using the diploidy enforced run were considered. This was done to exclude potential intragenomic variability due to multicopy genes. The genotypes of each of the strains were obtained using vcftools (Danecek et al., 2011), for SNPs (excluding indels) displaying two alleles, a quality criterion >40, and covered more than 10 times in each of the 37 strains (no missing data). This database was composed of 227,829 SNPs.

For these same SNPs, the allelic frequencies of the simulated population samples and of the natural populations based on the metaT samples were obtained using FreeBayes (Garrison & Marth, 2012) enforcing diploidy, and extracting coverage for each of the two alleles at the various SNP positions.

In addition, for the metaT samples, *A. minutum* allelic frequencies were also estimated using a *de novo* approach, i.e. without restricting the analysis to the SNPs identified using the strains. It was performed for *A. minutum* contigs using Freebayes (Garrison & Marth, 2012) and vcftools (Danecek et al., 2011) by enforcing diploidy, retaining positions with quality criterion >40, two alleles, and excluding indels.

The MetaT samples were analyzed considering several coverage thresholds, representing a total of 18 datasets (Table 1). For each dataset, no missing value was allowed. They consisted of 6 main datasets obtained after filtering metaT samples based on *A.minutum* read coverage thresholds and SNPs based on site specific coverage thresholds. Each was analyzed in three ways: 1. by restricting the analyses to the SNPs previously identified using the strains, 2. by restricting the analyses after pruning (see below) SNPs previously identified using the strains

based on linkage disequilibrium 2. by using SNPs identified *de novo* from the metaT samples. Table 1 summarizes these datasets, indicating the minimum number of reads that must cover a given SNP (SNP coverage), the minimum number of reads aligning to *A. minutum* reference transcriptome (Min. reads), the number of metaT samples considered (Samples). The number of strain validated SNPs (Reference SNPs), the number strain validated SNPs after pruning (Pruned ref. SNPs), the number of SNPs identified directly from the metaT datasets (De novo SNPs) and the number of SNPs displaying one allele in strains from a given clade (NE_A, NE_B, Vigo) and the alternative allele in the two other clades (Diagnostic SNPs). In the Pooled_5P6 dataset, the metaT samples from the 5P6 dataset were pooled per geographic site. Only SNPs displaying a minimal allelic frequency > 0.05 in one sample were considered.

5. Population genomic analyses

5.1. Nucleotide divergence between strains

Nucleotide divergence among strains was calculated as the proportion of variable sites divided by the number of sites, only considering sites covered more than 10 times in all the strains (total of 15,103,704 sites). Strains were clustered using the function “hclust” as implemented in R. For each SNP, the allelic frequencies in the cultivated population of strains was calculated by dividing the number of strains with the reference allele by the total number of strains (Supplementary Figure 1, Strains).

5.2. Genetic differentiation

For each simulated population and metaT dataset, SNP coverage and reference allele read counts were extracted from VCF files. For each SNP, observed allelic frequencies were calculated by dividing the reference allele read counts by the coverage (Supplementary Figure 1, Simulated and metaT). The anova method implemented in the R package poolfstat (Hivert et al., 2018), with poolsize parameter set to 10,000, was used to compute pairwise Fst estimates: 1. To investigate the precision of Fst estimates and how it is affected by the level

of coverage, pairwise Fst were calculated among the 10 simulated populations for each overall coverage level (1e+04, 5e+04, 1e+05, 5e+05, 1e+06, 5e+06 and 1e+07 reads) and considering seven SNP coverage levels (All SNPs, SNPs covered by more than 5, 10, 20, 30, 50, and 100 reads). 2. To investigate the potential bias that could result from an estimation of allelic frequencies using RNA and not DNA (for instance due to allele specific differential expression in populations), Fst between the simulated populations and the actual allelic frequencies of the cultivated population of strains (see above) was calculated at the various overall and SNP specific coverage levels (see above). 3. To determine the potential bias that may result from using several sequencing approaches, five metaT samples were sequenced using both 2x100bp and 2x150bp approaches. Fst was calculated within each of the five sample pairs considering SNPs covered by more than 5, 10, 20, 30, 50, 100 reads. 4. To determine the structure of the natural *A. minutum* populations, pairwise Fst were calculated between metaT samples for each dataset (Table 1). For points 1, 2 and 3 above, Fst=0 are expected in case of absolute precision and total absence of bias.

5.3. Comparing genetic structure inferred using strains and metaT data

The pattern of genetic variability inferred from the 37 strains and 77 metaT samples was explored using a PCA approach, starting from the vcf files, as implemented in PLINK (Purcell et al., 2007). Linkage disequilibrium pruning was performed on the strain dataset, using 50kb windows (meaning that each transcript is analyzed as a whole), a 10bp window step size, keeping SNPs displaying a $R^2 < 0.1$.

5.4. Obtaining a folded joint allele frequency spectrum (JAFS) from metaT

From the three versions of the 5P6 dataset (strain validated SNPs, pruned and *de novo* SNPs, Table 1), for each SNP, reference allele count and coverage were summed per geographic site, leading to a single pooled sample per site (pooled 5P6 dataset, Table 1). For all the SNPs

247 covered more than 30 times in each pooled sample, the rare allele frequency was computed.
248 SNPs with a minimal allele frequency systematically < 0.05 per site were discarded. The
249 distribution of the rare allele frequency was calculated per site and the folded joint allele
250 frequency spectrum plotted for each of the three pairs of populations using the package hexbin
251 in R.

252 ### 5. Investigating coding regions genome wide divergence

253 The genetic linkage between contigs was established following an *A. minutum* linkage map
254 previously developed (Mary et al., Submitted). Using the pooled 5P6 datasets, for each SNP
255 in each geographic site, haplotype diversity was calculated as $1-\sum p_i^2$ where p_i is the
256 frequency of each of the two alleles. The R package pcadapt with "pool" option and k=2 was
257 used to identify SNPs displaying extreme allele frequency differences between populations
258 (Luu et al., 2017). The rollapply function from the zoo package implemented in R was used to
259 determine moving average in terms of haplotype diversity and number of significant SNPs
260 (from pcadapt) along the linkage groups. Values below the 0.5th and above the 99.5th
261 percentiles were identified.

262

263 # Results

264 ## 1. Nucleotide divergence between cultivated strains

265 Nucleotide divergence between the 37 monoclonal *A. minutum* strains was investigated at
266 227,829 SNP positions using mRNA sequences. Strain clustering based on the nucleotide
267 divergence indicated the occurrence of three diverging clusters (Figure 2). The first one,
268 hereafter named NE_A, was composed of 27 strains isolated from the Bay of Brest, Penzé
269 and Rance Estuary in France, as well as from Cork harbor in Ireland. Strains from this cluster
270 were isolated during algal blooms between 1989 and 2013, as well as after germination of

cysts preserved in sediment cores dated from 1947 to 2006. The second one, hereafter named NE_B, was composed of three strains isolated outside of blooming periods in 2010 and 2011 in the Bay of Concarneau and Brest. The third one was composed of seven strains isolated from the Bay of Vigo during a red tide in 2018 (Supplementary Table 1).

2. Fst precision and bias based on RNAseq data

The impact of coverage levels on the precision of allele frequency estimation resulting from population wide mRNA sequencing was quantified by pooling and subsampling the strain mRNA sequences at different coverage levels. Fst were quantified between replicate simulated populations. This was done separately for PE100 and PE150 datasets (Supplementary Table 1). Fst precision considerably improved with coverage levels. Below a coverage of 1e+05 reads, Fst estimation was imprecise, with Fst estimate between replicate simulated populations higher than 0.2. At a coverage of 5e+05, precision dropped to ~0.1, especially if the analysis was restricted to SNPs with a coverage > 5. At a coverage level > 5e+06, Fst precision was around 0.01 (Figure 3A and B).

To quantify the potential bias of inferring allele frequencies from mRNA and not DNA (for instance due to potential allele specific expression patterns), allele frequencies from the simulated populations were compared to the ones calculated following independent genotyping of the strains. Fst bias was systematically lower than 0.05 and often as low as 0.015 (Figure 3C and D).

To determine whether read length may influence Fst estimates, five metaT samples were sequenced using both 2x100bp and 2x150bp reads (Supplementary Figure 2). For each sample, Fst values between 2x100bp and 2x150bp read datasets were at time higher than 0.1. This indicated that the sequencing strategy may moderately influence Fst estimates. However, when calculated based on highly covered SNPs (>20), Fst were always below 0.05, indicating that such a potential issue may be solved by focusing on highly covered SNPs.

3. Genetic structure inferred from strains and metaT samples are similar

The genetic variability determined using strain and metaT samples was compared using a PCA (Figure 4). The first axis separated strains from the three clades identified above (Figure 4A; NE_A, NE_B and Vigo). The second and third axes separated the strains belonging to the NE_B clade and highlighted the high genetic variability existing within this clade (Figure 4A, B). All the metaT samples from the Bay of Brest were grouped with NE_A strains, indicating that they are composed of NE_A cells. All the metaT samples from the Bay of Vigo were grouped with the Vigo strains, indicating that they are composed of cells belonging to the Vigo clade. The only metaT sample from the Penzé estuary was relatively close to the Vigo samples (strains and metaT) along the first axis, and close to the NE_A strains and Bay of Brest metaT samples along the third axis. It is worth mentioning that some strains from the NE_A clade were isolated from the Penzé estuary from 1989 to 2010 (Figure 2), while the Penzé metaT sample was sampled in 2015 (Supplementary Table 2). The Penzé estuary metaT sample may either be composed of cells belonging to a fourth distinct population or result from admixture from two or three of the previously identified populations. To investigate this, allelic frequencies at diagnostic SNP positions (SNPs displaying one allele in all strains from a given clade (NE_A, NE_B, Vigo) and the alternative allele in the two other clades) were analyzed in all metaT samples (Supplementary Figure 3). For a great majority of these SNPs, the Penzé metaT sample displayed a fixed allele. Depending on the SNP, the allele corresponded to the one identified in NE_A, Vigo, and more rarely NE_B clade. The absence of intermediate allelic frequencies indicated a new population, from which no strain had been isolated, and not an admixed one.

4. Strong genetic structure between the Bay of Brest and Vigo, but no genetic structure between samples within each site

Pairwise Fst were calculated between all metaT samples and results were summarized in Figure 5. Within each site, all Fst values were below 0.03, i.e. about at the precision limit of the method, indicating an absence of intra-site genetic differentiation. It should be noted that the 36 metaT datasets from the Bay of Brest were sampled during *A. minutum* blooms that occurred over three consecutive years (Supplementary Table 2), indicating both intra- and inter-annual stability of the population genetic composition. In sharp contrast, genetic differentiation between geographic sites was extremely strong, with median Fst values of 0.55, 0.59, and 0.71 between Bays of Brest and Vigo, Bay of Brest and Penzé Estuary, and Bay of Vigo and Penzé Estuary, respectively.

5. Folded joint allele frequency spectrum (JAFS) is compatible with ancient divergence without gene flow

The metaT samples coming from the same geographic site were pooled and allelic frequencies at the three geographic sites compared using folded JAFS (Figure 6). The three JAFS highlighted extreme allele frequency differences in the three populations. Most of the SNPs were distributed along the axes, indicating that alleles segregating at intermediate frequencies at a given site are often absent of the two other sites (private alleles). Moreover, at numerous SNP positions, alleles appearing fixed, or almost fixed, at a given site were totally absent from the two other sites. The JAFS also showed a strong excess of SNP positions displaying alleles restricted to the Bay of Brest, indicating that this population was genetically more diverse than the two others.

6. Gene specific structure may be inferred from metaT samples

For each population, haplotype diversity was investigated along *A. minutum* linkage groups (Figure 7A). As already identified using the JAFS (Figure 6), diversity was higher in the population from the Bay of Brest. Haplotype diversity fluctuated along the linkage groups, with transient increase or decrease. Haplotype diversity modifications were population specific rather than shared between the three populations. In agreement with the JAFS analysis, 10,099 SNPs were identified as displaying different allele frequencies in the three populations (pcadapt, adjusted p-value < 1e-10). These SNPs were widespread along the linkage groups, but a few genomic regions displayed higher proportions. Genomic regions enriched in significant SNPs often corresponded to regions displaying low haplotype diversity in one or two populations, but this was not systematically the case. Two genomic regions appeared of special interest in linkage groups L1 and L37 (Figure 7B, C). From 0 to 1.6 cM of L1, haplotype diversity was extremely low in the population from the Bay of Brest and displayed a slight increase in SNPs identified as displaying different allelic frequencies in the three populations (Figure 7B). This region contained 96 SNPs coming from 26 transcripts, including 12 annotated ones (Table 2). Haplotype diversity was extremely low in the Bay of Vigo population at 107.4 cM in L37 (Figure 7C). This region encompassed 327 SNPs in 43 transcripts, including 15 annotated (Table 3) and was characterized by a strong excess of SNPs displaying different allelic frequencies in the three populations.

7. Genetic structure may be inferred from metaT samples even in absence of pre-existing strain validated SNP database

Results presented above in terms of Fst, JAFS and genome wide divergence analyses from metaT datasets were restricted to strain validated SNPs. These same analyses were also performed using SNPs identified *de novo* using the metaT datasets. Overall results were extremely similar using strain validated and *de novo* SNPs, with a strong structure between

geographic sites and absence of structure within sites (Supplementary Figure 4), with the occurrence of numerous private alleles and of fixed polymorphism (Supplementary Figure 5). Fluctuating haplotype diversity along linkage groups and low diversity in the Bay of Brest at the beginning of L1 and in the Bay of Vigo at the end of L37 (Supplementary Figure 6) was also detected. A few differences may nevertheless be noticed. First, the number of SNPs considered was several times higher when using *de novo* SNPs (Table 1). Second, inter-population Fst values were lower (0.10<Fst<0.20 with de novo SNPs compared to 0.5<Fst<0.7 with strain validated SNPs). Third, a high proportion of SNPs displayed similar allelic frequencies in the three populations (Supplementary Figure 5). Finally, haplotype diversity was higher in the Bay of Vigo and Penzé Estuary populations when using *de novo* SNPs (but still slightly lower than in the Bay of Brest).

Discussion

In the present study, the genetic structure of the dinoflagellate *A. minutum* populations in Western Europe was investigated using strains and metaT samples. MetaT datasets were shown to be extremely insightful to decipher the complex genetic structure of such a microbial organism. Results highlighted very strong genetic structure probably resulting from an ancient divergence without gene flow, numerous SNP markers displaying private alleles spread-out in the genomes in the different populations, as well as a few genomic regions displaying very low genetic diversity in one population. These results are discussed below.

1. Using metaT samples to infer protist genetic structure

The estimation of allelic frequencies is not biased by gene expression. This could have been a major issue in case of widespread and systematic allele specific expression. To illustrate the potential issue, one may imagine two populations living in two distinct environments, *A* and *B,* each displaying two alleles, *x* and *y,* at a 50/50 allele frequency. However, if allele *x* is more expressed in environment *A* and allele *y* in environment *B*, estimated allele frequency from

mRNA sequences would be biased compared to actual allelic frequencies. Allele specific expression (ASE) is mostly studied in humans, by monitoring the relative expression of the two gene copies in heterozygotes across cell types and tissues. Results suggest unequal expression of gene copies may be widespread, but mostly for gene copies displaying genetic variation in cis-regulatory regions. Moreover the observed bias tends to vary from cell to cell or tissue to tissue (Cleary & Seoighe, 2021; Montgomery et al., 2011; Wagner et al., 2010). Our pooling and resampling approach clearly shows that population wide ASE is not a global issue. Nevertheless, we cannot rule out that for specific SNP markers, especially in strong linkage disequilibrium with a cis-regulatory variant, the observed allelic frequency inferred from metaT datasets would be influenced by population wide ASE.

In metaT samples, SNP coverage is difficult to control *a priori*, because it strongly depends upon the relative frequency of the species of interest in the sampled community, as well as upon the relative expression of the carrying gene relative to all other genes expressed by this species. However, it is a critical factor to properly estimate allele frequencies. Indeed, as SNP coverage decreases, observed allelic frequency would be strongly affected by sampling error. The pooling and resampling approach helped identify the relationship between species coverage and expected Fst precision. For *A. minutum*, the resampling approach suggested that a minimum coverage of 5e+05 reads is sufficient to detect genetic structure corresponding to Fst > 0.1 and coverage higher than 5e+06 reads enabled extremely precise Fst estimates (Fst ~ 0.01). The main consequence is that, depending on the sample sequencing depth, a high relative abundance of the species of interest in sampled communities may be required to accurately identify very low levels of genetic differentiation (for instance, Fst ~ 0.01 may be detected with a relative abundance of 0.25 if the sample sequencing depth is 2e+07, but only 0.05 if the sample sequencing depth is 1e+08). However, high levels of differentiation (Fst > 0.1) could be identified, even in very moderately abundant species (for instance a relative abundance of 0.025 or 0.005 if the sample sequencing depth is 2e+07 or 1e+08, respectively). As a matter of comparison, the average sequencing depth in the samples used in the present

418 study was 2e+07 (Metegnier et al., 2020) and 1.6e+08 in Tara Ocean samples (Carradec et
419 al., 2018).

420 The comparison between strain and metaT samples was extremely insightful. Population
421 structure inferred from the two approaches was extremely similar. This has several
422 implications. At the same time, it encourages the use of metaT samples to determine genetic
423 structure of protist populations, but also confirms that strain based population samples are not
424 necessarily biased compared to natural populations. This could have been the case if the
425 number of strains was too small to capture the genetic diversity of natural populations, or if
426 single cell isolation and subsequent culturing steps to obtain clonal strains selected an
427 unrepresentative subset of natural populations. We should note that, in the present system,
428 genetic structure was extremely strong, with numerous private alleles, making it easy to detect
429 even with a small strain sample size.

430 Another insightful result of the metaT sample analyses was that replicate samples from the
431 same population displayed very low temporal variability, indicating that the population genetic
432 diversity may be captured using a very low number of samples. As an illustration, we may refer
433 to the extreme similarity of the samples from the Bay of Vigo, sampled over a distance of ten
434 kilometers, or of the samples from the Bay of Brest obtained at the same locality during three
435 consecutive years. This has profound implications for future investigation of protist population
436 structure. Indeed, using strains, a strong sampling effort has to be taken to capture the genetic
437 diversity of the population of interest at a given locality, limiting the feasibility of genetic
438 structure analyses to a handful of sites. Using metaT samples, as the population genetic
439 diversity may be captured using a very limited number of easy to obtain samples, a finer spatial
440 sampling resolution may be considered. Increasing the spatial coverage with a relatively low
441 resolution can be an essential step towards a better understanding of the extremely complex
442 spatio-temporal protist population genetic structure (see below).

443 The present study highlighted that population genetic analyses can be performed from metaT
444 datasets even in absence of pre-validated SNPs. Indeed, results obtained from strain validated
445 SNPs and SNPs detected *de novo* from the metaT datasets were extremely similar. This was
446 true for the overall genetic structure, as well as for the genomic regions, genes and even SNPs
447 (identified as displaying different allelic frequencies across populations). We noted that Fst
448 values were much lower using *de novo* SNPs and that numerous SNPs identified *de novo*
449 appeared to display similar allelic frequencies across populations, while this was rarely the
450 case when using strain validated SNPs. Such differences may be explained by the peculiar
451 organization of dinoflagellate genes. Indeed, these are organized in tandem repeats
452 (Stephens et al., 2020; Wisecaver & Hackett, 2011). During mapping, reads belonging to
453 different gene copies align to the same reference transcript. However, the different gene
454 copies within a single cell may display SNPs, reflecting the genetic divergence of gene copies
455 following gene duplication in a given genome. As dinoflagellate vegetative cells are haploid, it
456 is possible to exclude these markers by restricting the analyses to SNPs displaying
457 monomorphism within a given genome but polymorphism across genomes when genotyping
458 individual strains (see methods). Using SNPs identified *de novo* from metaT datasets, such
459 restriction is impossible and the population genetic analyses include both markers of genetic
460 diversity between and within genomes. The observed difference between strain-validated and
461 *de novo* SNPs (lower Fst and shared polymorphism with *de novo* SNPs) thus probably
462 resulted from ancestral gene duplications and mutations of the various gene copies that
463 preceded the population splits. Despite the confounding effect of the genetically variable gene
464 copies within a given genome, the signal of genetic divergence between populations is still
465 captured from metaT samples. Using a *de novo* SNP approach is not only a possibility, but it
466 may also be a better choice than using strain validated SNPs. Indeed, in the present study,
467 the haplotype diversity of the Penzé Estuary populations was much lower using the strain
468 validated SNPs compared to the *de novo* approach, probably because no strains from this
469 population were isolated and genotyped, preventing the identification of the Penzé Estuary
470 population private polymorphism using the strain validated approach. These results are

especially promising for applying metaT based population genetic approaches to a wide range of species, including species difficult to grow in the lab but with a reference transcriptome or genome available, or even species not cultivated but for which reference transcriptomes could be computed directly from metaT or metaG datasets (Delmont et al., 2021; Vorobev et al., 2020).

2. Strong divergence between *A. minutum* populations in Western Europe

Within *A. minutum*, a strong genetic structure was inferred both from strains and metaT samples in Western Europe. A total of four highly divergent groups were identified. The first group (NE_B) was identified only using strains sampled outside of massive *A. minutum* developments, i.e. bloom events. The metaT samples were exclusively obtained from blooms and there was no sign of the presence of individuals belonging to the NE_B group in these samples. Strains from NE_B were phenotypically distinct from the ones from the other three populations, as they did not produce PSTs (Geffroy et al., 2021). Interestingly, the genetic diversity of this population was much higher than the other three populations. Theoretically, genetic diversity is directly related to the effective population size (Ellegren & Galtier, 2016), which would imply that the NE_B effective population size is much higher than the three others. Considering the hypothesis that the NE_B population is found at low cell density, this might seem counterintuitive. However, dense *A. minutum* blooms tend to occur transiently, in restricted geographic areas, like the one in Bay of Vigo in 2018, to which the strains used in this study belong (Nogueira et al., 2022). Moreover, massive development probably involves numerous rounds of mitotic division, contributing to decreasing the effective population size despite huge census population sizes (Ellegren & Galtier, 2016). Given the volumes at play at sea, very low density populations across broad areas could potentially outnumber both census and effective population sizes of bloom forming populations. The role of low density

populations is rarely considered in protist population genetics although we might speculate that they could play an important role in terms of metapopulation dynamics.

The other three populations were exclusively sampled during blooms. The first one, NE_A, was sampled thanks to strains isolated in the Atlantic Ocean and English Channel and corresponded to all metaT sampled from the Bay of Brest during three consecutive years. This population was the most genetically diverse bloom forming population and displayed an extremely stable genetic structure in the Bay of Brest, during the three years sampled using metaT, but also probably during several decades, as two strains isolated from sediment cores and dated to 1947 belonged to this same population. The second one corresponded to strains and metaT samples obtained from the Bay of Vigo during a huge red tide. Finally, the last population was sampled from a single metaT sample from the Penzé Estuary in 2015. Private alleles tend to indicate that this last population was truly distinct from the others and did not correspond to the admixture of other populations. Surprisingly, strains isolated from the Penzé Estuary from 1989 to 2013 all belonged to the NE_A clade. With such a limited sampling, it is difficult to determine whether this new population was transient or if it replaced the NE_A population.

Altogether, our study highlighted an extremely complex *A. minutum* population structure, with extremely high levels of divergence supported by the presence of numerous private alleles, including fixed polymorphism. The divergence between NE_A and NE_B had been characterized previously using a few strains (Le Gac et al., 2016c) and the present study showed that several *A. minitum* populations exhibit extensive divergence in Western Europe. The observed divergence pattern is compatible with ancestral divergence and an absence of contemporary gene flow between populations (Almeida et al., 2015; Ellegren, 2014; Gutenkunst et al., 2009).

Due to their extensive dispersal abilities in the marine environment, protist species are commonly expected to be cosmopolite and display homogeneous populations worldwide (De

522 Wit & Bouvier, 2006; Finlay, 2002). However, strong genetic structure is extremely common
523 in marine protist species (Casabianca et al., 2012; Craig et al., 2019; Gao et al., 2019; Godhe
524 & Rynearson, 2017; Paredes et al., 2019; Rengefors et al., 2017). This pattern is in sharp
525 contrast with marine macroorganism populations often displaying extremely low levels of
526 divergence between them (Gagnaire et al., 2015; Waples, 1998). Protist genetic structure is
527 sometimes correlated with geographic or hydrodynamic features (Casabianca et al., 2012;
528 Casteleyn et al., 2010; Godhe et al., 2016; Postel et al., 2020) as well as with various
529 environmental parameters (Sassenhagen et al., 2018; Sjöqvist et al., 2015; Whittaker &
530 Rynearson, 2017). Moreover divergent populations sometimes co-occur for extensive periods
531 of time (Härnström et al., 2011; Lundholm et al., 2017). The level of divergence identified in
532 the present study and in other species is on the order of what is expected for sibling species.
533 One possibility to explain the observed level of divergence would be the existence of
534 numerous cryptic species, evolving without exchanging genes, but displaying virtually no
535 morphological difference, within taxonomically recognized protist species. Hence numerous
536 cryptic species have been repeatedly identified from what were previously thought to be single
537 cosmopolitan protist species (De Luca et al., 2021; Smayda, 2011). Even using molecular
538 tools, identifying cryptic species may be especially difficult for protists. Indeed, the
539 coalescence time (the time since the most common ancestor of two gene copies) is expected
540 to be directly related to the effective population size (Hudson, 1990). With huge population
541 sizes, the coalescence time may easily be much higher than the speciation time. As a
542 consequence, ancestral polymorphism may remain in contemporary cryptic species, blurring
543 the phylogenetic signal of species split and preventing the identification of cryptic species
544 (Rannala et al., 2020).

545 Given their census population size, intra-population genetic diversity would be expected to be
546 extremely high. However, the present study and previous ones (Blanc-Mathieu et al., 2017;
547 Filatov, 2019), indicated that the observed levels of diversity in protists tend to be on the order
548 of what was observed for moderately diverse animal species displaying much lower census

549 population sizes (Romiguier et al., 2014). Previous studies calculated that given the observed
550 genetic diversity, the effective population size should be on the order of tens of millions of
551 cells, a number of cells that may be found in a few liters of sea water (Filatov, 2019). Such
552 discrepancy between the census population size and the level of diversity is known as the
553 Lewontin's paradox (Buffalo, 2021; Ellegren & Galtier, 2016; Filatov, 2019; Galtier &
554 Rousselle, 2020). Three main solutions have been proposed to solve this paradox. The first
555 one emphasizes the possibility of reduced mutation rate in species with large populations.
556 Mutation rate estimation are rather limited in protist species, but tend to indicate that it does
557 not deviate from rates of species displaying much lower census population sizes (Krasovec et
558 al., 2020). The second one emphasizes the importance of demographic fluctuations and
559 asexuality on reducing the effective population size. The third one, the reduction of genetic
560 diversity at the neighborhood of sites increasing in frequency in a population due to selection
561 (genetic hitchhiking or draft). A thorough investigation in the haptophyte *Emiliania huxleyi*
562 appeared compatible with a major importance of recurrent selective events in shaping the
563 genetic diversity of this species (Filatov, 2019). In the present study, the numerous SNPs
564 displaying extremely divergent allelic frequencies could result from adaptive evolution
565 occurring independently in the different populations. These SNPs appeared spread-out in the
566 genome and most of the time they were not surrounded by genomic regions of low haplotype
567 diversity. Such a signature would be compatible with selection from standing genetic variation
568 during which neutral alleles segregating in the population become adaptive following a
569 modification of the selective pressures (Barrett & Schluter, 2008). As they appear in various
570 genomic contexts, such mutations could increase in frequency in the population with minimal
571 effect on the genetic diversity at neighboring sites. Such events would correspond to soft
572 selective sweeps (Ellegren, 2014). Nevertheless, a few genomic regions tend to display a very
573 low level of diversity in one population compared to the others, but also compared to the other
574 genomic regions of the same population. Such events are compatible with hard selective
575 sweeps during which a *de novo* adaptive mutation appears in a population and rises to fixation
576 with the genomic context in which it appeared (Ellegren, 2014).

As a conclusion metaT datasets may be great to investigate the population genetics of protist species relatively abundant in given communities. By circumventing the time-consuming culturing steps required to genotype individual cells from natural populations, metaT but also any other population or community wide dataset (metaG, multiplex amplicon sequencing…) should be of primary importance to improve the spatio-temporal sampling of protist populations. Such improvements are mandatory if we want to better understand the population genetics of these species that display moderate diversity and complex spatio-temporal genetic structure despite theoretical expectations of extremely high diversity and low genetic structure. One of the major constraints is that most of the population genomic tools (linkage disequilibrium analyses, simulations…) were developed considering individual genotypes and not allelic frequencies estimated from population wide samples. A situation that may change given the growing interest in using eDNA/RNA to determine the population genetics of a wide range of organisms (Sigsgaard et al., 2020).

Acknowledgements

The study was funded by PRIMROSE (EC Interreg Atlantic Area $EAPA_{182}$/2016) and the Brittany Region as part of the *Paleoecology of* Alexandrium minutum *dans la Rade de Brest*– Marché n°2017-90292 project PALMIRA. We thank all the participants in the crew of the RV Ramón Margalef from the Remedios cruise (Research project - grant number CTM2016-75451-C2-1-R), particularly B. Mourino-Carballido, for their support to the sample collection. We thank the Ifremer Sebimer team for bioinformatic support as well as the INRAE GeT-PlaGe sequencing platform for sequencing.

References:

602 Alexander, H., Jenkins, B. D., Rynearson, T. A., & Dyhrman, S. T. (2015).
603 Metatranscriptome analyses indicate resource partitioning between diatoms in the
604 field. *Proceedings of the National Academy of Sciences*, *112*(17), E2182-E2190.
605 https://doi.org/10.1073/pnas.1421993112
606 Alexander, H., Rouco, M., Haley, S. T., Wilson, S. T., Karl, D. M., & Dyhrman, S. T. (2015).
607 Functional group-specific traits drive phytoplankton dynamics in the oligotrophic
608 ocean. *Proceedings of the National Academy of Sciences*, *112*(44), E5972-E5979.
609 https://doi.org/10.1073/pnas.1518165112
610 Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., Legras, J.-L.,
611 Serra, M., Dequin, S., Couloux, A., Guy, J., Bensasson, D., Gonçalves, P., &
612 Sampaio, J. P. (2015). A population genomics insight into the Mediterranean origins
613 of wine yeast domestication. *Molecular Ecology*, *24*(21), 5412-5427.
614 https://doi.org/10.1111/mec.13341
615 Barrett, R. D. H., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in*
616 *Ecology & Evolution*, *23*(1), 38-44. https://doi.org/10.1016/j.tree.2007.09.008
617 Ben-Gigirey, B., Rossignoli, A. E., Riobó, P., & Rodríguez, F. (2020). First Report of
618 Paralytic Shellfish Toxins in Marine Invertebrates and Fish in Spain. *Toxins*, *12*(11),
619 723. https://doi.org/10.3390/toxins12110723
620 Biller, S. J., Berube, P. M., Lindell, D., & Chisholm, S. W. (2015). Prochlorococcus : The
621 structure and function of collective diversity. *Nature Reviews Microbiology*, *13*(1),
622 13-27. https://doi.org/10.1038/nrmicro3378
623 Blanc-Mathieu, R., Krasovec, M., Hebrard, M., Yau, S., Desgranges, E., Martin, J.,
624 Schackwitz, W., Kuo, A., Salin, G., Donnadieu, C., Desdevises, Y., Sanchez-
625 Ferandin, S., Moreau, H., Rivals, E., Grigoriev, I. V., Grimsley, N., Eyre-Walker, A., &
626 Piganeau, G. (2017). Population genomics of picophytoplankton unveils novel
627 chromosome hypervariability. *Science Advances*, *3*(7), e1700239.
628 https://doi.org/10.1126/sciadv.1700239
629 Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic : A flexible trimmer for Illumina

630 sequence data. *Bioinformatics*, *30*(15), 2114-2120.
631 https://doi.org/10.1093/bioinformatics/btu170
632 Buffalo, V. (2021). Quantifying the relationship between genetic diversity and population size
633 suggests natural selection cannot explain Lewontin's Paradox. *eLife*, *10*, e67509.
634 https://doi.org/10.7554/eLife.67509
635 Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R.,
636 Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A.,
637 Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J.,
638 Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes.
639 *Nature Communications*, *9*(1), 373. https://doi.org/10.1038/s41467-017-02342-1
640 Casabianca, S., Penna, A., Pecchioli, E., Jordi, A., Basterretxea, G., & Vernesi, C. (2012).
641 Population genetic structure and connectivity of the harmful dinoflagellate
642 Alexandrium minutum in the Mediterranean Sea. *Proceedings of the Royal Society B:*
643 *Biological Sciences*, *279*(1726), 129-138. https://doi.org/10.1098/rspb.2011.0708
644 Casteleyn, G., Leliaert, F., Backeljau, T., Debeer, A.-E., Kotaki, Y., Rhodes, L., Lundholm,
645 N., Sabbe, K., & Vyverman, W. (2010). Limits to gene flow in a cosmopolitan marine
646 planktonic diatom. *Proceedings of the National Academy of Sciences*, *107*(29),
647 12952-12957. https://doi.org/10.1073/pnas.1001380107
648 Cleary, S., & Seoighe, C. (2021). Perspectives on Allele-Specific Expression. *Annual Review*
649 *of Biomedical Data Science*, *4*, 101-122. https://doi.org/10.1146/annurev-biodatasci-
650 021621-122219
651 Cohen, N. R., Ellis, K. A., Lampe, R. H., McNair, H., Twining, B. S., Maldonado, M. T.,
652 Brzezinski, M. A., Kuzminov, F. I., Thamatrakoln, K., Till, C. P., Bruland, K. W.,
653 Sunda, W. G., Bargu, S., & Marchetti, A. (2017). Diatom Transcriptional and
654 Physiological Responses to Changes in Iron Bioavailability across Ocean Provinces.
655 *Frontiers in Marine Science*, *4*, 360. https://doi.org/10.3389/fmars.2017.00360
656 Craig, R. J., Böndel, K. B., Arakawa, K., Nakada, T., Ito, T., Bell, G., Colegrave, N.,
657 Keightley, P. D., & Ness, R. W. (2019). Patterns of population structure and complex

haplotype sharing among field isolates of the green alga Chlamydomonas reinhardtii. *Molecular Ecology*, *28*(17), 3977-3993. https://doi.org/10.1111/mec.15193

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

De Luca, D., Piredda, R., Sarno, D., & Kooistra, W. H. C. F. (2021). Resolving cryptic species complexes in marine protists : Phylogenetic haplotype networks meet global DNA metabarcoding datasets. *The ISME Journal*, *15*(7), 1931-1942. https://doi.org/10.1038/s41396-021-00895-0

De Wit, R., & Bouvier, T. (2006). 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology*, *8*(4), 755-758. https://doi.org/10.1111/j.1462-2920.2006.01017.x

Delebecq, G., Schmidt, S., Ehrhold, A., Latimier, M., & Siano, R. (2020). Revival of Ancient Marine Dinoflagellates Using Molecular Biostimulation. *Journal of Phycology*, *56*(4), 1077-1089. https://doi.org/10.1111/jpy.13010

Delmont, T. O., Gaia, M., Hinsinger, D. D., Fremont, P., Vanni, C., Guerra, A. F., Eren, A. M., Kourlaiev, A., d'Agata, L., Clayssen, Q., Villar, E., Labadie, K., Cruaud, C., Poulain, J., Silva, C. D., Wessner, M., Noel, B., Aury, J.-M., Coordinators, T. O., … Jaillon, O. (2021). *Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics* (p. 2020.10.15.341214). https://doi.org/10.1101/2020.10.15.341214

Dia, A., Guillou, L., Mauger, S., Bigeard, E., Marie, D., Valero, M., & Destombe, C. (2014). Spatiotemporal changes in the genetic diversity of harmful algal blooms caused by the toxic dinoflagellate Alexandrium minutum. *Molecular Ecology*, *23*(3), 549-560. https://doi.org/10.1111/mec.12617

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, *29*(1), 51-63.

https://doi.org/10.1016/j.tree.2013.09.008

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, *17*(7), 422-433. https://doi.org/10.1038/nrg.2016.58

Filatov, D. A. (2019). Extreme Lewontin's Paradox in Ubiquitous Marine Phytoplankton Species. *Molecular Biology and Evolution*, *36*(1), 4-14. https://doi.org/10.1093/molbev/msy195

Finlay, B. J. (2002). Global Dispersal of Free-Living Microbial Eukaryote Species. *Science*, *296*(5570), 1061-1063. https://doi.org/10.1126/science.1070710

Gagnaire, P.-A., Broquet, T., Aurelle, D., Viard, F., Souissi, A., Bonhomme, F., Arnaud-Haond, S., & Bierne, N. (2015). Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evolutionary Applications*, *8*(8), 769-786. https://doi.org/10.1111/eva.12288

Galtier, N., & Rousselle, M. (2020). How Much Does Ne Vary Among Species? *Genetics*, *216*(2), 559-572. https://doi.org/10.1534/genetics.120.303622

Gao, Y., Sassenhagen, I., Richlen, M. L., Anderson, D. M., Martin, J. L., & Erdner, D. L. (2019). Spatiotemporal genetic structure of regional-scale Alexandrium catenella dinoflagellate blooms explained by extensive dispersal and environmental selection. *Harmful Algae*, *86*, 46-54. https://doi.org/10.1016/j.hal.2019.03.013

Garcés, E., Bravo, I., Vila, M., Figueroa, R. I., Masó, M., & Sampedro, N. (2004). Relationship between vegetative cells and cyst production during Alexandrium minutum bloom in Arenys de Mar harbour (NW Mediterranean). *Journal of Plankton Research*, *26*(6), 637-645. https://doi.org/10.1093/plankt/fbh065

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]*. http://arxiv.org/abs/1207.3907

Geffroy, S., Lechat, M.-M., Le Gac, M., Rovillon, G.-A., Marie, D., Bigeard, E., Malo, F., Amzil, Z., Guillou, L., & Caruana, A. M. N. (2021). From the sxtA4 Gene to Saxitoxin Production : What Controls the Variability Among Alexandrium minutum and Alexandrium pacificum Strains? *Frontiers in Microbiology*, *12*, 341.

https://doi.org/10.3389/fmicb.2021.613199

Godhe, A., & Rynearson, T. (2017). The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160399. https://doi.org/10.1098/rstb.2016.0399

Godhe, A., Sjöqvist, C., Sildever, S., Sefbom, J., Harðardóttir, S., Bertos-Fortis, M., Bunse, C., Gross, S., Johansson, E., Jonsson, P. R., Khandan, S., Legrand, C., Lips, I., Lundholm, N., Rengefors, K. E., Sassenhagen, I., Suikkanen, S., Sundqvist, L., & Kremp, A. (2016). Physical barriers and environmental gradients cause spatial and temporal genetic differentiation of an extensive algal bloom. *Journal of Biogeography*, *43*(6), 1130-1142. https://doi.org/10.1111/jbi.12722

Gong, W., Paerl, H., & Marchetti, A. (2018). Eukaryotic phytoplankton community spatiotemporal dynamics as identified through gene expression within a eutrophic estuary. *Environmental Microbiology*, *20*(3), 1095-1111. https://doi.org/10.1111/1462-2920.14049

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Härnström, K., Ellegaard, M., Andersen, T. J., & Godhe, A. (2011). Hundred years of genetic structure in a sediment revived diatom population. *Proceedings of the National Academy of Sciences*, *108*(10), 4252-4257. https://doi.org/10.1073/pnas.1013528108

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring Genetic Differentiation from Pool-seq Data. *Genetics*, *210*(1), 315-330. https://doi.org/10.1534/genetics.118.300900

Hu, S. K., Liu, Z., Alexander, H., Campbell, V., Connell, P. E., Dyhrman, S. T., Heidelberg, K. B., & Caron, D. A. (2018). Shifting metabolic priorities among key protistan taxa within and below the euphotic zone. *Environmental Microbiology*, *20*(8), 2865-2879.

https://doi.org/10.1111/1462-2920.14259

Hudson, R. R. (1990). Gene genealogies and the coalescent process. In *Oxford surveys in evolutionary biology* (Futuyama and Antonovics, Vol. 7, p. 1-44). http://home.uchicago.edu/~rhudson1/popgen356/OxfordSurveysEvolBiol7_1-44.pdf

Ji, N., Lin, L., Li, L., Yu, L., Zhang, Y., Luo, H., Li, M., Shi, X., Wang, D.-Z., & Lin, S. (2018). Metatranscriptome analysis reveals environmental and diel regulation of a Heterosigma akashiwo (raphidophyceae) bloom. *Environmental Microbiology*, *20*(3), 1078-1094. https://doi.org/10.1111/1462-2920.14045

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) : Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biology*, *12*(6), e1001889. https://doi.org/10.1371/journal.pbio.1001889

Krasovec, M., Rickaby, R. E. M., & Filatov, D. A. (2020). Evolution of Mutation Rate in Astronomically Large Phytoplankton Populations. *Genome Biology and Evolution*, *12*(7), 1051-1059. https://doi.org/10.1093/gbe/evaa131

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., White, A. E., & Armbrust, E. V. (2021). *The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics* (p. 2021.01.15.426851). https://doi.org/10.1101/2021.01.15.426851

Lampe, R. H., Cohen, N. R., Ellis, K. A., Bruland, K. W., Maldonado, M. T., Peterson, T. D., Till, C. P., Brzezinski, M. A., Bargu, S., Thamatrakoln, K., Kuzminov, F. I., Twining, B. S., & Marchetti, A. (2018). Divergent gene expression among phytoplankton taxa in response to upwelling. *Environmental Microbiology*, *20*(8), 3069-3082. https://doi.org/10.1111/1462-2920.14361

[dataset] Le Gac, M. Quéré, J. (2016a). A. minutum divergence. European Nucleotide

770 Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB15046
771 [dataset] Le Gac Mickael, Metegnier Gabriel, Chomerat Nicolas, Malestroit Pascale, Quere
772 Julien, Bouchez Olivier, Siano Raffaele, Destombe Christophe, Guillou Laure,
773 Chapelle Annie (2016b). Evolutionary processes and cellular functions underlying
774 divergence in Alexandrium minutum. SEANOE. https://doi.org/10.17882/45445
775 Le Gac, M., Metegnier, G., Chomérat, N., Malestroit, P., Quéré, J., Bouchez, O., Siano, R.,
776 Destombe, C., Guillou, L., & Chapelle, A. (2016c). Evolutionary processes and
777 cellular functions underlying divergence in Alexandrium minutum. *Molecular Ecology*,
778 *25*(20), 5129-5143. https://doi.org/10.1111/mec.13815
779 Lewis, A. M., Coates, L. N., Turner, A. D., Percy, L., & Lewis, J. (2018). A review of the
780 global distribution of Alexandrium minutum (Dinophyceae) and comments on ecology
781 and associated paralytic shellfish toxin profiles, with a focus on Northern Europe.
782 *Journal of Phycology*, *54*(5), 581-598. https://doi.org/10.1111/jpy.12768
783 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-
784 MEM. *arXiv:1303.3997 [q-bio]*. http://arxiv.org/abs/1303.3997
785 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,
786 G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The
787 Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.
788 https://doi.org/10.1093/bioinformatics/btp352
789 Lilly, E. L., Halanych, K. M., & Anderson, D. M. (2005). Phylogeny, biogeography, and
790 species boundaries within the Alexandrium minutum group. *Harmful Algae*, *4*(6),
791 1004-1020. https://doi.org/10.1016/j.hal.2005.02.001
792 Lundholm, N., Ribeiro, S., Godhe, A., Rostgaard Nielsen, L., & Ellegaard, M. (2017).
793 Exploring the impact of multidecadal environmental changes on the population
794 genetic structure of a marine primary producer. *Ecology and Evolution*, *7*(9),
795 3132-3142. https://doi.org/10.1002/ece3.2906
796 Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt : An R package to perform genome
797 scans for selection based on principal component analysis. *Molecular Ecology*

Resources, *17*(1), 67-77. https://doi.org/10.1111/1755-0998.12592

Marchetti, A., Schruth, D. M., Durkin, C. A., Parker, M. S., Kodner, R. B., Berthiaume, C. T., Morales, R., Allen, A. E., & Armbrust, E. V. (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences*, *109*(6), E317-E325. https://doi.org/10.1073/pnas.1118408109

Mary, L., Quéré, J., Latimier, M., Rovillon, G.-A., Hégaret, H., Réveillon, D., & Le Gac, M. (Submitted). Genetic association of toxin production in the dinoflagellate Alexandrium minutum. *Microbial Genomics*.

McCauley, L. A. R., Erdner, D. L., Nagai, S., Richlen, M. L., & Anderson, D. M. (2009). BIOGEOGRAPHIC ANALYSIS OF THE GLOBALLY DISTRIBUTED HARMFUL ALGAL BLOOM SPECIES ALEXANDRIUM MINUTUM (DINOPHYCEAE) BASED ON rRNA GENE SEQUENCES AND MICROSATELLITE MARKERS1. *Journal of Phycology*, *45*(2), 454-463. https://doi.org/10.1111/j.1529-8817.2009.00650.x

[dataset] Metegnier G., Quere J., Le Gac M. (2015). Metatranscriptomic sequences from Alexandrium minutum blooms sampled in situ in the bay of Brest (France) between 2013 and 2015. IFREMER. http://dx.doi.org/10.12770/9d4131da-b33b-429b-9cdd-e7325b06f7d8

Metegnier, G., Paulino, S., Ramond, P., Siano, R., Sourisseau, M., Destombe, C., & Le Gac, M. (2020). Species specific gene expression dynamics during harmful algal blooms. *Scientific Reports*, *10*(1), 6182. https://doi.org/10.1038/s41598-020-63326-8

Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M., & Dermitzakis, E. T. (2011). Rare and Common Regulatory Variation in Population-Scale Sequenced Human Genomes. *PLOS Genetics*, *7*(7), e1002144. https://doi.org/10.1371/journal.pgen.1002144

Nogueira, E., Bravo, I., Montero, P., Díaz-Tapia, P., Calvo, S., Ben-Gigirey, B., Figueroa, R. I., Garrido, J. L., Ramilo, I., Lluch, N., Rossignoli, A. E., Riobó, P., & Rodríguez, F.

(2022). HABs in coastal upwelling systems : Insights from an exceptional red tide of the toxigenic dinoflagellate Alexandrium minutum. *Ecological Indicators*, *137*, 108790. https://doi.org/10.1016/j.ecolind.2022.108790

Paredes, J., Varela, D., Martínez, C., Zúñiga, A., Correa, K., Villarroel, A., & Olivares, B. (2019). Population Genetic Structure at the Northern Edge of the Distribution of Alexandrium catenella in the Patagonian Fjords and Its Expansion Along the Open Pacific Ocean Coast. *Frontiers in Marine Science*, *5*, 532. https://doi.org/10.3389/fmars.2018.00532

Postel, U., Glemser, B., Alekseyeva, K. S., Eggers, S. L., Groth, M., Glöckner, G., John, U., Mock, T., Klemm, K., Valentin, K., & Beszteri, B. (2020). Adaptive divergence across Southern Ocean gradients in the pelagic diatom Fragilariopsis kerguelensis. *Molecular Ecology*, *29*(24), 4913-4924. https://doi.org/10.1111/mec.15554

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559-575. https://doi.org/10.1086/519795

Rannala, B., Edwards, S. V. S. V., Leaché, A., & Yang, Z. (2020). The Multi-species Coalescent Model and Species Tree Inference. In C. Scornavacca, F. Delsuc, & N. Galtier (Éds.), *Phylogenetics in the Genomic Era* (p. 3.3:1-3.3:21). No commercial publisher | Authors open access book. https://hal.archives-ouvertes.fr/hal-02535622

Rengefors, K., Kremp, A., Reusch, T. B. H., & Wood, A. M. (2017). Genetic diversity and evolution in eukaryotic phytoplankton : Revelations from population genetic studies. *Journal of Plankton Research*, *39*(2), 165-179. https://doi.org/10.1093/plankt/fbw098

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., Chiari, Y., Dernat, R., Duret, L., Faivre, N., Loire, E., Lourenco, J. M., Nabholz, B., Roux, C., Tsagkogeorga, G., Weber, A. a.-T., Weinert, L. A., Belkhir, K., Bierne, N., … Galtier, N. (2014). Comparative population genomics in animals uncovers the determinants

of genetic diversity. *Nature*, *515*(7526), 261-263. https://doi.org/10.1038/nature13685

Sassenhagen, I., Gao, Y., Lozano-Duque, Y., Parsons, M. L., Smith, T. B., & Erdner, D. L. (2018). Comparison of Spatial and Temporal Genetic Differentiation in a Harmful Dinoflagellate Species Emphasizes Impact of Local Processes. *Frontiers in Marine Science*, *5*. https://doi.org/10.3389/fmars.2018.00393

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, *13*(2), 245-262. https://doi.org/10.1111/eva.12882

Sjöqvist, C., Godhe, A., Jonsson, P. R., Sundqvist, L., & Kremp, A. (2015). Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea–Baltic Sea salinity gradient. *Molecular Ecology*, *24*(11), 2871-2885. https://doi.org/10.1111/mec.13208

Smayda, T. J. (2011). Cryptic planktonic diatom challenges phytoplankton ecologists. *Proceedings of the National Academy of Sciences*, *108*(11), 4269-4270. https://doi.org/10.1073/pnas.1100997108

Stephens, T. G., González-Pech, R. A., Cheng, Y., Mohamed, A. R., Burt, D. W., Bhattacharya, D., Ragan, M. A., & Chan, C. X. (2020). Genomes of the dinoflagellate Polarella glacialis encode tandemly repeated single-exon genes with adaptive functions. *BMC Biology*, *18*(1), 56. https://doi.org/10.1186/s12915-020-00782-8

Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T. O., Annamalé, A., Wincker, P., & Pelletier, E. (2020). Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Research*, *30*(4), 647-659. https://doi.org/10.1101/gr.253070.119

Wagner, J. R., Ge, B., Pokholok, D., Gunderson, K. L., Pastinen, T., & Blanchette, M. (2010). Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human. *PLOS Computational Biology*, *6*(7), e1000849.

https://doi.org/10.1371/journal.pcbi.1000849

Waples, R. (1998). Separating the wheat from the chaff : Patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, *89*(5), 438-450. https://doi.org/10.1093/jhered/89.5.438

Whittaker, K. A., & Rynearson, T. A. (2017). Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. *Proceedings of the National Academy of Sciences*, *114*(10), 2651-2656. https://doi.org/10.1073/pnas.1612346114

Wisecaver, J. H., & Hackett, J. D. (2011). Dinoflagellate Genome Evolution. *Annual Review of Microbiology*, *65*(1), 369-387. https://doi.org/10.1146/annurev-micro-090110-102841

Wurch, L. L., Alexander, H., Frischkorn, K. R., Haley, S. T., Gobler, C. J., & Dyhrman, S. T. (2019). Transcriptional Shifts Highlight the Role of Nutrients in Harmful Brown Tide Dynamics. *Frontiers in Microbiology*, *10*, 136. https://doi.org/10.3389/fmicb.2019.00136

Zhang, Y., Lin, X., Shi, X., Lin, L., Luo, H., Li, L., & Lin, S. (2019). Metatranscriptomic Signatures Associated With Phytoplankton Regime Shift From Diatom Dominance to a Dinoflagellate Bloom. *Frontiers in Microbiology*, *10*, 590. https://doi.org/10.3389/fmicb.2019.00590

Data Accessibility and Benefit-Sharing

Raw sequence reads are available at https://doi.org/10.12770/9d4131da-b33b-429b-9cdd-e7325b06f7d8 (metaT from Penzé and Bay of Brest), ENAXXX (metaT from Bay of Vigo) http://www.ebi.ac.uk/ena/data/view/PRJEB15046 and ENAXXX (strains). A. minutum Reference transcriptome is available at https://doi.org/10.17882/45445

Benefits Generated: Benefits from this research accrue from the sharing of our data on public databases as described above. The data that supports the findings of this study are available in the supplementary material of this article, as well as from the public database cited above.

Author Contributions

The research was designed by MLG, CD, and MS. The sampling was performed by MLG, MS, GM, JQ, RS, FR. The molecular biology was performed by MLG, JQ, GM. The bioinformatics and population genomic analyses were done by MLG, GM, LM. Writing of the article was carried out by MLG, CD, LM, MS, RS, FR.

Tables

Table 1: Summary of the metaT datasets considered.

Dataset	SNP coverage	Min. reads	Samples	Reference SNPs	Pruned ref. SNPs	De novo SNPs	Diagnostic SNPs (NE_A, NE_B, Vigo)
5P5	5	5e+05	63	727	348	10,362	18, 3, 20
P6	10	1e+06	59	1,063	449	11,801	28, 8, 43
5P6	30	5e+06	40	4,591	1,905	23,586	112, 31, 76
P7	50	1e+07	25	8,004	1,622	28,374	210, 38, 116
P7_20	20	1e+07	25	68,920	19,749	135,711	1344, 231, 570

Pooled_5 P6	30	5e+06 per initial sample	3 (1 per site)	87,873	40,441	371,087	2340, 178, 1027

Table 2. Annotated genes in Linkage group L1 from 0 to 1.6 cM

Contig name	Homolog symbol	Homolog name
comp15150_c0_seq1	TCR8_PASMD	Tetracycline resistance protein, class H
comp26455_c0_seq1	EXD1_MOUSE	Exonuclease 3'-5' domain-containing protein 1
comp40771_c0_seq1	FKBP_YEAST	FK506-binding protein 1
comp61927_c0_seq1	SL9A8_CHICK	Sodium/hydrogen exchanger 8
comp67798_c0_seq1	CAF1_EPHMU	Collagen EMF1-alpha
comp72384_c0_seq1	AGAL_CYATE	Alpha-galactosidase
comp82361_c0_seq1	CDPK1_ARATH	Calcium-dependent protein kinase 1
comp85427_c0_seq1	YR811_MIMIV	Putative ariadne-like RING finger protein R811
comp96664_c1_seq4	KAPR_BLAEM	cAMP-dependent protein kinase regulatory subunit
comp100730_c0_seq1	DLPC_DICDI	Dynamin-like protein C
comp103028_c0_seq2	NUMA1_HUMAN	Nuclear mitotic apparatus protein 1
comp104417_c0_seq4	CLCN7_MOUSE	H(+)/Cl(-) exchange transporter 7

Table 3: Annotated genes in Linkage group L37 at 107.4 cM

Contig name	Homolog symbol	Homolog name
comp25787_c0_seq1	SPSC_BACSU	Spore coat polysaccharide biosynthesis protein SpsC
comp60283_c0_seq1	PUM5_ARATH	Pumilio homolog 5
comp73475_c0_seq1	TYLE_STRFR	Demethylmacrocin O-methyltransferase
comp97198_c0_seq1	NADE_YEAST	Glutamine-dependent NAD(+) synthetase
comp101309_c0_seq1	Y4233_RHOPA	Putative potassium channel protein RPA4233
comp103360_c0_seq1	RBSK_HUMAN	Ribokinase
comp103541_c0_seq1	TBL41_ARATH	Protein trichome birefringence-like 41
comp106479_c0_seq1	KLHL4_HUMAN	Kelch-like protein 4
comp108302_c0_seq1	MYH3_MOUSE	Myosin-3
comp108536_c0_seq2	CBWD2_HUMAN	COBW domain-containing protein 2
comp110804_c0_seq1	MTG1_MOUSE	Mitochondrial ribosome-associated GTPase 1
comp125204_c0_seq1	RS17_CORA7	30S ribosomal protein S17
comp126348_c1_seq1	UBP26_ORYSI	Ubiquitin carboxyl-terminal hydrolase 26
comp128295_c0_seq1	PKHL1_HUMAN	Fibrocystin-L
comp130956_c0_seq1	CLCN3_CAVPO	H(+)/Cl(-) exchange transporter 3

Figure Legends:

Figure 1: Geographic origin of the strains and metaT samples. Dots and stars indicate the geographic origin of metaT samples and strains, respectively. Several strains and metaT samples may have the same geographic origin, see Supplementary Table 1 and 2.

Figure 2: Strain clustering. Strains are clustered based on nucleotide divergence. colored dots indicate the geographic origin of the strains. Isolation date for strains isolated from water samples or sediment core datations for strains isolated from sediment cores (Supplementary Table 1).The three strain populations NE_A, NE_B and Vigo are indicated.

Figure 3: Precision and bias of Fst estimates based on mRNA sequences. A and B correspond to pairwise Fst among pairs of replicate simulated populations based on PE100 (A) and PE150 (B) strain samples for various coverage levels. C and D correspond to Fst between the population of strains (based on individual strain genotypes) and each of the ten simulated populations, based on PE100 (A) and PE150 (B) strain samples for various coverage levels. R indicates the number of reads subsampled for the simulated populations, C is coverage level used to filter out SNPs (All indicates no filtering, all SNPs are considered), S is the number of SNP analyzed.

Figure 4: Population structure for strain and metaT samples. Genetic variation along four axes following PCA analysis. The results correspond to the pruned 5P6 dataset, but similar results were obtained for all the other datasets (not shown). PC1 and 2 are represented on A and PC3 and PC4 on B. The inset indicates the percentage of variance explained by each principal component. Colors indicate the geographic origin of the strains and symbols indicate strains (circles) and metaT datasets (triangles).

Figure 5: Boxplot indicating the pairwise Fst between and within populations based on metaT samples. The results correspond to the 5P6 dataset, but similar results were obtained for all the other datasets (not shown).

Figure 6: Folded Joint Allele Frequency Spectrum (JAFS) comparing the rare allele frequencies in the three populations using metaT samples. Colors indicate the number of SNPs falling in each bin defined by a unique combination of allele frequency in the two populations considered. For each of the three population pairs, The results correspond to the 5P6 dataset, but similar results were obtained for all the other datasets (not shown).

Figure 7: Genome wide diversity inferred from metatT. A. Haplotype diversity (H) moving average (width=100SNPs, steps=20 SNPs) along *A. minutum* linkage groups in the Bay of Brest (purple), Bay of Vigo (red) and Penze Estuary (green) populations. The black line represents the moving average proportion of SNPs displaying different allelic frequencies across the three populations (adjusted p-values < 1e-10). Dots indicate regions falling above or below of the 95th and 5th quartiles, respectively. Background colors correspond to the various linkage groups. B. and C. Haplotype diversity (purple, red and green, color code as in A.) and proportion of significant SNPs (black broken lines) moving average (width=50SNPs, steps=10 SNPs) in linkage groups L1 and L37, respectively. The number of SNPs in each linkage group is indicated

Supplementary Figure Legends

Supplementary Figure 1: Graphical summary illustrating the estimation of allelic frequencies based on the population of strains (Strains), simulated populations (Simulated) and metaT samples (metaT). For the population of strains, strain genotypes and allelic frequencies were obtained after individually aligning strain RNAseq reads on *A. minutum* reference transcriptome. For the simulated populations, strain RNAseq reads were pooled, subsampled at several coverage levels (see Material and Methods) and aligned to *A. minutum* reference transcriptome before extracting allelic frequencies. For metaT, metaT reads were aligned to a metareference transcriptome (see Material and Methods) and only reads aligning to A. minutum contigs were considered to analyze allelic frequencies.

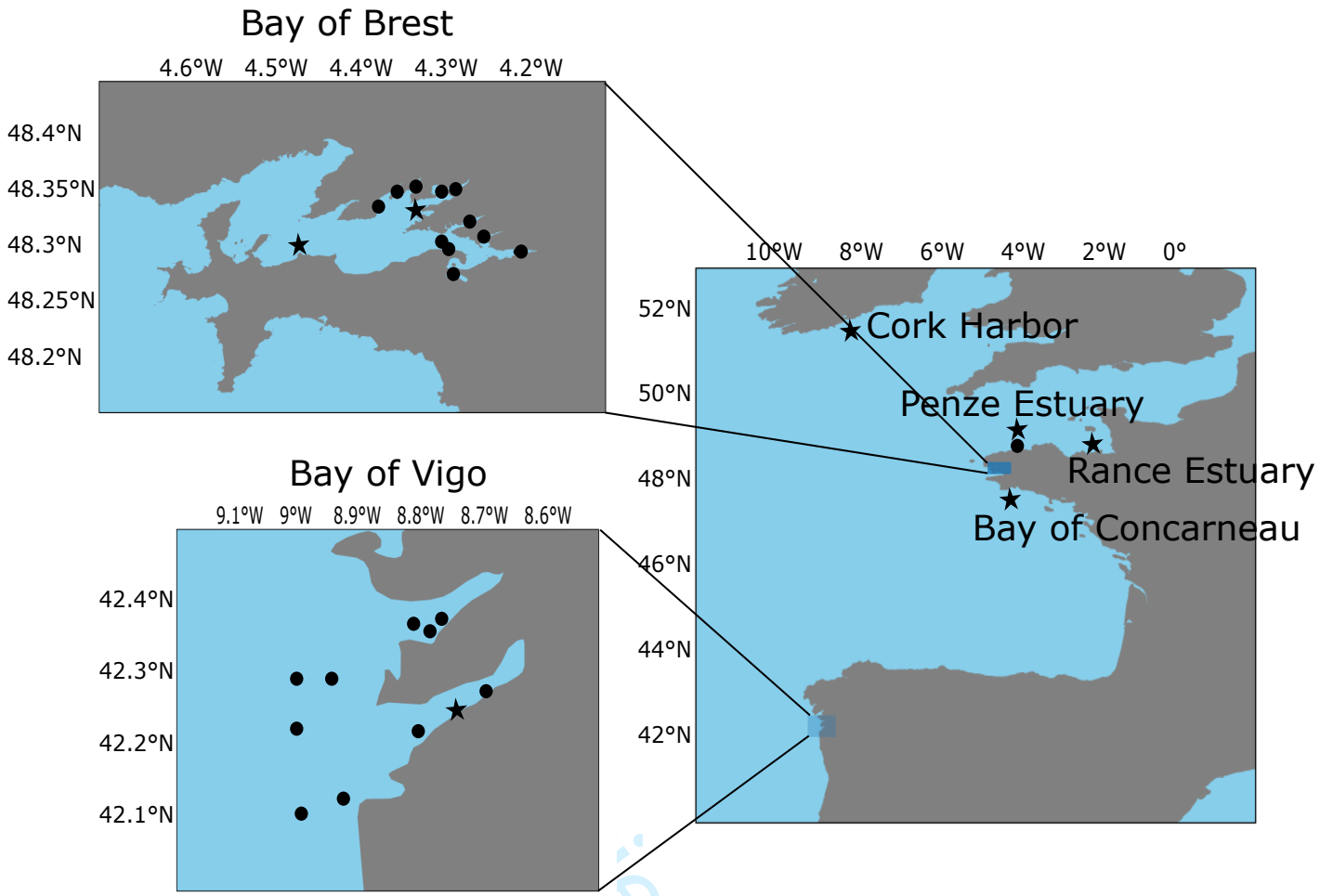
Supplementary Figure 2: Relationship between sequencing strategy and Fst estimates. Five metaT samples were sequenced using both 2x100bp and 2x150bp reads. A. For each sample, Fst between 2x100bp and 2x150bp datasets as a function of SNP coverage. C5, C10, C20, C30, C50, C100, C250, indicate that only SNP with a coverage higher than 5, 10, 20, 30, 50, 100 and 250 were considered. B. Number of SNP considered for each metaT sample pair at the seven coverage levels.

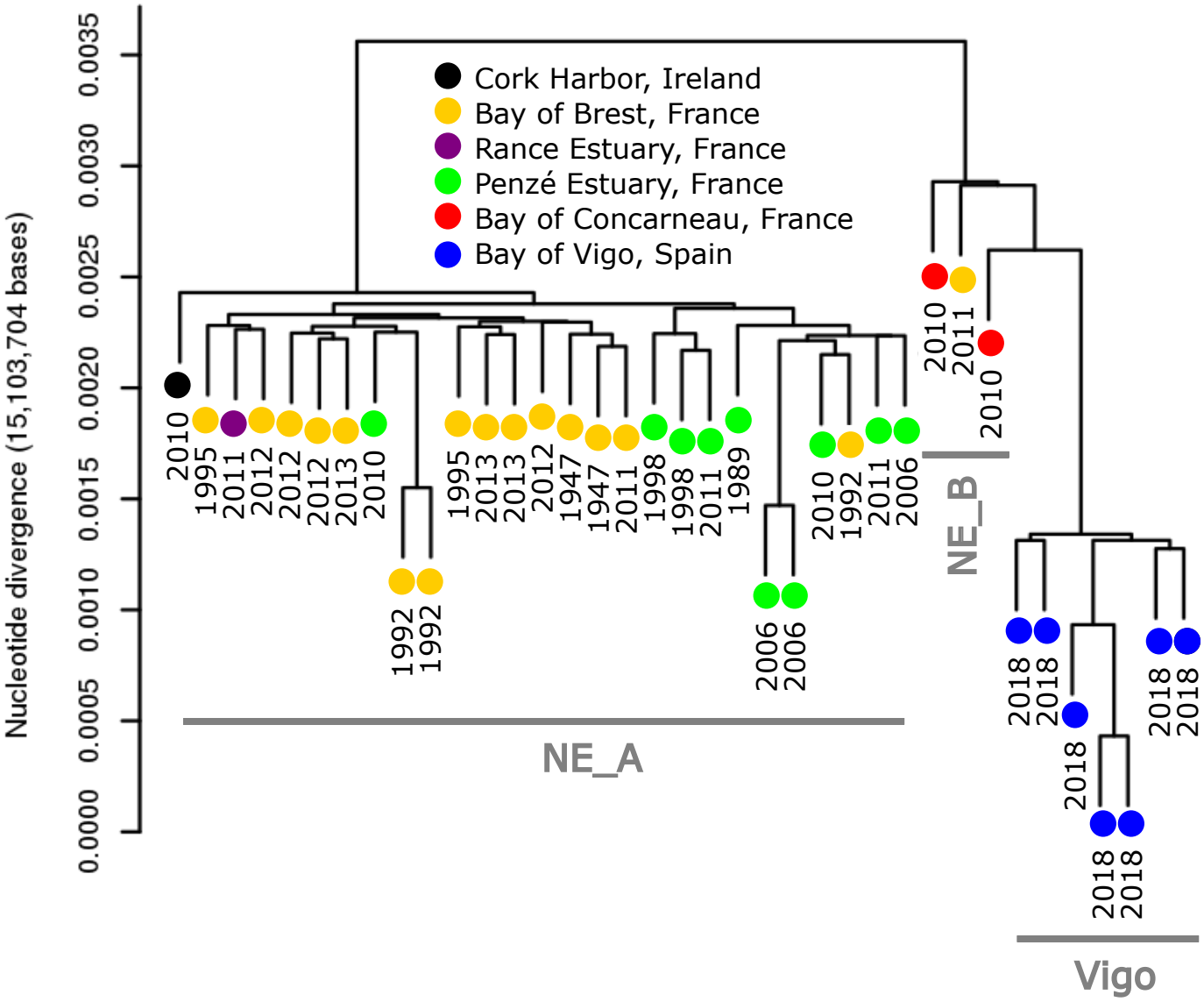
Supplementary Figure 3: Reference allele frequency of population diagnostic SNPs. Population diagnostic SNPs (SNPs displaying one allele in all strains from a given population and the alternative allele in the two other clades) were identified for NE_A (112 SNPs), NE_B (31 SNPs) and Vigo populations (76 SNPs). For each diagnostic SNP (x-axis), the reference (as observed in the reference transcriptome) allele frequency (averaged for metaT samples from the Bay of Brest and Vigo) obtained from metaT samples is indicated for the Penze Estuary (Blue), the Bay of Vigo (Red) and the Bay of Brest (Black) metaT samples. The results correspond to the 5P6 dataset, but similar results were obtained for all the other datasets (not shown).

Supplementary Figure 4 : Boxplot indicating the pairwise Fst between and within populations based on metaT samples using de novo SNPs. The results correspond to the 5P6 dataset, but similar results were obtained for all the other datasets (not shown).

Supplementary Figure 5 : Folded Joint Allele Frequency Spectrum (JAFS) comparing the rare allele frequencies in the three populations using de novo SNPs identified from metaT samples. Colors indicate the number of SNPs falling in each bin defined by a unique combination of allele frequency in the two populations considered. For each of the three population pairs, The results correspond to the 5P6 dataset, but similar results were obtained for all the other datasets (not shown).

Supplementary Figure 6: Genome wide diversity using de novo SNPs identified from metaT samples. A. Haplotype diversity (H) moving average (width=100SNPs, steps=20 SNPs) along *A. minutum* linkage groups in the Bay of Brest (purple), Bay of Vigo (red) and Penze Estuary (green) populations. The black line represents the moving average proportion of SNPs displaying different allelic frequencies across the three populations (adjusted p-values < 0.05). Dots indicate regions falling above or below of the 95th and 5th quartiles, respectively. Background colors correspond to the various linkage groups. B. and C. Haplotype diversity (purple, red and green, color code as in A.) and proportion of significant SNPs (black broken lines) moving average (width=50SNPs, steps=10 SNPs) in linkage groups L1 and L37, respectively. The number of SNPs in each linkage group is indicated




Cork Harbor, Ireland
Bay of Brest, France
Rance Estuary, France
Penzé Estuary, France
Bay of Concarneau, France
Bay of Vigo, Spain
Nucleotide divergence (15,103,704 bases)
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0.0035
2010
1995
2011
2012
2012
2012
2013
2010
1992
1992
1995
2013
2013
2012
1947
1947
2011
1998
1998
2011
1989
2006
2006
2010
1992
2011
2006
NE_A
2010
2011
2010
NE_B
2018
2018
2018
2018
2018
2018
Vigo