Received: 9 December 2021 | Accepted: 9 August 2022

DOI: 10.1111/1462-2920.16171

RESEARCH ARTICLE

environmental microbiology sfam

Genomic differentiation of three pico-phytoplankton species in the Mediterranean Sea

Ophélie Da Silva[1,2] | **Sakina-Dorothée Ayata**[1,2,3] | **Enrico Ser-Giacomi**[3,4] | **Jade Leconte**[5] | **Eric Pelletier**[5,6] | **Cécile Fauvelot**[1,7] | **Mohammed-Amin Madoui**[8] | **Lionel Guidi**[1,6] | **Fabien Lombard**[1,6,9] | **Lucie Bittner**[2,9]

[1]Sorbonne Université, CNRS, Laboratoire d'Océanographie de Villefranche, LOV, Villefranche-sur-Mer, France

[2]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France

[3]Sorbonne Université, UMR 7159 CNRS-IRD-MNHN, LOCEAN-IPSL, Paris, France

[4]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[5]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

[6]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, Paris, France

[7]Institut de Recherche pour le Développement (IRD), UMR ENTROPIE, Nouméa, New Caledonia

[8]Service d'Etude des Prions et des Infections Atypiques (SEPIA), Institut François Jacob, Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Université Paris Saclay, Fontenay-aux-Roses, France

[9]Institut Universitaire de France (IUF), Paris, France

Correspondence
Ophélie Da Silva, Laboratoire d'Océanographie de Villefranche, 181, chemin du Lazaret, 06230 Villefranche-sur-Mer, France.
Email: oph.dasilva@gmail.com.

Funding information
Simons Foundation, Grant/Award Number: CBIOMES #549931; French Ministry of Higher Education, Research and Innovation; MASTODON program MITI, CNRS; Émergence program of Sorbonne Université

Abstract

For more than a decade, high-throughput sequencing has transformed the study of marine planktonic communities and has highlighted the extent of protist diversity in these ecosystems. Nevertheless, little is known relative to their genomic diversity at the species-scale as well as their major speciation mechanisms. An increasing number of data obtained from global scale sampling campaigns is becoming publicly available, and we postulate that metagenomic data could contribute to deciphering the processes shaping protist genomic differentiation in the marine realm. As a proof of concept, we developed a findable, accessible, interoperable and reusable (FAIR) pipeline and focused on the Mediterranean Sea to study three a priori abundant protist species: *Bathycoccus prasinos*, *Pelagomonas calceolata* and *Phaeocystis cordata*. We compared the genomic differentiation of each species in light of geographic, environmental and oceanographic distances. We highlighted that isolation-by-environment shapes the genomic differentiation of *B. prasinos*, whereas *P. cordata* is impacted by geographic distance (i.e. isolation-by-distance). At present time, the use of metagenomics to accurately estimate the genomic differentiation of protists remains challenging since coverages are lower compared to traditional population surveys. However, our approach sheds light on ecological and evolutionary processes occurring within natural marine populations and paves the way for future protist population metagenomic studies.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Environmental Microbiology* published by Society for Applied Microbiology and John Wiley & Sons Ltd.

INTRODUCTION

Single-celled eukaryotes or protists are major contributors to the diversity of plankton in the oceans (de Vargas et al., 2015; Moon-van der Staay et al., 2001). They encompass a myriad of lifestyles, trophic modes, as well as morphological characteristics (Caron et al., 2012, 2017) and play key roles in the functioning of marine pelagic ecosystems, impacting trophic dynamics and global biogeochemical cycles (Biard et al., 2016; Gasol & Kirchman, 2018). Protists have long been under-explored especially from a genomic point of view (Del Campo et al., 2014; Sibbald & Archibald, 2017). The scarcity of reference genomic data for protists results in a misunderstanding of the processes that underpin the contemporary distribution of genetic diversity in natural populations (Lebret et al., 2012; Logares, 2011). Protists are supposed to have vast population sizes and the potential for long-distance dispersal (Dolan, 2005; Watts et al., 2013), which results in reduced evolutionary diversification processes (Finlay, 2002). In comparison with most macro-organisms, protists are thus expected to have little opportunity for allopatric divergence and could show low levels of spatial genetic structure. Studies from the last decade, based on high-throughput sequencing from natural communities via metabarcoding, have unveiled a high diversity of protist species, revealing both endemic and cosmopolitan species (Bittner et al., 2013; Forster et al., 2015; Logares et al., 2014; Malviya et al., 2016). Moreover, the very few genomic studies based on protist 'model micro-organisms' such as *Emiliana huxleyi* (Read et al., 2013) or *Ostreococcus tauri* (Blanc-Mathieu et al., 2017) highlighted a large intraspecific diversity in marine ecosystems. Consequently, even if marine protists have the potential for high dispersal through the currents, protist population structure has been frequently described at local (Evans et al., 2005), regional (Casteleyn et al., 2009) and even at global (Casteleyn et al., 2010) scales, and several processes were reported as drivers of their diversification (Sjöqvist et al., 2015).

In the marine realm, gene flow among planktonic populations can be driven by marine currents, abiotic (i.e. physico-chemical) environmental conditions as well as biotic factors. Oceanographic currents support directional dispersal, conditioning the physical connectivity between distant populations patterns (Godhe et al., 2013; Riginos et al., 2016). They have been identified as major drivers for the structuring of marine populations (Casabianca et al., 2012; Nagai et al., 2007). The asymmetric migration patterns associated could additionally favour local adaptation (Kawecki & Holt, 2002; Sjöqvist et al., 2015). Genetic differentiation could also be driven by natural selection through environmental conditions such as silicate and nitrate/nitrite concentrations (Gao et al., 2019), or changes in salinity (Godhe et al., 2016; Sjöqvist et al., 2015), light or temperature (Latorre et al., 2021; Mena et al., 2019).

To date, population genetic studies have focused on a restricted number of protist species and on sparse genomic markers. With the expansion of high-throughput sequencing, single nucleotide polymorphisms (SNPs) analysis is emerging as a powerful approach to infer population genetic differentiation among natural populations. SNPs are abundant, randomly distributed in genomes, and show low mutation rates and low false genotyping rates, while representing fair statistical power (Selkoe et al., 2016). SNP detection methods consist in mapping short sequences (reads) obtained by high-throughput sequencing on longer reference sequences. Recent studies started to provide a metagenome-level description of the ecological preferences for a few protists (Leconte et al., 2020; Seeleuthner et al., 2018; Vannier et al., 2016) and one of them analysed proxies of species obtained from a genomic reference-free method (metavariant species; Laso-Jadart et al., 2021). However, to our knowledge, genetic differentiation of protists from metagenomes has not been explored in a systematic way and there are no guidelines for the implementation of the ecological and evolutionary processes that are shaping their diversity at the species scale.

The objective of our study was to develop an original bioinformatics pipeline aiming to exploit the currently available metagenomic data for the characterization of genomic differentiation of protists in the marine ecosystems. To address this issue, we focused on the Mediterranean Sea, which is an ideal location to study population genomics (i.e. a semi-enclosed marginal sea characterized by tortuous coastlines, with several environmental gradients despite a highly dynamic circulation, Ayata et al., 2018), on three a priori abundant planktonic protistan species in this area (de Vargas et al., 2015). We gathered reference sequences and metagenomic data previously published for which genomic differentiation could be highlighted at the species scale and tentatively explained by external drivers, such as geography, environmental conditions and oceanographic circulation. We hypothesized that genomic differentiation is greater among distant populations (Wright, 1943) and/or among populations sampled in different hydrological environments (Wang & Bradburd, 2014). We obtained contrasted results for all three species, which allowed us to discuss how current metagenomic data could support and provide new resources for population genomics studies for overlooked but abundant and ecologically relevant organisms.

EXPERIMENTAL PROCEDURES

Our global analysis strategy for population genomics based on metagenomic data is summarized in

Supporting Information S8 and all scripts and data are openly available on https://github.com/opheliedasilva/popmetag.

Selection of protist species for the study of genomic populations

In order to study genomic differentiation within marine protist populations, we chose to exploit metagenomic data collected in the Mediterranean Sea during the Tara Oceans (TO) expedition (Alberti et al., 2017), in which protists prevailed from pico- to microplankton size fractions (i.e. from 0.8 up to 180 μm). We first analysed the abundance of the V9 eukaryotic metabarcodes in the TO Mediterranean samples (de Vargas et al., 2015) to identify dominant taxa over all stations for which a genomic/transcriptomic reference was available (Supporting Information S1). From there, we selected three phylogenetically distinct planktonic species: *B. prasinos* (Eikrem & Throndsen, 1990), *P. calceolata* (Andersen et al., 1993) and *P. cordata* (Zingone et al., 1999). While transcriptomes were used for *P. calceolata* and *P. cordata* (RCC969 and RCC1383, respectively), a reference genome was used for *B. prasinos* (RCC1105). The high gene density of *B. prasinos* genome (Moreau et al., 2012) was assumed to limit the impact of intergenic regions in the analysis (Supporting Information S7). Given the size range of these organisms, we focused on the 0.8–5 μm size fraction in the TO data, consisting of 13 metagenomics samples available from the surface layer (accession number PRJEB4352, Carradec et al., 2018; Supporting Information S1), containing on average 185 million sequence reads.

We built a bioinformatic pipeline to extract single nucleotide polymorphisms (SNPs) from metagenomic sequences in comparison to reference sequences (here genome or transcriptome). It consisted of five steps (detailed in the Supporting Information S2, and the whole bioinformatic pipeline is available on GitHub): (1) checking the quality of the metagenomic reads to remove the low-quality ones (Trimmomatic; Bolger et al., 2014), (2) mapping the metagenomic reads on the reference sequences to pull out reads of the targeted species (bwa mem; Li, 2013), (3) filtering aligned reads, first to remove low complexity sequences and avoid spurious alignments (PRINSEQ; Schmieder & Edwards, 2011) and second to reduce the recruitment of reads from a closely related species (reads aligned with less than 95% identity were removed; Vannier et al., 2016), (4) detecting genomic variants in comparison with the reference sequence, and (5) filtering the variants in order to only keep the SNPs (e.g. indels were removed). To minimize false positives, SNPs were filtered based on their vertical coverage (i.e. mean number of reads aligned at each position of the assembly): we only kept SNPs supported by at least four reads but less than $\mu + 2\sigma$ of vertical coverage (μ is the mean and σ is the standard deviation of SNP vertical coverage, in order to remove SNPs also abundant in closely related species). The output of our pipeline corresponded to an abundance table of the SNPs in each station. Based on this output, we computed for each SNP the frequency of alleles in each station. F_{ST} were calculated for each SNP at each pair of stations. As the number of allelic frequencies greatly varied from one locus to another, average frequency used in F_{ST} calculation was always computed for the two stations considered and not for all of them. Finally, the genomic differentiation between each pair of stations (pairwise F_{ST}) corresponded to the median F_{ST} and was used as the genomic distance between populations. For *B. prasinos*, 20 pairs of stations had no shared SNP and therefore no genomic differentiation was computed for these pairs of stations. For *P. calceolata*, one station had no common SNPs with all the others and was therefore removed from the analysis. For each species, the global F_{ST} was computed as the mean pairwise F_{ST}. For each species, we created heatmaps to visualize genomic distances between pairs of stations (Figure 5). Dendrograms, built by hierarchical clustering (complete linkage), were added on the heatmaps to help identifying groups of stations genetically close. The missing genomic distances (20 values out of 78 for *B. prasinos*) have been replaced by the mean value to perform the clustering.

Calculations of geographic, environmental and oceanic distances

Firstly, the latitude and longitude of sampling (metadata from PANGAEA; Pesant et al., 2015) were used to compute geographic distances among pairs of stations (i.e. minimal distances between two stations without crossing the lands). Secondly, each TO sample was associated with its hydrological and biogeochemical environment based on geographic coordinates and depth (TO metadata, Pesant et al., 2015). The environmental variables extracted from Medatlas-II climatologies (Fichaut et al., 2003) included surface temperature (°C), surface salinity (PSU) and surface concentrations of ammonium (mmol.m^{-3}), oxygen (ml.l^{-1}), nitrate (mmol.m^{-3}), nitrite (mmol.m^{-3}), phosphate (mmol.m^{-3}), silicate (mmol.m^{-3}) and chlorophyll a (mg.m^{-3}). A PCA was performed on these normalized and standardized variables (Legendre & Legendre, 2012) for a total of 22 TO stations (13 stations corresponding to our metagenomic sample and nine additional TO stations). The environmental distances between pairs of stations were computed as their Euclidean distances in the PCA space (using only significant axes based on the Kaiser–Guttman criterion) and the stations were

clustered using hierarchical clustering. Thirdly, physical transport by ocean circulation was estimated with Lagrangian model simulations to compute oceanographic distance between stations. As no assumption about mechanisms of dispersal and underlying model settings for each organism could be established, data from two types of existing models were used to assess Lagrangian transport in the Mediterranean Sea. The first Lagrangian dataset was a product of Berline et al. (2014) providing the mean connection time (MCT, in days) for each pair of stations. The second dataset was obtained by performing simulations of the model of Ser-Giacomi et al. (2015) based on the Lagrangian flow network approach. The dataset corresponded to a connectivity matrix estimating the probability of connection (PC) for a particle leaving a station to end up in another station in a given time period. To estimate PC over time, three dispersal durations were used (3, 6, and 12 months) and averaged to integrate dispersal characteristics at various temporal scales (seasonal, biannual, and annual circulation). As Lagrangian matrices are asymmetric, we chose the maximum PC and the minimum MCT as oceanographic distance for each pair of stations. More details about Lagrangian datasets are provided in Supporting Information S5.

Statistical analyses

The links between genomic distances and geographic, environmental and oceanographic circulation constraints were assessed through linear regressions using the following model:

$$y = \beta_0 + \beta_{geo} x_{geo} + \beta_{env} x_{env} + \beta_{pc} x_{pc} + \beta_{mct} x_{mct} + \varepsilon$$

where y corresponds to the normalized genomic distances $F_{ST}/(1-F_{ST})$, β_0 is the intercept coefficient (i.e. predicted response when pairs of stations are not distant in terms of geography, environment or oceanographic circulation). The geographic (β_{geo}), environmental (β_{env}) and oceanographic (β_{pc} and β_{mct}) coefficients quantify the effects of geographic, environmental and oceanographic distances (respectively x_{geo}, x_{env}, x_{pc} and x_{mct}) on the genomic distances. ε is the error term (i.e. random component between the variable to explain and the explanatory variable). For each species, we selected the optimal model by an exhaustive search procedure using the Bayesian information criterion (BIC). Fisher tests (null vs. selected model) were carried out to test the overall significance of the linear models. Variables impacting genomic differentiation were identified with Student tests using a threshold of 0.05.

All analyses were conducted in R (v3.5.0; R Core Team, 2020) using the packages: tidyverse (v1.2.1; Wickham et al., 2019), ggreprel (v0.8.1; Slowikowski, 2020), ggpubr (v0.3.0; Kassambara, 2020), maps (v3.3.0; Becker et al., 2018), and viridis (v0.5.1; Garnier, 2018) for graphs and maps; and FactoMineR (v1.42; Lê et al., 2008); gdistance (v1.3; van Etten, 2017), and leaps (v3.0, Lumley & Miller, 2013) for statistical analysis.

RESULTS

Genomic distances from metagenomic samples based on a selection of three protist species

The genomic distances were computed for three phylogenetically distinct planktonic species (Supporting Information S1): *Bathycoccus prasinos* (Chlorophyta; Eikrem & Throndsen, 1990), *Pelagomonas calceolata* (Ochrophyta; Andersen et al., 1993) and *Phaeocystis cordata* (Haptophyta; Zingone et al., 1999). These three species are widespread at global scale, with an ubiquitous distribution for *B. prasinos* (Moreau et al., 2012; Vannier et al., 2016) and *P. calceolata* (Worden et al., 2012), or present in many areas for *P. cordata* (i.e. Red Sea, Indian Ocean, Mediterranean Sea; Decelle et al., 2012). *B. prasinos* is a major contributor of the primary production and shows a seasonal cycle in the Mediterranean Sea (Lambert et al., 2019; Moreau et al., 2012). *P. calceolata*, involved in nitrate assimilation (Dupont et al., 2015), has been overlooked in the Mediterranean Sea. Finally, *P. cordata* has been detected in free-living mode and in symbiosis with Acantharia in the Mediterranean Sea (Decelle et al., 2012). A reference genome was selected for *B. prasinos* (Moreau et al., 2012), whereas, at the time of our analysis, only transcriptomes were publicly available for *P. calceolata* and *P. cordata* (Johnson et al., 2019; Keeling et al., 2014). The cumulative length of these assemblies varied by twofold from *P. cordata* to *P. calceolata* (respectively, 9.4 and 21 Mb), while the length of *B. prasinos* assembly was intermediate (15 Mb). Based on the cell size of these organisms (Supporting Information S1), we focused on the 0.8–5 μm size fraction of the Tara Oceans (TO) dataset, consisting of 13 metagenomic samples collected at the surface layer from 13 stations in the Mediterranean Sea (Figure 1, accession number PRJEB4352, Carradec et al., 2018; see Supporting Information S2 for details). We mapped the metagenomic reads (on average 185 million reads/sample) on reference assemblies and obtained horizontal and vertical coverages between species (Figure 2; percentage of the reference covered by at least one read and mean number of reads aligned at each position of the reference, respectively). *B. prasinos* and *P. calceolata* displayed higher horizontal coverages than *P. cordata* despite their longer reference sizes (maximal horizontal coverage, *B. prasinos*: 98.1%, *P. calceolata*: 68.2%,

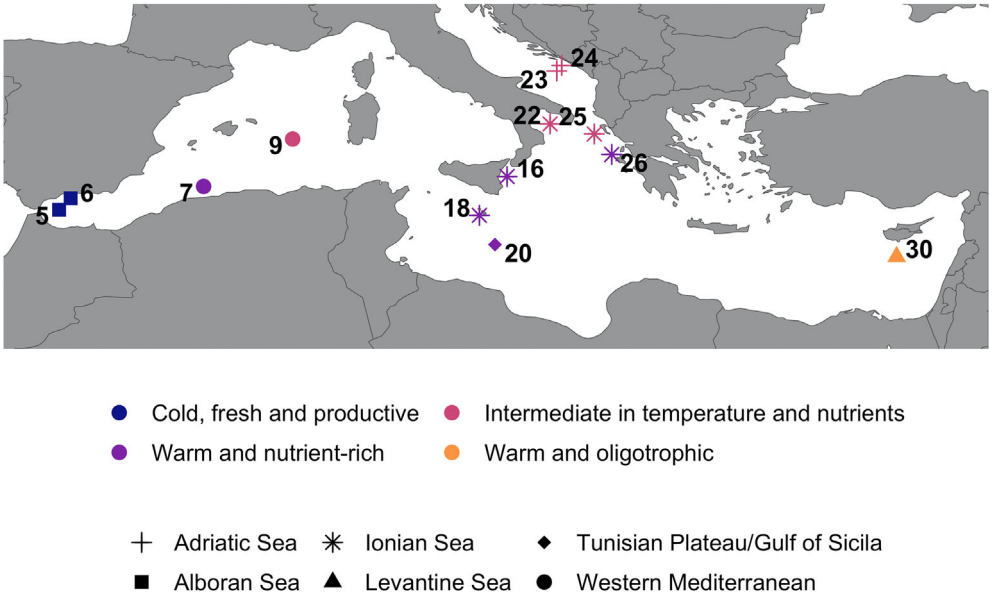

FIGURE 1 Geographic location of the 13 stations sampled during the Tara oceans for metagenomic analysis. Stations are indicated by numbers (with increasing numbers from west to east, following the Tara oceans cruise trajectory). Geographic entities are based on the marine ecoregions of the world (Spalding et al., 2007) and represented by different shapes. Environmental entities determined through principal component analysis (PCA) are indicated by colours (defined in Figure 4).

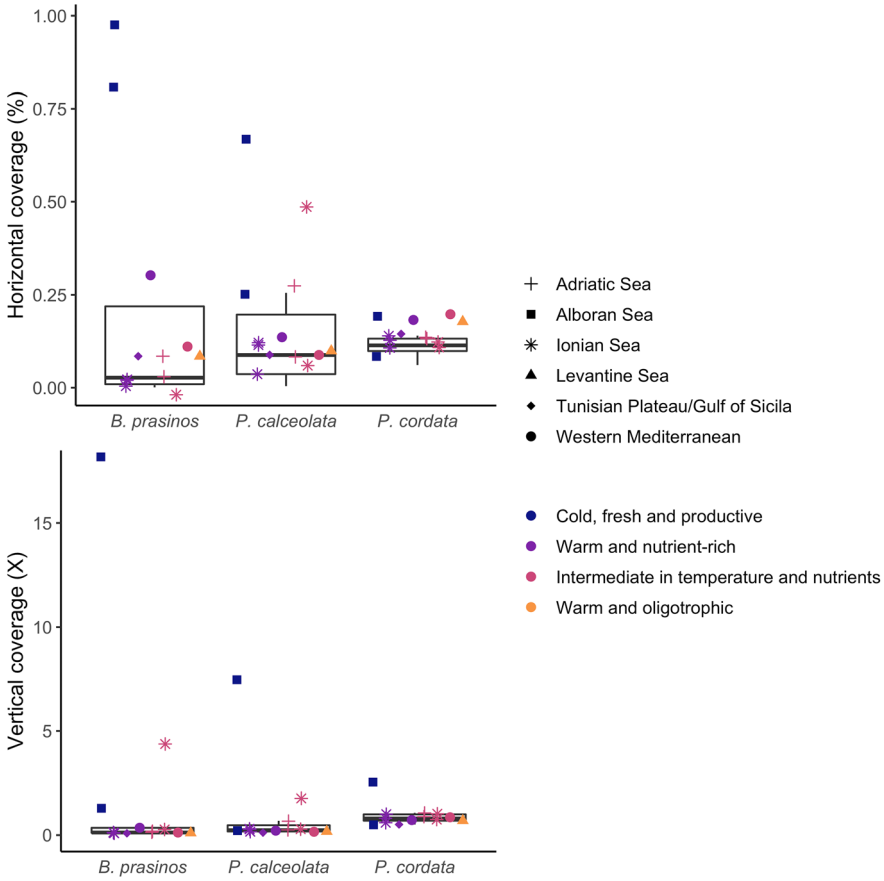

FIGURE 2 Distributions of horizontal and vertical coverages for each species within the 13 Mediterranean Sea stations (i.e. percentage of reference covered by at least one read and mean number of reads aligned at each position of the reference, respectively). Type of shape: geographic entities (defined in Figure 1). Colours: environmental entities (defined in Figure 4).

P. cordata: 19.6%). The mean vertical coverage was also more heterogeneous for *B. prasinos* and for *P. calceolata* than for *P. cordata* (min–max mean vertical coverage, *B. prasinos*: 0.085–17.6 X, *P. calceolata*: 0.119–7.2 X, *P. cordata*: 0.51–2.47 X). *B. prasinos* showed highest coverages at stations 5 and 6, located in the Western part of the Mediterranean Sea (Alboran Sea).

SNPs were detected from aligned and filtered reads (Experimental procedures, Supporting Information S2). *B. prasinos*, *P. calceolata* and *P. cordata* showed different total numbers of SNPs over reference size ratio (respectively, 3.4, 51.4 and 0.5 SNPs/Mb). In average, only 9.26%, 9.67% and 18.74% of the total number of SNPs were observed in each station for *B. prasinos*, *P. calceolata* and *P. cordata*, respectively. Moreover, a filtering on the vertical coverage led to the removal of SNPs (i.e. SNPs having between 4 and $\mu + 2\sigma$ vertical coverage were kept; Experimental procedures, Supporting Information S2). Consequently, pairwise F_{ST} (i.e. median genomic distance as defined in the Wright's formulation, where 0 indicates no genomic differentiation and 1 means maximal genomic differentiation) between all station pairs were computable for *P. cordata*, whereas for *P. calceolata*, the station 30 had to be removed, and for *B. prasinos* 20 pairwise F_{ST} were not computable. We obtained 78 pairwise genomic distances (pairwise F_{ST}) for *P. cordata*, 66 for *P. calceolata* and 58 for *B. prasinos*.

Our results show that the Mediterranean Sea populations of *B. prasinos* had a stronger global genomic differentiation (global $F_{ST} = 0.136$) than *P. calceolata* (global $F_{ST} = 0.066$) and *P. cordata* (global $F_{ST} = 0.045$) (Figure 3). *B. prasinos* also showed the most contrasted genomic differentiation (pairwise F_{ST}) ranging from little (0.012) to very high (0.476)

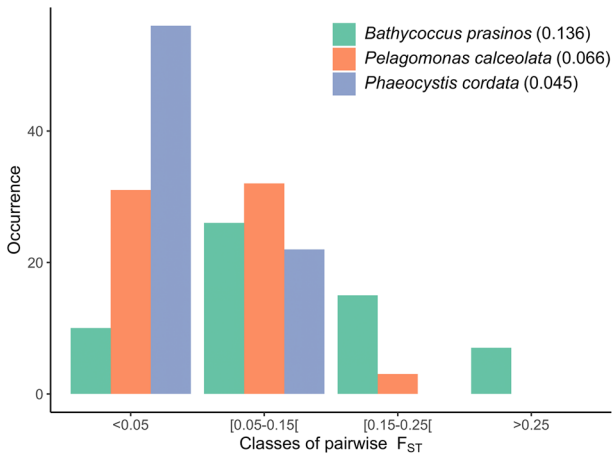

FIGURE 3 Distribution the genomic distances for each species between the 13 Mediterranean Sea stations. The pairwise F_{ST} were grouped following Hartl and Clark (1997) by four classes of genomic differentiation (<0.05: little, [0.05–0.15]: moderate, [0.15–0.25]: high, >0.25: very high).

(Figure 3). *B. prasinos* was also the only species showing very high genomic differentiation values (seven F_{ST} values > 0.25). *P. calceolata* showed little to high genomic differentiation (F_{ST} ranging from 0.019 to 0.181), whereas *P. cordata* had the lowest F_{ST} values (ranging from 0.026 to 0.069). The maximum F_{ST} for *P. cordata* was observed in the little genomic differentiation class, whereas *P. calceolata* and *B. prasinos* F_{ST} peaked in the moderate genomic differentiation class (Figure 3).

Geographic, environmental and oceanographic distances between stations

Geographic distances were computed as the minimal distances between each pair of stations, and ranged between 33.55 and 3624.49 km (mean: 1255.13 km; Figure 1, Supporting Information S3).

A principal component analysis (PCA) was performed on the environmental conditions of the stations, in order to infer environmental distances between each pair of stations. To strengthen the statistics, we analysed the 13 stations corresponding to our metagenomic samples and nine additional TO stations also sampled in the Mediterranean Sea (Figure 4). Three significant axes were identified: the first PCA axis (Dim1, 48.6% of the total variance) distinguished the warmer and more oligotrophic stations (Dim1 < 0) from the colder and nutrient richer ones (Dim1 > 0); the second PCA axis (Dim2, 18.8% of the total variance) separated the saltier and ammonium-rich stations (Dim2 > 0) from the less salty and ammonium-poor stations (Dim2 < 0); and the third PCA axis (Dim3, 15.2% of the variance) divided the silicate-rich stations (Dim3 < 0) from the phosphate and nitrite-rich ones (Dim3 > 0). Environmental distances were calculated as the variance-weighted Euclidean distances on the first three dimensions (Supporting Information S4).

Oceanographic distances were inferred from Lagrangian modelling both as the mean connection times (data from Berline et al., 2014) and as connection probabilities among stations (Ser-Giacomi et al., 2015). They ranged between 46.15 and 545.9 km (mean: 219.9 km) and 1.72×10^{-7} and 9.07×10^{-3} (mean: 5.98×10^{-4}), respectively (Supporting Information S5).

Deciphering drivers of genomic differentiation

For each of the three planktonic protist species, the link between genomic differentiation and geography, environment or oceanographic circulation was assessed through linear regression models. We used Fisher tests to show that our linear models explain the genomic differentiation of *B. prasinos* and *P. cordata* better than the null models (Table 1). In contrast, our model did not

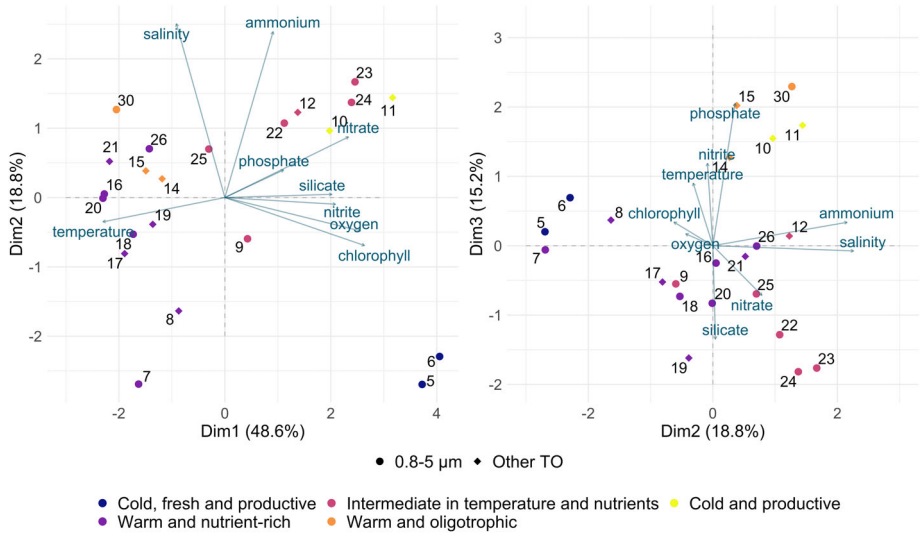

FIGURE 4 Environmental characterization of the Mediterranean Sea stations. Principal component analysis (PCA) of the environmental variables retrieved at 22 stations indicated by points and numbers. The 13 TO stations studied, reported with dots, and 9 extra stations are reported with diamonds. Colours of the points indicate the environmental clusters the stations belong to (hierarchical clustering). Environmental variables are represented with blue arrows. The significant axes (Dim1, Dim2, Dim3) are presented with their corresponding percentages of variance.

TABLE 1 Results of statistical models testing the impact of geography, environment, and ocean circulation on genomic distances of three planktonic species

Species	Selected versus null model	Selected model	R^2
Bathycoccus prasinos	p value = 0.0025	$\beta_{env} = 0.0382$ p value$_{env} = 0.0025$	15.24
Pelagomonas calceolata	p value = 0.344		
Phaeocystis cordata	p value = 0.0019	$\beta_{geo} = 4.624 \times 10^{-6}$ p value$_{geo} = 0.0019$	12.04

Note: The p value of the overall significance test is provided (selected vs. null model). For each selected model, the value of the regression coefficient β and the associated p values are indicated for the selected variables (env: Environmental distances, geo: Geographic distances), as well as the determination coefficient R^2.

provide a better fit than the null model for *P. calceolata*. The optimal selection model procedure led to the selection of environmental distances for *B. prasinos* and of geographic distances for *P. cordata* (Table 1). Selected models explained 15.24% of the genomic differentiation for *B. prasinos* and 12.04% for *P. cordata*. The genomic differentiation of *B. prasinos* significantly increased with environmental distances, whereas the genomic differentiation of *P. cordata* significantly increased with geographic distances.

The highest genomic differentiations of *B. prasinos* were observed between stations 5 (Alboran Sea) and 26 (Ionian Sea) and between stations 16 and 24 (Ionian Sea and Adriatic Sea, respectively; Figures 4 and 5A), whereas the smallest genomic differentiations were observed among stations 16, 20, 25 and 30, all located in the Eastern basin of the Mediterranean Sea. This group of stations was genetically distant from stations 23 and 24 (Adriatic Sea), from stations 5 and 6 (Alboran Sea) and also from station 22 (Ionian Sea) (Figure 5A). These two groups of stations (16, 20, 25, 30 vs. 5, 6, 22, 23, 24) were environmentally separated by the nutrient-temperature gradient (Dim1, Figure 4), with the former stations characterized by warmer and more oligotrophic conditions (Eastern Basin of the Mediterranean Sea).

For *P. cordata*, the highest genomic differentiations were observed between stations from the Western Basin (stations 5, 6, 7, 9, Alboran Sea and Western Mediterranean) and from the Eastern Basin (stations 18, 20, 24, 25, 26, 30, Ionian Sea, Tunisian Plateau/Gulf of Sicilia, Adriatic Sea and Levantine Sea) (Figures 1 and 5B). Exceptions were however observed, with spatially close stations that show high genomic differences (e.g., stations 9 and 23, or 18 and 24), confirming that geographic distances are not the only factor driving the genomic differentiation of this species.

DISCUSSION

Our population genomics approach based on metagenomic data and applied to marine planktonic protists

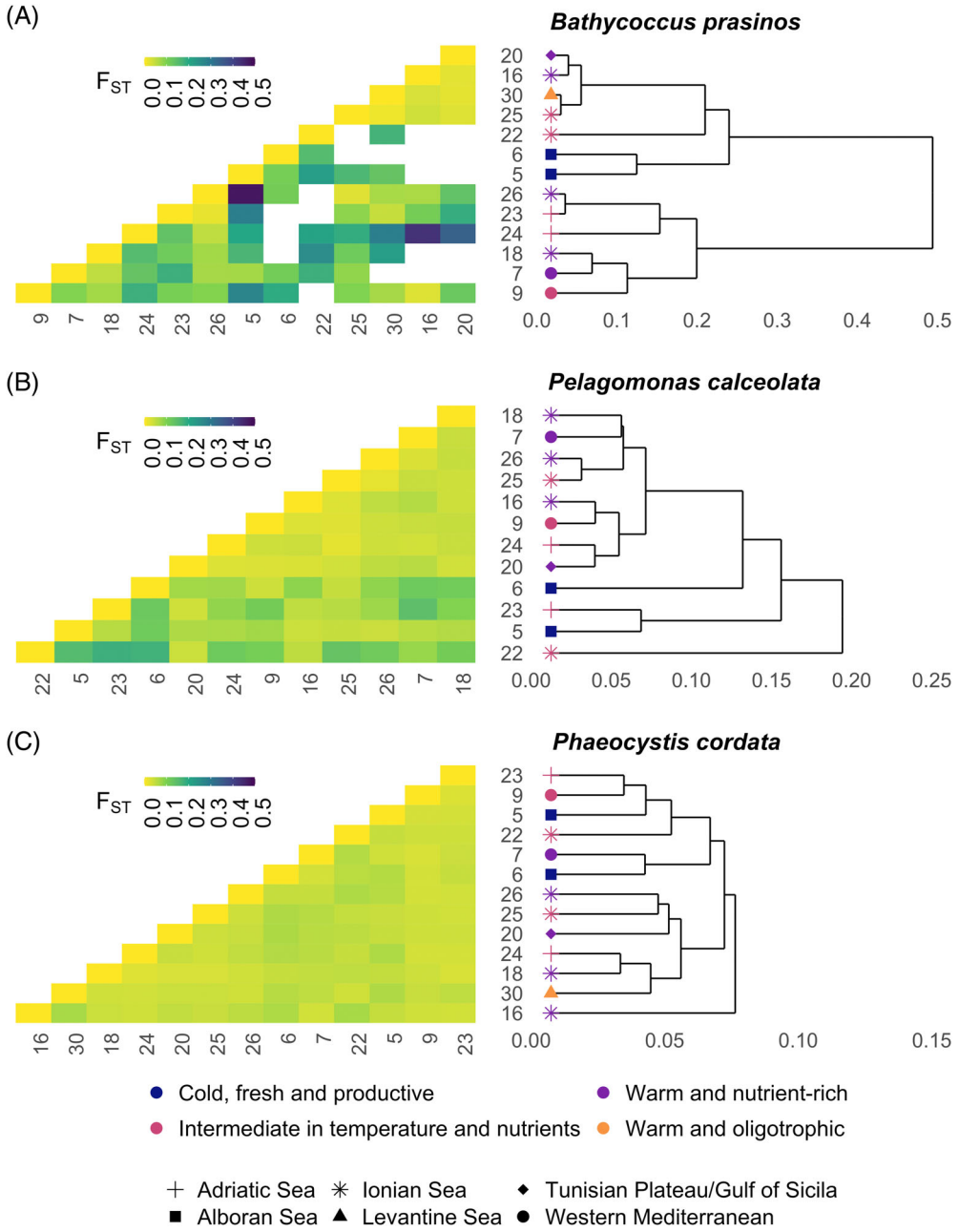

FIGURE 5 Genomic differentiation among the Mediterranean Sea populations. Heatmaps of genomic distances (pairwise F_{ST}) with associated dendrograms obtained by hierarchical clustering for (A) *B. prasinos*, (B) *P. calceolata* and (C) *P. cordata*. For *B. prasinos*, missing values (due to absence of common SNPs) were replaced by the mean value of F_{ST} of the stations. Type of shape: geographic entities (defined in Figure 1). Colours: environmental entities (defined in Figure 4).

offers new insights into the genomic differentiation of understudied organisms and describes the main abiotic drivers shaping it.

Main challenges for population genomics approach based on metagenomics

In this study, we developed a computational method in order to apply population genetic concepts from metagenomic data. This pipeline was designed according to the FAIR principles (Wilkinson et al., 2016) and it can be transposed to any protist lineages as long as reference sequences are available. The number of species that can be investigated with this pipeline is thereby limited by the number of available reference assemblies. Historically, sequencing projects prioritized species of biomedical and biotechnological interest, or species that were easy to cultivate. This has led to strong biases in public molecular databases in which protist reference sequences are under-represented (Del Campo et al., 2014; Keeling & Del Campo, 2017;

Sibbald & Archibald, 2017). In the last decade, an increasing number of sequencing initiatives ha tried to cope for this limitation (Keeling et al., 2014), but most protist lineages remain unrepresented in public databases (e.g. about 200 genomes in the Genome Online Database in February 2017, mostly parasitic; Sibbald & Archibald, 2017). So far, only a few dozen species are available for a focused analysis of the Mediterranean Sea. It is, however, very likely that the increasing availability of reference sequences will allow population genomics' analyses based on metagenomic data in the near future.

Based on metabarcoding data, our three targeted species were potentially abundant in the Mediterranean Sea (Supporting Information S1). However, the reference sequences selected may include biases, as references do not come from the exact same geographic location as the metagenomic data. The *P. calceolata* reference strain (RCC969) was sampled in the southern Pacific Ocean (Lê et al., 2008), which is, as the Mediterranean Sea, an oligotrophic area. However, variations among *Pelagomonas* strains have already been reported, notably driven by different light-acclimation strategies (Kulk et al., 2012; Worden et al., 2012). The two other references, *P. cordata* (RCC1383) and *B. prasinos* (RCC1105), were both isolated in the Mediterranean Sea, respectively, from the Tyrrhenian Sea (Zingone et al., 1999) and from Banyuls' Bay (northwestern Mediterranean Sea; Moreau et al., 2012), but our 13 metagenomic samples were not collected from these areas. The relatively low mapping coverages obtained might reflect the fact that dominant lineages in the natural communities differ significantly from the current reference lineages. Reference lineages come mainly from species in culture or more rarely from species that have been isolated locally from the environment (e.g. single amplified genomes; Del Campo et al., 2014). Therefore, references are to date poorly indicative of the genomic variability in natural populations (Bittner et al., 2013; Laso-Jadart et al., 2021; Worden et al., 2012). Using metagenomic-assembled genomes (MAGs) as new references can partly circumvent this issue, because they correspond to abundant biological units in the studied ecosystems. MAGs have been recently built for microbial eukaryotes, either from metagenomes (Delmont et al., 2022) or from metatranscriptomes (Vorobev et al., 2020). But, as they result from 'consensus assemblies', they can integrate the variability of several sampled organisms or populations, and the biological scale at which the study is carried out (species, genus, biological association/holobiont) remains uncertain (Shaiber & Eren, 2019). The representativity of current MAGs is high, but likely also far from the phylogenetic diversity highlighted by metagenomic studies based on classical de novo assemblies (Carradec et al., 2018 vs. Delmont et al., 2022). Highly abundant but genetically complex lineages (e.g. Dinoflagellates) still fail to be reconstructed from the environment, mainly due to the globally shallow sequencing depth generally achieved for environmental samples.

The main difference between traditional population genetics studies and our metagenomics-based approach is that metagenomic data provides occurrences for the whole community. Hence, this imposes a first analytical procedure during which the sequences of targeted species have to be extracted from the bulk data. Even if the TO metagenomic samples were obtained by the filtration of large seawater volumes and were sequenced to a consequent depth (160 million reads per sample; Alberti et al., 2017), it has previously been shown that the sampling effort for the smallest planktonic fraction did not result in a saturation of the eukaryotic genes (Carradec et al., 2018). Our targeted species, even if theoretically abundant in the Mediterranean Sea, might thus be underrepresented in the metagenomic samples. Nonetheless, we assumed that the direct mapping of metagenomic reads on references (instead of their use through de novo assembly of long sequences) could depict the genomic structure of protist populations (Leconte et al., 2020; Vannier et al., 2016). Filtering parameters were tuned to ensure the detection of good quality variants of our targeted species. Firstly, during the read recruitment step, alignments with less than 95% mean identity were discarded in order to obtain a proper genome abundance estimate for the targeted species. This identity threshold is comparable with previous estimations based on Chlorophyta lineages (Blanc-Mathieu et al., 2017; Leconte et al., 2020) and ensures that all reads recruited belong to the same species, despite intraspecific variation. Future new genomic references on Haptophyta and Ochrophyta will allow refining this filtering step. Secondly, defining a minimal coverage threshold was not straightforward. In studies based on model organisms (i.e. mainly human, mice, few others Metazoa, as well as pathogenic Eubacteria) at least 30 X vertical coverage thresholds are usually expected (Davide & Donati, 2017; Sims et al., 2014), while in our study vertical coverages ranged between 0.085 X and 17.6 X. However, low minimal coverage thresholds (e.g. 4 X) have already been used (e.g. on copepods; Madoui et al., 2017), allowing for first population genomic inferences based on metagenomic data for non-model planktonic species. In addition, the coverage threshold was centred around the mean vertical coverage (maximal coverage threshold of $\mu + 2\sigma$) leading to the removal of SNPs due to the read recruitment of closely related species (Madoui et al., 2017; Laso-Jadart et al., 2020, 2021; Supporting Information S6). Therefore, the three picoeukaryotes well illustrate how far the coverages obtained based on metagenomics are from classical population genomics approaches. Low coverages should be interpreted with caution. Nonetheless,

even if F_{ST} estimates are based on a restricted amount of data, we believe that our results are valuable for the observed trends for genomic differentiation from natural populations of protists.

Protist genomic differentiation in the Mediterranean Sea and its different drivers

A large number of population genetics studies have demonstrated how dispersal and environment impact macro-organisms' population genetic diversity in the Mediterranean Sea (e.g., the striped red mullet, Dalongeville et al., 2018). In contrast, only a few studies have focused on protists. Two studies used microsatellites data for two dinoflagellate species and highlighted the existence of a genetic structuring between the Atlantic Ocean and the Mediterranean Sea (Lowe et al., 2012) as well as through circulation patterns in the Mediterranean Sea (Casabianca et al., 2012). However, most seascape genetic studies rely on the study of pluricellular, macroscopic organisms (Selkoe et al., 2016). High-throughput metagenomics now enable to investigate highly abundant components of the ecosystems (e.g. the *Oithona nana* copepod, Madoui et al., 2017; *Oithona similis*, Laso-Jadart et al., 2020) involving more and more microbes (Delmont et al., 2019; Faure et al., 2021; Laso-Jadart et al., 2021; Leconte et al., 2020; Seeleuthner et al., 2018; Vannier et al., 2016). By analogy, we conducted a comparative study of three picoeukaryotes. While a reference genome was used for *B. prasinos*, reference transcriptomes were used for *P. calceolata* and *P. cordata*. Even if that may impact the observed patterns and their direct comparisons, we hypothesized that the bias is limited due to the compact genome of *B. prasinos*, which is a peculiar characteristic of Mamiellales (Moreau et al., 2012; Supporting Information S7). Our study highlighted that *B. prasinos* (Chlorophyta), *P. calceolata* (Ochrophyta) and *P. cordata* (Haptophyta) exhibit distinct spatial patterns, forced by different external constraints. As our statistical models did not explain a large part of the genomic differentiation for *B. prasinos* and *P. cordata*, and none for *P. calceolata*, our results must be interpreted with caution. For *B. prasinos*, gene flow appeared stronger among populations from similar environments, which suggested an isolation-by-environment pattern (Wang & Bradburd, 2014). Several mechanisms may also explain the genetic differentiation due to environmental forcing, especially local adaptation, non-random mating due to adaptation or phenotypic plasticity (Sexton et al., 2014). Nonetheless, Vannier et al. (2016) identified that environmental conditions such as temperature and light may influence the distribution of *Bathycoccus*. Our isolation-by-environment hypothesis is in line with this assumption and suggests that the environmental conditions might drive the genomic differentiation of this lineage. For *P. cordata*, gene flow decreased with geographic distances, supporting a hypothesis of isolation-by-distance for this species (Wright, 1943). Isolation-by-distance has frequently been suggested as an important driver of genomic differentiation for protists (Casteleyn et al., 2010; Demura et al., 2014; Nagai et al., 2007) and can act at different scales. Indeed, isolation-by-distance has been described as a driver of genomic differentiation for the diatom *Pseudo-Nitschia pungens* at global (Casteleyn et al., 2010), regional (Casteleyn et al., 2009) and local scales (Evans et al., 2005). Finally, the spatial scale of our sampling impacts the detection of genomic isolation in our datasets. Dalongeville et al. (2018) assessed the importance of geographic distances at long-distance spatial scales (>1000 km) and of dispersal constraints at shorter spatial scales (<1000 km) for structuring the genetic diversity of the red mullet fish in the Mediterranean Sea. For the Dinoflagellates, *Oxyrrhis marina* and *O. maritima*, Lowe et al. (2010) identified contrasted genetic structures between the Atlantic and the Mediterranean Sea subpopulations. Future studies could benefit from sampling at finer and higher resolution to better decipher processes shaping protist population differentiation in natural environments.

Protist dispersal and oceanographic circulation

Since planktonic protists are supposed to have potential for long distance dispersal and large population sizes (Cermeño & Falkowski, 2009; Finlay, 2002), the circulation may not impact their genomic differentiation (Cermeño & Falkowski, 2009; Gibbons et al., 2013; Hellweger et al., 2014). However, *P. cordata* showed a pattern of isolation-by-distance at the scale of the Mediterranean Sea, which was surprisingly stronger with geographic distance rather than with oceanographic distance. Indeed, oceanographic distance better represents the asymmetric dispersal of plankton. We thus expected to better explain protists differentiation, in the Mediterranean Sea where several frontal structures (i.e. boundaries between distinct water masses) correspond to switches in plankton community composition (Ayata et al., 2018). In our study, the computed geographic distance was correlated with both circulation and environmental distance and could then integrate both pieces of information. As a consequence, isolation-by-distance and isolation-by-environment may indirectly contribute together to the genomic patterns observed for *P. cordata*. In addition, for the three species studied, the statistical models did not explain most of the genomic differentiation, suggesting that other parameters should also be tested in order to improve the prediction of genomic differentiation within each of

our protist species. In particular, these parameters include historic factors, such as population size, and biotic factors, such as competition between *B. prasinos* and other Mamiellales (as questioned by Leconte et al., 2020) or free-living versus symbiont states for *P. cordata* (Uwizeye et al., 2021).

CONCLUSION

Metagenomic data represent an opportunity to improve knowledge on genomic differentiation of marine plankton, in particular for protists, which play a crucial role in ecosystem functioning but remain poorly investigated mainly due to difficulties to maintain them in culture. In this study, the genomic differentiation of three protists species with contrasted life history traits was characterized in the Mediterranean Sea based on both metagenomic samples and reference assemblies. Although relatively weak genomic differentiation was observed, we were able to identify distinct drivers explaining the observed patterns. Our results suggest that, at the scale of the Mediterranean Sea, *B. prasinos* differentiation was constrained by isolation-by-environment process, whereas *P. cordata* differentiation was constrained by isolation-by-distance process. These identified processes cannot be extrapolated to the global ocean or to another basin. Although several limitations still remain for the use of metagenomic data for population genomics studies, for example, the need of a reference assembly close to wild population, our results describe a promising approach for future studies targeting uncultured but abundant and ecologically important species.

ACKNOWLEDGEMENTS

The authors thank J. O. Irisson for help in calculating geographic distances and L. Berline for sharing oceanographic distances. The authors are also thankful to P. Debeljak for proofreading the manuscript and to F. Riquet for interesting discussions about population genetics. The authors want to thank all people involved in the Tara Oceans project for making data publicly available. This work was funded mainly by our salaries as French state employees and therefore by French taxpayers. Funding for this research was provided by the ModelOmics project of the Émergence program of Sorbonne Université, and partly supported by the project MEGALODOM, part of the MASTODON program from the MITI, CNRS France. Ophélie Da Da Silva was supported by a PhD grant from the French Ministry of Higher Education, Research and Innovation for 3 years. Enrico Ser-Giacomi is very grateful for support from the Simons Foundation: the Simons Collaboration on Computational BIOgeochemical modelling of Marine EcosystemS (CBIOMES #549931). Sakina-Dorothée Ayata acknowledges the CNRS for her two sabbatical years as visiting researcher at ISYEB in 2018–2020 and the Institut des Sciences du Calcul et des Données (ISCD) of Sorbonne Université for its supports through the "FORMAL - From ObseRving to Modeling oceAn Life" project-team (2020–2022). Lucie Bittner acknowledges the Institut Universitaire de France for her 5-year nomination as Junior Member (2020–2025).

CONFLICT OF INTEREST

The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

Pipeline, analysis scripts and contextual data are available on GitHub https://github.com/opheliedasilva/popmetag. Genomic data are available on Zenodo https://zenodo.org/record/6434681#.Yv6ngXZByiM.

ORCID

Ophélie Da Silva https://orcid.org/0000-0003-0240-2954
Sakina-Dorothée Ayata https://orcid.org/0000-0003-3226-9779
Enrico Ser-Giacomi https://orcid.org/0000-0002-2994-9514
Jade Leconte https://orcid.org/0000-0003-3132-0114
Eric Pelletier https://orcid.org/0000-0003-4228-1712
Cécile Fauvelot https://orcid.org/0000-0003-0806-1222
Mohammed-Amin Madoui https://orcid.org/0000-0003-4809-2971
Lionel Guidi https://orcid.org/0000-0002-6669-5744
Fabien Lombard https://orcid.org/0000-0002-8626-8782
Lucie Bittner https://orcid.org/0000-0001-8291-7063

REFERENCES

Alberti, A., Julie, P., Stefan, E., Labadie, K., Romac, S., Ferrera, I. et al. (2017) Viral to metazoan marine plankton nucleotide sequences from the Tara oceans expedition. *Scientific Data*, 4(1), 1–20.

Andersen, R.A., Saunders, G.W., Paskind, M.P. & Sexton, J.P. (1993) Ultrastructure and 18s rRNA gene sequence for *Pelagomonas calceolata* gen. et sp. nov. and the description of a new algal class, the Pelagophyceae classis nov. *Journal of Phycology*, 29(5), 701–715.

Ayata, S.D., Irisson, J.O., Aubert, A., Berline, L., Dutay, J.C., Mayot, N. et al. (2018) Regionalisation of the Mediterranean basin, a MERMEX synthesis. *Progress in Oceanography*, 163, 7–20.

Becker, R.A., Wilks, A.R., Brownrigg, R., Minka, T.P. & Deckmyn, A. (2018) Maps: draw geographical maps. R Package Version 3 (0).

Berline, L., Rammou, A.M., Doglioli, A., Molcard, A. & Petrenko, A. (2014) A connectivity-based eco-regionalization method of the Mediterranean Sea. *PLoS One*, 9(11), e111978.

Biard, T., Stemmann, L., Picheral, M., Mayot, N., Vandromme, P., Hauss, H. et al. (2016) In situ imaging reveals the biomass of Giant protists in the Global Ocean. *Nature*, 532(7600), 504–507.

Bittner, L., Gobet, A., Audic, S., Romac, S., Egge, E.S., Santini, S. et al. (2013) Diversity patterns of uncultured haptophytes unravelled by pyrosequencing in Naples Bay. *Molecular Ecology*, 22(1), 87–101.

Blanc-Mathieu, R., Krasovec, M., Hebrard, M., Yau, S., Desgranges, E., Martin, J. et al. (2017) Population genomics of Picophytoplankton unveils novel chromosome Hypervariability. *Science Advances*, 3(7), e1700239.

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120.

Caron, D.A., Alexander, H., Allen, A.E., Archibald, J.M., Armbrust, E. V., Bachy, C. et al. (2017) Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews. Microbiology*, 15(1), 6–20.

Caron, D.A., Countway, P.D., Jones, A.C., Kim, D.Y. & Schnetzer, A. (2012) Marine protistan diversity. *Annual Review of Marine Science*, 4, 467–493.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R. et al. (2018) A Global Ocean atlas of eukaryotic genes. *Nature Communications*, 9(1), 1–13.

Casabianca, S., Penna, A., Pecchioli, E., Jordi, A., Basterretxea, G. & Vernesi, C. (2012) Population genetic structure and connectivity of the harmful dinoflagellate *Alexandrium minutum* in the Mediterranean Sea. *Proceedings of the Royal Society B: Biological Sciences*, 279(1726), 129–138.

Casteleyn, G., Adams, N.G., Vanormelingen, P., Debeer, A.-E., Sabbe, K. & Vyverman, W. (2009) Natural hybrids in the marine diatom *Pseudo-nitzschia pungens* (Bacillariophyceae): genetic and morphological evidence. *Protist*, 160(2), 343–354.

Casteleyn, G., Leliaert, F., Backeljau, T., Debeer, A.-E., Kotaki, Y., Rhodes, L. et al. (2010) Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings National Academy of Sciences. United States of America*, 107(29), 12952–12957.

Cermeño, P. & Falkowski, P.G. (2009) Controls on diatom biogeography in the ocean. *Science*, 325(5947), 1539–1541.

Dalongeville, A., Andrello, M., Mouillot, D., Lobreaux, S., Fortin, M.-J., Lasram, F. et al. (2018) Geographic isolation and larval dispersal shape seascape genetic patterns differently according to spatial scale. *Evolutionary Applications*, 11(8), 1437–1447.

Davide, A. & Donati, C. (2017) Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nature Communications*, 8(1), 1–14.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R. et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 1–12.

Decelle, J., Probert, I., Bittner, L., Desdevises, Y., Colin, S., de Vargas, C. et al. (2012) An original mode of symbiosis in open ocean plankton. *Proc Natl Acad Sci U S A*, 109(44), 18000–18005.

Del Campo, J., Sieracki, M.E., Molestina, R., Keeling, P., Massana, R. & Ruiz-Trillo, I. (2014) The others: our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution*, 29(5), 252–259.

Delmont, T.O., Gaia, M., Hinsinger, D.D., Frémont, P., Vanni, C., Fernandez-Guerra, A. et al. (2022) Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, 2(5), 100123.

Delmont, T.O., Kiefl, E., Kilinc, O., Esen, O.C., Uysal, I., Rappe, M.S. et al. (2019) Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global Sar11 subclade. *eLife*, 8, e46497.

Demura, M., Nakayama, T., Kasai, F. & Kawachi, M. (2014) Genetic structure of Japanese *Chattonella marina* (Raphidophyceae) populations revealed using microsatellite markers. *Phycological Research*, 62(2), 102–108.

Dolan, J.R. (2005) An introduction to the biogeography of aquatic microbes. *Aquatic Microbial Ecology*, 41(1), 39–48.

Dupont, C.L., McCrow, J.P., Valas, R., Moustafa, A., Walworth, N., Goodenough, U. et al. (2015) Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *The ISME Journal*, 9(5), 1076–1092.

Eikrem, W. & Throndsen, J. (1990) The ultrastructure of Bathycoccus gen. nov. and B. Prasinos sp. nov., a non-motile Picoplanktonic alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic. *Phycologia*, 29(3), 344–350.

Evans, K.M., Kühn, S.F. & Hayes, P.K. (2005) High levels of genetic diversity and low levels of genetic differentiation in North Sea *Pseudo-nitzchia Pungens* (Bacillariophyceae) populations. *Journal of Phycology*, 41(3), 506–514.

Faure, E., Ayata, S.D. & Bittner, L. (2021) Towards omics-based predictions of planktonic functional composition from environmental data. *Nature Communications*, 12(1), 1–15.

Fichaut, M., Garcia, M.J., Giorgetti, A., Iona, A., Kuznetsov, A., Rixen, M. et al. (2003) MEDAR/MEDATLAS 2002: a Mediterranean and Black Sea database for operational oceanography. *Elsevier Oceanography Series*, 69, 645–648.

Finlay, B.J. (2002) Global dispersal of free-living microbial eukaryote species. *Science*, 296(5570), 1061–1063.

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S. et al. (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biology*, 13(1), 1–16.

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.

Garnier, S. (2018) viridis: default color maps from 'matplotlib'. R package version 0.5.1 Available at: https://CRAN.R-project.org/package=viridis

Gasol, J.M. & Kirchman, D.L. (Eds.). (2018) *Microbial ecology of the ocean*. Hoboken, NJ: Wiley, pp. 1–46.

Gibbons, S.M., Caporaso, J.G., Pirrung, M., Field, D., Knight, R. & Gilbert, J.A. (2013) Evidence for a persistent microbial seed bank throughout the global ocean. *Proceedings of the National Academy of Sciences of the United States of America*, 110(12), 4651–4655.

Godhe, A., Egardt, J., Kleinhans, D., Sundqvist, L., Hordoir, R. & Jonsson, P.R. (2013) Seascape analysis reveals regional gene flow patterns among populations of a marine planktonic diatom. *Proceedings of the Royal Society B: Biological Sciences*, 280(1773), 20131599.

Godhe, A., Sjöqvist, C., Sildever, S., Sefbom, J., Harardóttir, S., Bertos-Fortis, M. et al. (2016) Physical barriers and environmental gradients cause spatial and temporal genetic differentiation of an extensive algal bloom. *Journal of Biogeography*, 43(6), 1130–1142.

Hartl, D.L., & Clark, A.G. (1997) *Principles of Population Genetics*, 3rd edition. Sunderland, MA: Sinauer Associates.

Hellweger, F.L., van Sebille, E. & Fredrick, N.D. (2014) Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science*, 345(6202), 1346–1349.

Johnson, L.K., Alexander, H. & Brown, C.T. (2019) Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*, 8(4), giy158.

Kassambara, A. (2020) Ggpubr: 'Ggplot2' based publication ready plots. Available at: https://CRAN.R-project.org/package=ggpubr.

Kawecki, T.J. & Holt, R.D. (2002) Evolutionary consequences of asymmetric dispersal rates. *The American Naturalist*, 160(3), 333–347.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A. et al. (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.

Keeling, P.J. & Del Campo, J. (2017) Marine protists are not just big bacteria. *Current Biology*, 27(11), R541–R549.

Kulk, G., de Vries, P., van de Poll, W.H., Visser, R.J.W. & Buma, A.G. J. (2012) Temperature-dependent growth and photophysiology of prokaryotic and eukaryotic oceanic picophytoplankton. *Marine Ecology Progress Series*, 466, 43–55.

Lambert, S., Tragin, M., Lozano, J.C., Ghiglione, J.F., Vaulot, D., Bouget, F.Y. et al. (2019) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *The ISME Journal*, 13(2), 388–401.

Laso-Jadart, R., O'Malley, M., Sykulski, A., Ambroise, C. & Madoui, M.-A. (2021) How marine currents and environment shape plankton genomic differentiation: a mosaic view from Tara oceans metagenomic data. *Biorxiv*, 2021.04.29.441957.

Laso-Jadart, R., Sugier, K., Petit, E., Labadie, K., Peterlongo, P., Ambroise, C. et al. (2020) Investigating population-scale allelic differential expression in wild populations of *Oithona similis* (Cyclopoida, Claus, 1866). *Ecology and Evolution*, 10(16), 8894–8905.

Latorre, F., Deutschmann, I.M., Labarre, A., Obiol, A., Krabberød, A. K., Pelletier, E. et al. (2021) Niche adaptation promoted the evolutionary diversification of Tiny ocean predators. *Proceedings of the National Academy of Sciences of the United States of America*, 118(25), e2020955118.

Lê, S., Josse, J. & Husson, F. (2008) FactoMineR: an r package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18.

Lebret, K., Kritzberg, E.S., Figueroa, R. & Rengefors, K. (2012) Genetic diversity within and genetic differentiation between blooms of a microalgal species. *Environmental Microbiology*, 14(9), 2395–2404.

Leconte, J., Benites, L.F., Vannier, T., Wincker, P., Piganeau, G. & Jaillon, O. (2020) Genome resolved biogeography of Mamiellales. *Genes*, 11(1), 66.

Legendre, P. & Legendre, L. (2012) *Numerical ecology*, 3rd edition. Amsterdam: Elsevier.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* Preprint arXiv:1303.3997.

Logares, R. (2011) Population genetics: the next stop for microbial ecologists? *Open Life Sciences*, 6(6), 887–892.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R. et al. (2014) Patterns of rare and abundant marine microbial eukaryotes. *Current Biology*, 24(8), 813–821.

Lowe, C.D., Martin, L.E., Montagnes, D.J.S. & Watts, P.C. (2012) A legacy of contrasting spatial genetic structure on either side of the Atlantic–Mediterranean transition zone in a marine protist. *Proceedings of the National Academy of Sciences of the United States of America*, 109(51), 20998–21003.

Lowe, C.D., Montagnes, D.J.S., Martin, L.E. & Watts, P.C. (2010) Patterns of genetic diversity in the marine heterotrophic flagellate *Oxyrrhis marina* (Alveolata: Dinophyceae). *Protist*, 161(2), 212–221.

Lumley, T. & Miller, A. (2013) Package 'leaps': regression subset selection. Available at: http://CRAN.R-Project.Org/Package=Leaps [Accessed 18th March 2018].

Madoui, M.-A., Poulain, J., Sugier, K., Wessner, M., Noel, B., Berline, L. et al. (2017) New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Molecular Ecology*, 26(17), 4467–4482.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J. et al. (2016) Insights into global diatom distribution and diversity in the World's ocean. *Proceedings of the National Academy of Sciences of the United States of America*, 113(11), E1516–E1525.

Mena, C., Reglero, P., Hidalgo, M., Sintes, E., Santiago, R., Martín, M. et al. (2019) Phytoplankton community structure is driven by stratification in the oligotrophic Mediterranean Sea. *Frontiers in Microbiology*, 10, 1698.

Moon-van der Staay, S.Y., De Wachter, R. & Vaulot, D. (2001) Oceanic 18s rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, 409(6820), 607–610.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N. et al. (2012) Gene functionalities and genome structure in *Bathycoccus Prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biology*, 13(8), 1–16.

Nagai, S., Lian, C., Yamaguchi, S., Hamaguchi, M., Matsuyama, Y., Itakura, S. et al. (2007) Microsatellite markers reveal population genetic structure of the toxic dinoflagellate *Alexandrium Tamarense* (Dinophyceae) in Japanese coastal waters. *Journal of Phycology*, 43(1), 43–54.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G. et al. (2015) Open science resources for the discovery and analysis of Tara oceans data. *Scientific Data*, 2(1), 1–16.

R Core Team. (2020) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at:. https://www.R-project.org/

Read, B.A., Kegel, J., Klute, M.J., Kuo, A., Lefebvre, S.C., Maumus, F. et al. (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature*, 499(7457), 209–213.

Riginos, C., Crandall, E.D., Liggins, L., Bongaerts, P. & Treml, E.A. (2016) Navigating the currents of seascape genomics: how spatial analyses can augment population genomic studies. *Current Zoology*, 62(6), 581–601.

Schmieder, R. & Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6), 863–864.

Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M. et al. (2018) Single-cell genomics of multiple uncultured Stramenopiles reveals underestimated functional diversity across oceans. *Nature Communications*, 9(1), 1–10.

Selkoe, K.A., Aloia, C.C.D., Crandall, E.D., Iacchei, M., Liggins, L., Puritz, J.B. et al. (2016) A decade of seascape genetics: contributions to basic and applied marine connectivity. *Marine Ecology Progress Series*, 554, 1–19.

Ser-Giacomi, E., Rossi, V., López, C. & Hernández-García, E. (2015) Flow networks: a characterization of geophysical fluid transport. *Chaos*, 25(3), 036404.

Sexton, J.P., Hangartner, S.B. & Hoffmann, A.A. (2014) Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*, 68(1), 1–15.

Shaiber, A. & Eren, A.M. (2019) Composite metagenome-assembled genomes reduce the quality of public genome repositories. *MBio*, 10(3), e00725–e00719.

Sibbald, S.J. & Archibald, J.M. (2017) More protist genomes needed. *Nature Ecology and Evolution*, 1(5), 1–3.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A. & Ponting, C.P. (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews. Genetics*, 15(2), 121–132.

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.

Slowikowski, K. (2020) Ggrepel: automatically position non-overlapping text labels with 'Ggplot2'. Available at: https://CRAN.R-project.org/package=ggrepel.

Spalding, M.D., Fox, H.E., Allen, G.R., Davidson, N., Ferdaña, Z.A., Finlayson, M.A.X. et al. (2007) Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. *Bioscience*, 57(7), 573–583.

Uwizeye, C., Brisbin, M.M., Gallet, B., Chevalier, F., Lekieffre, C., Schieber, N.L. et al. (2021) Cytoklepty in the plankton: a host strategy to optimize the bioenergetic machinery of endosymbiotic algae. *Proceedings of the National Academy of Sciences of the United States of America*, 118(27), e2025252118.

van Etten, J. (2017) R package Gdistance: distances and routes on geographical grids. *Journal of Statistical Software*, 76(13), 21.

Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J.-M. et al. (2016) Survey of the green Picoalga *Bathycoccus* genomes in the Global Ocean. *Scientific Reports*, 6(1), 1–11.

Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T.O., Annamalé, A., Wincker, P. et al. (2020) Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Research*, 30(4), 647–659.

Wang, I.J. & Bradburd, G.S. (2014) Isolation by environment. *Molecular Ecology*, 23(23), 5649–5662.

Watts, P.C., Lundholm, N., Ribeiro, S. & Ellegaard, M. (2013) A century-long genetic record reveals that protist effective population sizes are comparable to those of macroscopic species. *Biology Letters*, 9(6), 20130849.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R. et al. (2019) Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A. et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1–9.

Worden, A.Z., Janouskovec, J., McRose, D., Engman, A., Welsh, R. M., Malfatti, S. et al. (2012) Global distribution of a wild alga revealed by targeted metagenomics. *Current Biology*, 22(17), R675–R677.

Wright, S. (1943) Isolation by distance. *Genetics*, 28(2), 114–138.

Zingone, A., Chrétiennot-Dinet, M.-J., Lange, M. & Medlin, L. (1999) Morphological and genetic characterization of *Phaeocystis cordata* and *p. Jahnii* (Prymnesiophyceae), two new species from the Mediterranean Sea. *Journal of Phycology*, 35(6), 1322–1337.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Da Silva, O., Ayata, S.-D., Ser-Giacomi, E., Leconte, J., Pelletier, E., Fauvelot, C. et al. (2022) Genomic differentiation of three pico-phytoplankton species in the Mediterranean Sea. *Environmental Microbiology*, 1–14. Available from: https://doi.org/10.1111/1462-2920.16171