Decoding Populations in the Ocean Microbiome

Ramiro Logares

Institute of Marine Sciences (ICM), CSIC, Barcelona, E-08003, Catalonia, Spain.

Contact

Institute of Marine Sciences (ICM), CSIC,

Passeig Marítim de la Barceloneta, 37-49

E-08003, Barcelona, Spain

Email: ramiro.logares@icm.csic.es

Article type: Review

ABSTRACT

Understanding the characteristics and structure of populations is fundamental to comprehending ecosystem processes and evolutionary adaptations. While the study of animal and plant populations has spanned a few centuries, microbial populations have been under scientific scrutiny for a considerably shorter period. In the ocean, analyzing the genetic composition of microbial populations and their adaptations to multiple niches can yield important insights into ecosystem function and the microbiome's response to global change. However, microbial populations have remained elusive to the scientific community due to the challenges associated with isolating microorganisms in the laboratory. Today, advancements in large-scale metagenomics and metatranscriptomics facilitate the investigation of populations from many uncultured microbial species directly from their habitats. The knowledge acquired thus far reveals substantial genetic diversity among various microbial species, showcasing distinct patterns of population differentiation and adaptations, and highlighting the significant role of selection in structuring populations. In the coming years, population genomics is expected to significantly increase our understanding of the architecture and functioning of the ocean microbiome, providing insights into its vulnerability or resilience in the face of ongoing global change.

Keywords: microbes, populations, ocean, metagenomics, metatranscriptomics

MAIN TEXT

Ocean microbes are key for the functioning of the Earth's system

The ocean microbiome is one of the main engines of the biosphere and, to a large extent, responsible for the conditions we live in [1]. This microbiome is populated by an astronomical number of cells. Gross estimates indicate that the global ocean harbors ~10^{29} prokaryotic cells and ~10^{30} viruses [2,3], while in one milliliter of open ocean water, there are typically 10^3 protists, 10^6 prokaryotes, and 10^7 viruses [4]. Microbes account for ~70% of the biomass in the ocean, representing ~4.2 gigatons of carbon [5]. This biomass is distributed in at least 10^{10} species [6] that belong to a wide array of phylogenetic lineages, several of which have been diversifying in the ocean for eons [7]. Thus, the ocean microbiome is a large reservoir of taxonomic and functional diversity.

The ocean microbiome is crucial in global biogeochemical cycles [1,8]. In the sunlit ocean, the tiniest microbes, the picoplankton, are responsible for an important fraction of the total atmospheric carbon and nitrogen fixation [9–11], representing ~46% of the global primary productivity [12]. Surface ocean picoplankton plays a fundamental role in processing organic matter by recycling nutrients and carbon to support additional production and channeling organic carbon to upper trophic levels in food webs [11,13,14].

Two key components of the ocean microbiome, prokaryotes (bacteria and archaea) and unicellular eukaryotes or protists (including marine fungi), have fundamental differences in cellular structure, feeding habits, metabolic diversity, growth rates, and behavior [15]. Prokaryotic metabolisms are diverse and have major roles in global biogeochemical cycles [1,8]. In contrast, protists' metabolisms are less diverse, but instead, they show major innovations in morphology and behavior [15]. A

substantial fraction of the ocean microbiome biomass seems to comprise protists (and fungi) [5], including many heterotrophic groups that transfer carbon from prokaryotes or other protists to upper trophic levels.

What is the total diversity of the ocean microbiome?

This is a recurrent question in marine microbial ecology that has been addressed in multiple works [6,16–21] and that, so far, does not have a definitive answer. Current estimates of the total prokaryotic diversity on the planet vary significantly, with some differing by orders of magnitude [6,19–21]. Nevertheless, over the past twenty years, we have made significant progress in understanding and delimiting the diversity of the vast array of microorganisms in the ocean. This is, in part, a consequence of the omics revolution that allowed retrieving microbes directly from the environment. Pioneering surveys ~20 years ago pointed to a large diversity of microbial genes and taxa in the ocean [22]. Subsequent large-scale oceanographic campaigns, such as *Malaspina* [23], TARA Oceans [24], Bio-GO-SHIP [25], and GEOTRACES cruises [26], significantly expanded our comprehension of the magnitude of the ocean's microbiome diversity. These campaigns indicated ~50,000 - 100,000 protists and ~10,000 - 35,000 bacterial "species" or taxonomic units [16,27,28] in the open ocean plankton using High Throughput DNA Sequencing (HTS). From the metabolic-function perspective, TARA Oceans, based on sequencing microbial genome fragments (hereafter metagenomics), has cataloged ~47 million predominantly prokaryotic genes [29] and ~116 million eukaryotic genes [30] at the global-ocean plankton scale. Similarly, the *Malaspina* consortium reported ~4 million predominantly prokaryotic genes from the deep ocean plankton [31].

100 The previous estimates show substantial variability, but over the next few years,
101 they will likely improve, providing us with more accurate estimates of the diversity of
102 the ocean microbiome. Yet, these estimates are bound to the evolutionary divergence
103 captured by the rRNA-gene or functional genes, which may miss fine-grained diversity
104 or could introduce biases. For example, the rRNA gene is a slow-evolving marker that
105 may not capture differences between microbial species or populations. Similarly,
106 different microbial species may share large identical regions of their genomes [32],
107 and if we focus on those areas, species will be indistinguishable.

108 Microdiversity refers to small-scale genetic variations (e.g., Single Nucleotide
109 Variants or SNVs) within a microbial species or population or among closely related
110 species and can be crucial for comprehending ecosystem function and the
111 vulnerability or resilience of communities to global change [33,34], contemporary
112 evolution [35], and ecological interactions [36]. Besides SNVs, horizontal gene transfer
113 and homologous recombination can also contribute to microdiversity and confer new
114 traits to different members of the same species [32,37].

115 Crossing the boundaries between microbial species and populations, and
116 comprehending the intra-species vs. the inter-species genetic variation is a current
117 challenge for microbial ecologists. There has been a tendency in microbial ecology to
118 specialize either in population-level (e.g., population genetics or genomics) or
119 community-level studies (e.g., community ecology). One of the main reasons is that
120 many population-level studies have been performed using cultures, while researchers
121 focusing on community ecology normally work with uncultured species [38]. Yet, we
122 will need to get used to moving across the species and population boundaries when
123 investigating the ocean microbiome, that is, between populations and communities, to

increase our understanding of its structure, the ecological interactions it contains, and its links with ecosystem processes.

Microbial species and populations

How do we define a microbial species? This is perhaps among the oldest questions in microbial ecology and remains so far partially answered. It has generated much debate, and abundant literature exists elsewhere [1,39,40]. Therefore, I will not address this question here. For this piece, I will consider microbial species as coherent genetic and ecological units composed of individuals that are phenotypically and ecologically more similar to themselves than to other species [32]. Speciation and diversification seem to require both divergent selection and gene flow barriers to occur [38]. Selective diversification and speciation would align with the *Ecological Species Concept*, where natural selection drives the process of divergence towards different niches [41], being the mechanism of speciation envisioned by Darwin. In turn, the *Biological Species Concept* [42] emphasizes the restrictions on gene flow as the main mechanism of diversification and speciation. Even though both concepts emerged from the study of animals and plants, and their validity in understanding microbial diversification is still under debate (especially due to Horizontal Gene Transfer or HGT), it is likely that both processes have a role in the adaptive diversification of microorganisms. Adaptive diversification is of interest as it is expected to generate microdiversity that reflects niche adaptations that may not be detected in regular surveys of the ocean microbiome using rRNA-genes or functional gene markers.

In animals and plants, it is expected that most genes flowing in one species do not affect those in another. Yet, in prokaryotes and, to some extent, microbial eukaryotes, the horizontal exchange of genomic information makes it difficult to make

clear separations between eco-genetic units. The transferred DNA can give new capabilities to the cells that receive it (for example, antibiotic resistance), and change its niche dimension, leading to a new differentially adapted population featuring a specific trait. Overall, despite the potentially leaky boundaries between eco-genetic units due to HGT, analyses of environmental isolates and metagenomes indicated that genotypic clusters of closely related organisms display cohesive responses to environmental heterogeneity, distinguishing them from other coexisting clusters [32].

The interplay of selection (*s*) and recombination (*r*) (i.e., horizontal exchange of DNA between cells) has been proposed as a key mechanism to explain the spread of new adaptive gene variants among and within eco-genetic units [32]. Here, I will mention two possible scenarios deriving from models. In the first, recombination within populations is low, and selection is high for a given gene or locus. Then, individuals with the advantageous trait (gene) will increase in abundance due to clonal expansion taking over the entire population leading to a *genome-wide selective sweep* (GWSS) [43]. This process purges genetic variation from populations, and different eco-genetic clusters may form after ecologically different populations experience multiple genome-wide sweeps [32,43]. Alternatively, in the second scenario, high recombination rates compared to selection are expected to promote the exchange of selectively advantageous genes among different population members without purging diversity, leading to *gene-specific selective sweeps* (GSSS). In this second case, eco-genetic clusters may take longer to form [32]. In the former process (GWSS), an adaptive gene or locus will tend to appear in a specific selective background, while in the latter (GSSS), the selective gene is expected to be present in multiple backgrounds. Even though the previous models may be oversimplified, they generate hypotheses to explain some observed characteristics of microbial populations in the ocean. For

example, the amount of genetic variation in populations and its distribution. In addition, these models may help predict the reactions of microbial populations to global change in the ocean by, for example, pointing to changes in the prevalence of GWSS or GSSS.

One challenge when investigating microbial populations is determining what organisms belong to the same species. One operative approach is to use genome similarity thresholds (e.g., the 95% threshold in the Average Nucleotide Identity [44,45]) to delineate species. This is particularly useful in studies without multiple genomes from cultures to compare, as in marine metagenomic studies. Although these thresholds are practical and popular, they require an *a priori* decision on the cut-off level to delineate different Operational Taxonomic Units (OTUs). The chosen threshold may or may not correspond with natural eco-genetic clusters.

An alternative to using arbitrary thresholds is to search for natural discontinuities in genomic diversity that could be linked to eco-genetic clusters that may represent populations or species. This approach has been recently referred to as *reverse ecology* [46,47]. One example of its implementation is the methodology that uses recent gene flow to delineate eco-genetic units, which could be linked to populations or species [46,47]. Here, gene flow discontinuities are identified and used to delineate species ("gene flow units") that can be subdivided into populations ("adaptively optimized gene flow clusters") without using any prior environmental knowledge [47]. The rationale is that recent gene flow will leave a higher number of identical regions in genomes exchanging genes horizontally compared to what would be expected if mutations had accumulated without gene transfer [47] **(Figure 1)**. The reason is that horizontally exchanged DNA would not have had enough time to accumulate mutations compared to other regions shared by descent or vertically. Then, pairwise measurements of recent gene flow among genomes can be used to

construct gene-flow networks to identify gene flow units (species) and gene flow clusters (populations) within them. A test of this approach produced genome clusters corresponding to previously identified populations of *Vibrio*, *Sulfolobus,* and *Prochlorococcus* [47]. Furthermore, results indicated strong discontinuities in the gene flow between species (gene flow units), aligning with the classic Biological Species Concept [42].

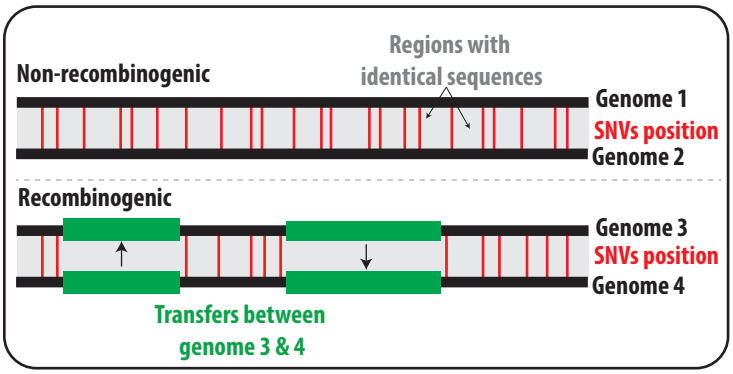


Figure 1. Microbial genomes that recombine (recombinogenic) and, therefore, belong to the same population or species would share longer identical regions than non-recombinogenic counterparts. Modified from Arevalo et al. [47]

From population genetics to population genomics

Population genetics investigate the evolutionary forces that generate, assort, and remove variation within species using specific marker genes or genomic areas. *Population genomics* is basically population genetics but using entire genomes [38]. While population genetics is an established field, population genomics is still an emerging field in microbiology, which has been boosted by decreasing DNA sequencing costs. Population genomics has a huge potential for a deeper understanding of the ocean microbiome, as it can reveal the fine-grained adaptive variation among populations and the genotypes that produce disease or dysbiosis [38,48].

The main forces determining the genetic composition of populations are *mutation*, *selection*, *gene flow*, and *genetic drift. Mutation* is the emergence of new and random gene variants and is the ultimate source of diversity. *Selection* changes allele frequencies due to their fitness impact on the phenotype, while *gene flow* is related to the exchange of genes between individuals. Lastly, *genetic drift* refers to the random fluctuations in allele frequencies from one generation to the next due to the stochastic sampling of individuals contributing offspring to the next generation [49].

Despite microbial population genetics and genomics being growing fields [38,50,51], our understanding of *mutation*, *selection*, *gene flow,* and *genetic drift* is still predominantly based on the study of animals and plants. The previous is especially true for environmental microbes. Yet, microbes typically differ from multicellular organisms in at least three fundamental aspects: *dispersal*, *reproductive rates,* and *population size* [52,53]. Even though the dispersal rate of most microbes is still unknown, indirect evidence points to high dispersal rates [52,54] that could be substantially higher than in multicellular organisms. Nevertheless, while it has been argued that organisms with <1mm of body size have virtually no barriers to dispersal [55], multiple studies during the last two decades point to dispersal limitation in microbes [28,52,54,56]. Furthermore, the reproductive rates of multicellular organisms tend to be lower than those of microbes. For example, generation times in small mammals can be in the order of months, while in some bacteria, it can be in the order of minutes/hours. Faster generation time implies that mutation, adaptation, and divergence can occur faster in microbes than in multicellular organisms.

Census vs. effective population size

Census population size (N), together with effective population size (N_e), are key parameters in population genetics that can affect population adaptation, drift, and dispersal. Census population size refers to the total number of individuals or cells and can affect random dispersal as more cells increase the chances of arriving at new locations. In turn, the effective population size N_e represents the number of individuals in a theoretical population that would experience the same amount of genetic drift as the population under consideration. N_e plays a pivotal role in population genetics. It influences the magnitude of genetic drift, the extent of genetic variability within a population, and the balance between the efficacy of selection and the random effects of drift [49]. Specifically, a population's neutral genetic diversity, which refers to genetic variations without fitness effects, is determined by the product of the effective population size N_e and the mutation rate. Furthermore, N_e is intrinsically tied to the efficacy of selection. It dictates whether a beneficial mutation proliferates or a deleterious one is purged, with the outcome governed by the product of N_e and the intensity of selection [49]. Small N_e can increase genetic drift, which can lead to reduced genetic diversity over time, increase the likelihood of the fixation of deleterious alleles, and increase the chances of losing advantageous alleles [57].

While N can be huge in microbes, N_e is usually smaller due to the variance in reproductive success and potential selective sweeps. Lynch and colleagues calculated N_e ~10^5 for vertebrates, ~10^6 for invertebrates and land plants, ~10^7 for unicellular eukaryotes, including fungi, and 10^8 for free-living prokaryotes [58]. These estimates imply that drift is about three orders of magnitude higher in large multicellular eukaryotes than in prokaryotes and that the effective population sizes are far below the census population sizes. Furthermore, the previous estimates indicate that

276 selection will be more efficient in large microbial populations than in animals and plants
277 [49]. This has been proposed as the basic tenet for the increasing number of genes
278 (due to retention of duplicates), introns, and mobile genetic elements in larger
279 eukaryotes, compared to prokaryotes [59]. The rationale is that the increase in
280 organism size would have led to smaller N_e and, therefore, a higher drift that allowed
281 the proliferation of the mentioned elements in eukaryotes. This hypothesis was initially
282 used to explain the genome streamlining (that is, the process by which non-essential
283 DNA is eliminated from genomes) of specific microbial genomes with crucial
284 importance in the ocean ecosystem, such as *Prochlorococcus*. Yet, later studies have
285 shown that other factors, such as niche complexity, must be considered to explain
286 streamlining [60].

287 Substantial variability in effective population size has been reported for
288 prokaryotes, ranging between 10^6 (host-associated) and 10^{10} (free-living), typically
289 being > 10^8 [61,62]. Similarly, the N_e of microbial eukaryotes has been found to range
290 between 10^6 (host-associated) and 10^8 (free-living) [62]. Despite the existing estimates
291 of N_e, this key variable remains unknown for most marine microbial species [63],
292 limiting our capability to understand how they may adapt to a changing ocean.
293 Measuring the N_e of marine microbes could also reveal unexpected results. For
294 example, the marine *Prochlorococcus* is one of the planet's most prolific
295 photosynthetic organisms, playing a pivotal role in global biogeochemical cycles.
296 *Prochlorococcus* features an average global abundance of 3×10^{27} cells annually and
297 contributes to a net primary production of 4 gigatons of carbon each year (~8% of the
298 ocean net primary production) [64]. The small genome size (~1.6 to ~2.7 Mbp [65]) of
299 *Prochlorococcus* suggested a large N_e, yet a recent study estimated the N_e of
300 *Prochlorococcus* to be ~1.7 x 10^7, being surprisingly smaller than that of other free-

living bacteria and suggesting that drift could be the key driver of evolution in this lineage [66]. This finding also raises questions about *Prochlorococcus*'s adaptability to global change. Other marine bacteria with massive census population sizes could also have smaller N_e than expected. For SAR11, which has a census population size of $2.4x10^{28}$ [60], reports indicate an effective population size that is smaller than that of *Roseobacter* [63]. Considering the crucial role of N_e in discerning the adaptive potential of marine microbial populations to climate change, it is imperative to determine this parameter, at least for those species with key roles in ocean ecosystem function.

Microbial population diversity, structure, and adaptations in the omics era

Characterizing and understanding the genomic diversity within microbial species and the genomic differences between strains, their phenotypes, and their ecological relevance is a primary challenge for microbial ecologists [67]**.** Specifically, comprehending the ecological differences between strains is highly relevant for understanding ecosystem function due to the different phenotypes and ecological roles that strains could have [68]. For example, both commensal and pathogenic strains can be found in *Escherichia coli* [69,70], *Enterococcus cecorum* [71]*,* and *Bacteroides fragilis* [72]. The study of strain-level heterogeneity can also contribute to characterizing pathogens and their ecosystemic impact [38,48]. Linking the diversity within species with environmental heterogeneity may also provide insights into short-term evolutionary processes (i.e., occurring before speciation) and the genomic differences that led to differential adaptation.

Even though our understanding of the genomic diversity and structure of environmental microbial populations and the genetic basis of strain differentiation is

limited, multiple studies reflect the fast progress of the field, fueled by the decreasing sequencing costs [67]. This is particularly evident in studies of the human microbiome [67]. Fewer studies are available for aquatic microbes. Still, a number of pioneering works pointed to high genomic diversity within microbial species and correlations between population genomic differentiation and niche adaptation. For example, populations adapted to different light intensities [73], and temperatures [74] were found among *Prochlorococcus* ecotypes. Further studies indicated that *Prochlorococcus* includes an enormous population variation, with potentially hundreds of subpopulations coexisting in small seawater samples [75]. These subpopulations displayed a substantial allelic variation in their core genome (including housekeeping and ecologically relevant genes), delineating different genomic backbones. Furthermore, each subpopulation genomic backbone was linked to distinct sets of flexible genes that may reflect different metabolic functions, thus pointing to adaptive evolution [75]. Another study [32] compared the patterns of population divergence in marine strains of *Vibrio cyclitrophicus* [76] as well as in the hot-spring archaeon *Sulfolobus islandicus* [77]. Both species displayed substantial diversity, and populations differentiated by Single Nucleotide Variants (SNVs) in specific areas of their genomes. While in *Vibrio* the SNVs were localized in genomic “islands”, in *Sulfolobus* they were spread across genomic “continents”. Genomic islands in *Vibrio* contain ecologically relevant genes, suggesting that SNVs are likely involved in ecological adaptation. Outside these islands or continents, populations were not differentiated [32].

The majority of the previous studies have used cultured isolates of microbial strains to investigate population diversity and structure. Yet, most of the microbial diversity cannot be cultured [78]. Therefore, researchers have started to use culture-

independent approaches to investigate wild microbial populations, such as Single-Cell Genomics **(Figure 2)** and Metagenomics [67,75,79,80]. A number of studies have recently started to leverage the power of Metagenome-Based Population Genomics [67,80] and the availability of large public datasets to investigate microdiversity in aquatic microbes **(Figure 3)**. These studies can be divided into two main classes: 1) those that compare metagenomic information against a collection of genomes or sequences of interest (e.g., POGENOM [81], MIDAS [82], metaSNV [83], StrainPhlAn [84], and inStrain [85]; **Figure 3**) and 2) reference-free approaches that investigate fine-grained variation among metagenomic reads (e.g., metaVaR [86]). Furthermore, and linked to the first approach, there are methods that aim at reconstructing strains or haplotypes from the metagenomic data (e.g., ConStrains [87], DESMAN [88], STRONG [89], InStrain [85], and Strain-GeMS [90]). Given the space limitations, below, I will provide a few examples of some of these approaches applied to marine microbes to convey the central message without aiming for a comprehensive review.

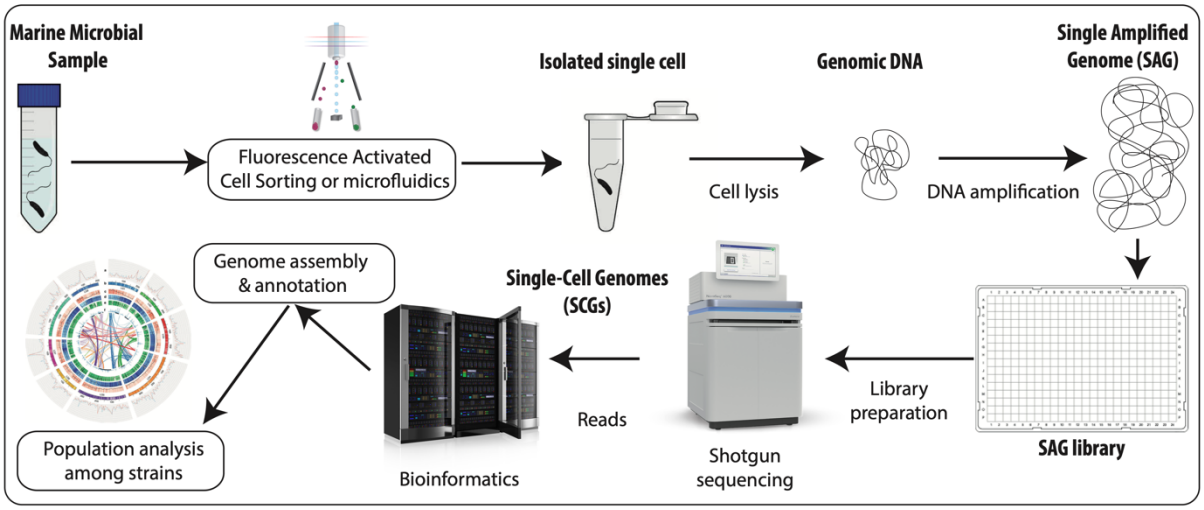


Figure 2. Single Cell Genomics [79]. In a nutshell, this approach starts with isolating single microbial cells, typically using Fluorescence Activated Cell Sorting (FACS) or microfluidics. Then, cells are lysed, and their genomic DNA is amplified, generating Single Amplified Genomes (SAGs). SAGs are subsequently shotgun sequenced, and the produced reads (DNA sequences) are assembled and annotated. Those SAGs from the same species can then be used for population genomics analyses (as in Kashtan et al. [75]). Furthermore, SAGs can be used as genomic templates in metagenome-based population genomics analyses [80] (Figure 3).

One pioneering study compared the information present in metagenomes against a compiled database of ca. 30,000 reference bacterial genomes using a tailored bioinformatics pipeline (MIDAS) [82]. This approach was used to investigate the population-level variation in 198 marine metagenomes from TARA Oceans coming from 66 stations in the global ocean [91]. Not surprisingly, it was found that, in general, the reference bacterial genomes used in MIDAS had low coverage in the ocean samples. Nevertheless, sufficient recruitment was evidenced for reference genomes of the genera *Pelagibacter*, *Alteromonas*, *Synechococcus*, and *Marinobacter* [82]. Pan-genome analyses showed a substantial variability of gene content in these species across the marine metagenomes. When all species were considered, an average of 19% of the genes differed between metagenomes [82], indicating substantial variability in gene content between strains across marine stations. Based on the variability in gene content of each bacterial species, authors found that the populations of different species were grouped by ocean region. For instance, SAR11 (*Pelagibacter*) was segregated into three distinct clusters, each aligning with a specific geographic region: the Mediterranean Sea, the South Atlantic Ocean, and the South Pacific Ocean. Each cluster encompassed samples from multiple water layers [82]. Furthermore, geographic distance decay in gene content was detected for most of the species examined. Hence, there appears to be a correlation between strain gene content and geographical distribution for several marine bacterial species.

As one of the most abundant lineages in the ocean, SAR11 [60] serves as an ideal model species for population genomics studies, facilitating the exploration of fine-grained microbial adaptations to the marine environment. SAR11 features sub-clades with specific ecological preferences and contains a large microdiversity [60,92–94]. Large amounts of microdiversity and frequent recombination [95] seem to reduce the

399 recovery of SAR11 contigs from metagenomes, even when the number of reads is
400 high, which limits the number of recovered Metagenome-Assembled Genomes
401 (MAGs) [94,96]. The low recovery of MAGs complicates population genomics
402 analyses, yet, a number of studies have found ways to leverage the large amounts of
403 SAR11 information in marine metagenomes. Haro-Moreno and colleagues
404 investigated the diversity and distribution of SAR11 using a large collection of Single-
405 Amplified Genomes (SAGs), cultures, and MAGs, together with a collection of 620
406 metagenomes [94]. A large population-level diversity was detected, indicating that this
407 is a characteristic of Pelagibacterales. Furthermore, population-level diversity was
408 conserved across a broad horizontal dimension of the ocean, pointing to a limited
409 influence of horizontal biogeography in the structure of microdiversity for the
410 investigated lineage. In turn, population-level diversity displayed marked changes
411 across the water column at single locations, indicating that the vertical dimension of
412 the ocean has a larger impact on microdiversity than the horizontal, despite their large
413 differences in geographic scale (a few kilometers vs. hundreds or thousands of
414 kilometers, respectively). This study also reports many synonymous Single Nucleotide
415 Variants (SNVs) in the investigated genomes, which aligns with a strong purifying
416 selection. Only a few genes displayed positive selection, which could be the basis of
417 strain or population adaptation [94]. Similarly, Delmont and colleagues [96] examined
418 the population variation of an abundant isolate of SAR11 in the surface global ocean
419 using metagenomics and found a large amount of variation in terms of Single Amino-
420 Acid Variants (SAAVs). More protein variants were detected in cold than in warm
421 currents, suggesting different adaptive patterns in populations. By clustering
422 metagenomes based on the SAAVs they feature (i.e., the populations that
423 metagenomes represent) revealed two main SAR11 clusters corresponding to warm

or cold large-scale ocean currents, suggesting two main niches for this SAR11 isolate [96]. At a finer scale, 6 proteotypes were identified, grouping samples with similar amino acid variants; these tended to display specific distributions in the global ocean linked to temperature, basins, and/or currents. Altogether, the correlation between SAR11 population-level diversity and environmental variables, particularly temperature, suggests that selection plays a more important role than dispersal in shaping the population structure of this key marine lineage. Another study has reported evidence of two subspecies for a SAR11 genome [83]. These subspecies had specific distributions, with one dominating in the Atlantic, Indian, and North-Pacific oceans and the other dominating in the South-Pacific Ocean. The correspondence between these subspecies with previous findings needs further investigation due to the different levels at which within-species diversity was investigated [67], as well as the likely use of different reference genomes.

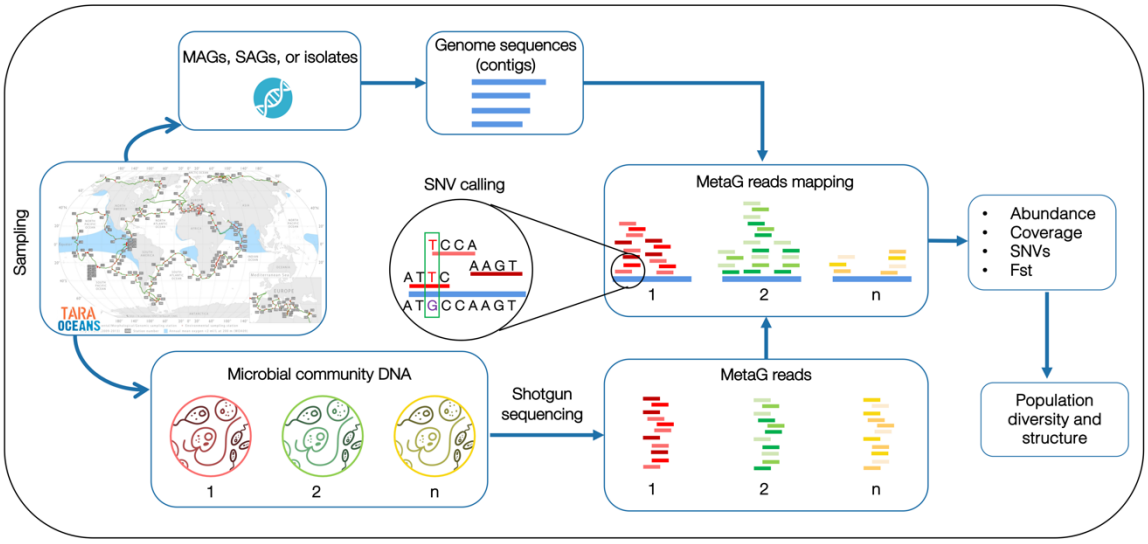


Figure 3. Metagenome-Based Population Genomics [80]**.** Metagenome-Assembled Genomes (MAGs), Single Amplified Genomes (SAGs; Figure 2), or genomes from isolates are generated after sampling or retrieved from collections. In parallel, marine metagenomes (MetaG) are produced from community DNA or retrieved from databases. Subsequently, unassembled metagenomes (reads) are mapped against MAGs, SAGs, or sequenced isolates. After mapping, the abundance and the horizontal

443 and vertical coverages of each MAG, SAG, or isolate are calculated, and Single Nucleotide Variants
444 (SNVs) are called. Based on the SNVs, population-level diversity, and structure (based on the Fst index)
445 can be assessed. The trajectory of the TARA Oceans sampling campaign is shown as an example. See
446 an application of this approach in Figure 4.
447

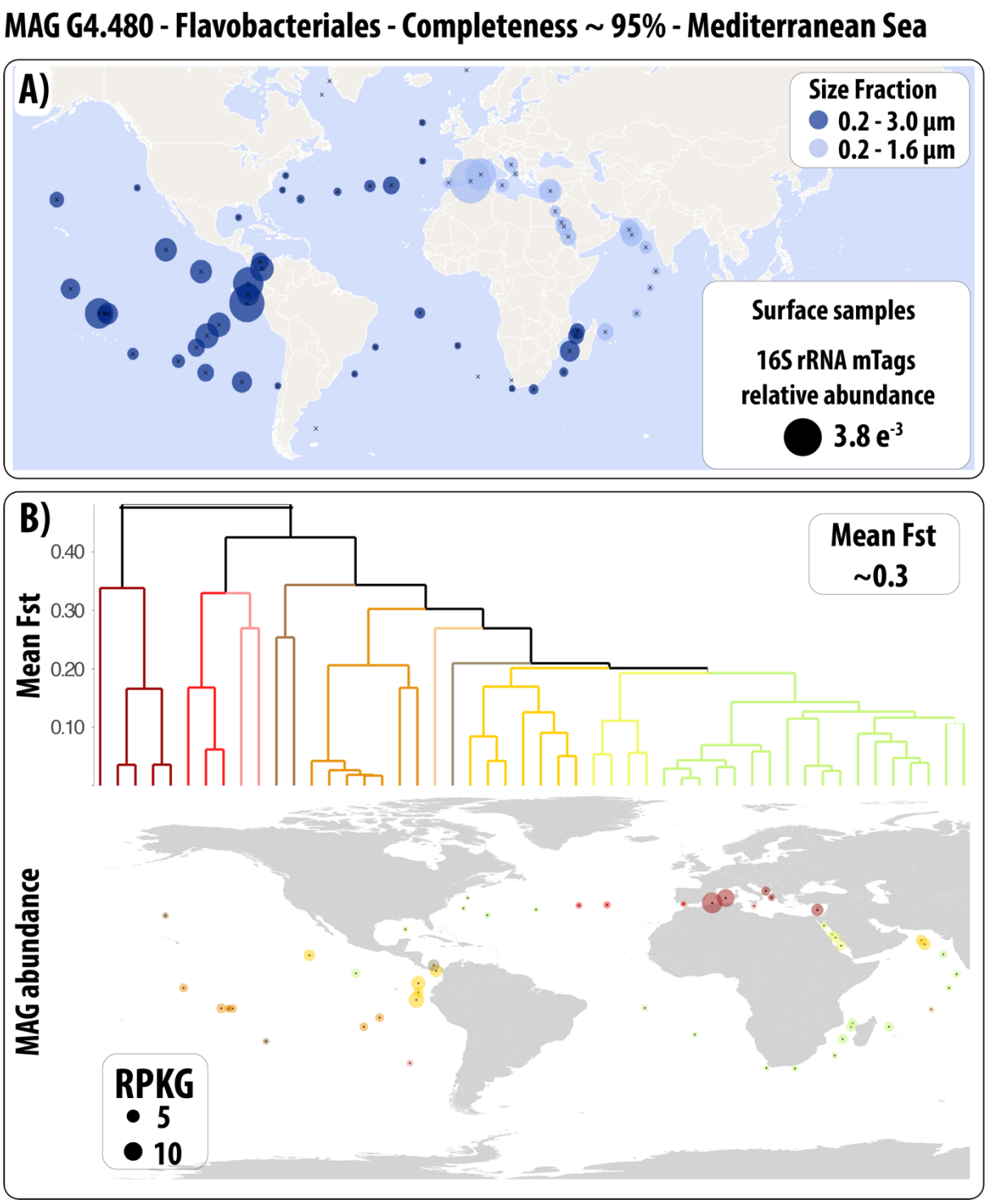
448 Population-level variation correlating with environmental heterogeneity was
449 also reported in a study of bacterioplankton in the Baltic Sea [81]. Here, Sjöqvist and
450 colleagues investigated the population-level diversity and structure of 22 MAGs that
451 were representative of genomic clusters by using metagenomes from a 1700km
452 transect and a time series. A substantial number of SNVs were detected for the 22
453 MAGs. Intra-sample mean nucleotide diversity (representing the probability that two
454 metagenomic reads covering a genomic position differ) displayed specific patterns for
455 some MAGs in the spatial dimension, while no temporal trends were observed [81].
456 Most MAGs displayed a non-random population structure across the Baltic Sea, as
457 measured by the fixation index (Fst, a measure of population differentiation). Salinity
458 and temperature emerged as the first and second spatial drivers of population
459 structure, respectively [81]. In four MAGs, evidence of isolation by distance
460 (geographic effects) was detected. Temporal temperature variation was a significant
461 population structuring driver for two MAGs (out of the four that could be analyzed).
462 Overall, population differentiation was higher across the Baltic Sea than temporally,
463 suggesting that spatial differences in salinity and temperature are a stronger driver of
464 population differentiation than seasonal variation of environmental variables.
465 Differentially adapted genes were detected in populations present at different
466 salinities, suggesting they may be the basis of population adaptation. Unlike the global
467 ocean, where temperature appears to be the central factor influencing population
468 structure [96], this study [81] identifies salinity as the primary driver in the Baltic Sea,
469 a region characterized by substantial salinity gradients.

Metagenome-based population genomics approaches have also been used to investigate marine protists. Leconte and colleagues investigated the population genomics of the picophytoplankton *Bathycoccus* RCC1105 isolated in January 2006 from the SOLA station (Banyuls-sur-Mer, France) in the Western Mediterranean Sea at 3m depth [97]. Broad population-level variation patterns were assessed using surface and deep chlorophyll maximum metagenomes from the TARA Oceans campaign corresponding to the 0.8-5 µm organismal size fraction. Of the original 162 TARA Oceans metagenomes, only 27 (ca. 17%) from diverse geographic locations and different ocean basins displayed enough coverage of the reference genome for downstream analyses [97]. Even though *Bathycoccus* has a relatively small genome (~15 Mb [98]) and displays widespread geographic distributions [99], the previous results evidence the greater difficulties of applying the metagenome-based population genomics approach to protists compared to prokaryotes [100]. The primary reason is that marine metagenomes generally encompass more prokaryotic than eukaryotic information, compounded by the inherently larger size and complexity of eukaryotic genomes. Nevertheless, when comparing the 27 metagenomes based on the SNVs they contain, it was found a clear separation between those originating from Arctic and temperate regions [97]. In addition, Arctic populations displayed a clear separation from Austral ones. A positive correlation between population and temperature differences was found [97], indicating, as in the previous example of SAR11, the relevant role of temperature in structuring the genomic variation of microbial populations in the ocean. Furthermore, 2742 SNVs and 13 SAAVs were detected that differentiate temperate from cold populations. The structure of protein variants from mesophilic and psychrophilic populations was compared, which provided insights into

494 the structural changes that may underpin adaptation to different temperature niches
495 and that are responsible for changes in functional and physical properties [97].
496 In another work, Da Silva and colleagues investigated the genomic
497 differentiation within three species of pico-phytoplankton in the Mediterranean Sea:
498 *Bathycoccus prasinos*, *Pelagomonas calceolata*, and *Phaeocystis cordata* [100].
499 Here, metagenomic reads from TARA Oceans stations in the Mediterranean Sea were
500 mapped to either reference genomes (*B. prasinos*), or transcriptomes (*P. calceolata*
501 and *P. cordata*) retrieved from the Mediterranean Sea or other regions. In general, *B.*
502 *prasinos* displayed a higher population differentiation than *P. calceolata* and *P. cordata*
503 in the Mediterranean Sea. In addition, results indicated that environmental selection
504 seems to shape the population-level diversity of *B. prasinos* in the Mediterranean Sea,
505 while *P. cordata* populations appear to be shaped by geographic distance (isolation
506 by distance) [100]. This study demonstrates that populations of different protist species
507 within the same functional group and with similar morphologies can exhibit varying
508 degrees of differentiation and be influenced by distinct mechanisms, such as selection
509 versus dispersal.
510 The studies discussed above required reference genomes or transcriptomes to
511 map against metagenomic reads. This is clearly a limitation, given that, at the moment,
512 there is no genomic or transcriptomic information for most microbial species.
513 Therefore, alternative reference-free approaches have been developed, which do not
514 need an alignment to a reference and can detect variants directly on unassembled
515 metagenomic reads. One such approach is metaVaR, which introduces the concept
516 of metavariant, which are variants detected in metagenomic reads [101]. Then,
517 metavariant species, or MVS, can be defined by clustering metavariants. Thus, an
518 MVS includes metavariants from the same species. MVSs can then be taxonomically

519 assigned by aligning variable loci against sequence databases [101]. Despite the
520 potential of this approach to investigate the population genomics of species with no
521 reference sequences, in reality, only a number of species are expected to present
522 enough metagenomic coverage and the number of metavariants needed to pass the
523 quality thresholds. For example, this approach was tested in a large dataset derived
524 from TARA Oceans that included millions of metavariants from 114 geographically
525 widespread marine samples, and only 113 MVSs were retrieved [101,102]. The 113
526 MSVs belonging to Metazoa, Chromista, Chlorophyta, Bacteria, and viruses were
527 analyzed across the North and South Atlantic Oceans, Southern Ocean, and the
528 Mediterranean Sea [86]. Population differentiation (as measured by the Fst index) was
529 higher among ocean basins than within basins for the analyzed species, which could
530 be attributed to higher connectivity within basins. Furthermore, unicellular organisms
531 (bacteria, unicellular eukaryotes, and viruses) displayed more population structure
532 than larger multicellular counterparts (zooplankton), which could be linked to different
533 dispersal capabilities affecting gene flow or different demographic histories (population
534 size, generation time). The primary drivers of population structure for the studied
535 species were oceanic currents (Lagrangian travel time), temperature, and salinity [86].
536 Yet, in this work, a large fraction of the population genomic differentiation could not be
537 explained, pointing to other abiotic (e.g., additional inorganic nutrients and pH) and
538 biotic variables (ecological interactions) that could contribute to population structure
539 [86]. All in all, this approach represents a valuable option for metagenome-based
540 population genomics when no reference genomes are available. Yet, this methodology
541 does not intend to replace reference-based methods, which according to the authors,
542 should be used whenever a reference is available [101].

Altogether, the previous studies show that a large complexity in terms of population-level diversity, structure, and fine-grained adaptations can be present within environmental microbial species. We can now access this underexplored dimension of diversity thanks to metagenome-based population genomics [80] (or metatranscriptomics) **(Figures 3 & 4)**. In addition, in multiple studies, selection seems to be central in structuring microdiversity, pointing to the fine-tuning of the ocean microbiome to environmental heterogeneity.



551

552 **Figure 4. Accessing the population-level dimension of diversity in marine microbes using**
553 **metagenomics.** The figure aims to provide a simple example of the additional information on population
554 structure that the metagenome-based population genomics approach can produce compared to 16S
555 rRNA surveys. Here, I use the MAG G4.480 (uncultured Flavobacteriales, ~95% completeness, and
556 <10% contamination) that we retrieved from the Mediterranean Sea (LTER Blanes Bay Microbial
557 Observatory; http://bbmo.icm.csic.es/). From this MAG, a fragment of the 16S rRNA gene (770 base
558 pairs) was extracted and then used to estimate the MAG abundance in the global ocean and the
559 Mediterranean Sea using the Ocean Barcode Atlas (OBA) [103] (https://oba.mio.osupytheas.fr/ocean-
560 atlas/); results are shown in Panel A. Only two 16S mTag [104] references from the OBA with >99%
561 sequence similarity with MAG G4.480 were considered (references AACY020490277.719.2228 &

EF572435.1.1502; both Flavobacteriales, Flavobacteriaceae, NS5 marine group). Furthermore, only surface samples originating from two size fractions (0.2-1.6 & 0.2-3.0 μm) from the TARA Oceans cruise were included. In sum, in Panel A, we observe the distribution of the MAG G4.480 as one single taxonomic entity. In Panel B, the diversity within this entity is explored using metagenome-based population genomics (Figure 3), and we notice that additional patterns emerge. In the upper section of Panel B, the Fst values (measuring population differentiation) among the investigated stations were clustered, and different clusters, which may correspond to populations, were colored (Fst ~0.2 was used to delineate clusters). Note that some clusters correspond to geographic regions (Panel B, lower section), for example, the clusters in the Mediterranean Sea, Red Sea, and Indian Ocean, suggesting that they could represent geographically delineated populations. These patterns are missed by the 16S rRNA gene (Panel A). The abundance of the Mediterranean MAG G4.480 across the global ocean and the Mediterranean Sea based on metagenomic read recruitment is shown in the lower section of Panel B. MAG abundances are indicated in RPKG (Reads Per Kilobase of MAG and Gigabase of metagenomic data). To obtain the Fst values and the abundances of the MAG G4.480 (Panel B), we followed the procedure indicated in Figure 3**,** which is partially implemented in POGENOM [81]. Only surface metagenomes from TARA Oceans with enough coverage (horizontal and vertical) of MAG G4.480 were used in downstream analyses, which explains the different numbers of stations included in Panels A and B.

Populations and contemporary evolution

Given the large population sizes of many microbial species, the ocean microbiome could evolve relatively fast compared to multicellular organisms with smaller populations [35]. Thus, substantial evolution could be expected at contemporary timescales (e.g., decades, centuries) [35]. Yet, we do not have a clear estimate of how fast the ocean microbiome may evolve. Understanding the tempo and mode of adaptation of marine microbes is essential in the context of global change, as evolutionary adaptation is one of the expected reactions of microbes to changing environmental conditions [105].

Evolutionary adaptation occurs by the accumulation of beneficial mutations over time. Populations of a given species could be depicted as entities that move in an adaptive landscape [106], which normally resemble mountain ranges, with local or global peaks that indicate areas of high fitness and valleys between them, which are areas of lower fitness **(Figure 5)**. Selection tends to push populations uphill in the adaptive landscape, and as populations climb different peaks, they become adapted **(Figure 5).**

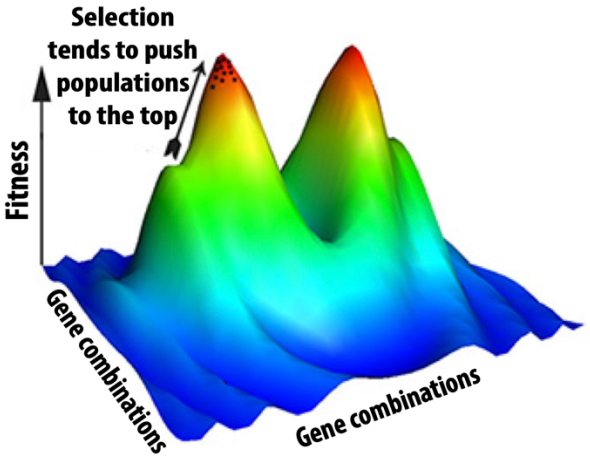


Figure 5. Adaptive landscape with two peaks. Adapted from © Laurence Loewe, 2016, CC-BY 4.0

In large microbial populations, many combinations of genotypes are potentially possible, which could explore adaptive landscapes more thoroughly than larger organisms with smaller populations **(Figure 5)**. The evolution of specific genotypes is determined by selection and drift, and the relative role of each process is predominantly dictated by the effective population size N_e and the selection coefficient s [107]. As mentioned, in species with large N_e, selection is expected to be more effective in fixing or removing mutations than in species with smaller N_e [49], potentially in a shorter amount of time. The estimated time to fixation of a neutral mutation is proportional to population size, being on average N_e (haploid) or $2N_e$ (diploid) generations. Thus, neutral mutations may remain for a long time in large microbial populations before being fixed or lost through drift, which aligns with results that were previously discussed [94,97]. While the probability of fixation for a beneficial mutation is approximately $2s$, where s is the selection coefficient, mutations with a slight fitness advantage may face challenges in increasing frequency, particularly in smaller populations where genetic drift is more influential [108]. Yet, in species with a large N_e, the likelihood of such mutations increasing in frequency is enhanced due to the reduced impact of genetic drift [49]. Even though the effective population sizes of

617 marine microbes are largely unknown [63], it is expected that for many species, it is
618 sufficiently large so that selection drives adaptation. All in all, due to the large N_e, small
619 or large changes in environmental conditions could facilitate the contemporary
620 adaptation of different microbial lineages.

621 Changing environmental conditions challenge the ocean microbiome [105].
622 New selective regimes (a changing adaptive landscape, **Figure 5**) are expected to
623 select from the available genetic diversity of microbial populations and from emerging
624 *de novo* mutations. This process is expected to promote evolution in contemporary
625 timescales. So far, microbial evolution experiments (in contemporary timescales) have
626 indicated three major trends: 1) significant phenotypic innovations can emerge (e.g.,
627 new metabolisms, growth rates), 2) high levels of evolutionary *parallelism* (i.e.,
628 repeated evolutionary changes), and 3) emergence of population structure, such as
629 genetically differentiated cell sub-groups [109,110].

630 In contrast to laboratory experiments, relatively little is known about microbial
631 evolution in the wild, and the interested reader is referred to Brennan & Logares for an
632 in-depth discussion [35]. Here, I will briefly mention two examples from aquatic (non-
633 marine) environments that illustrate the importance of metagenome-based population
634 genomics coupled with time-series metagenomics for understanding microbial
635 evolution in the wild. These studies typically use a DNA archive, including samples
636 from various time points, to track the evolutionary process. In the first example, Denef
637 and Banfield investigated the evolution of a natural acidophilic biofilm over 9 years in
638 Acid Mine Drainage (AMD) ecosystems [111]. An evolutionary rate of 1.3×10^{-9}
639 substitutions per nucleotide per generation was estimated for one MAG, and further
640 analyses showed how mutations could emerge and become fixed as a product of
641 selection and drift. Given the extreme nature of AMD environments and the low

642 immigration rates, it can be considered that mutations emerged *in-situ.* Determining
643 whether a mutation emerges in one location *de novo* or has arrived through
644 immigration is a challenge in these types of studies.

645 Another study examined 30 bacterial MAGs that were derived from
646 metagenomic samples collected over a nine-year period in a freshwater lake [112]. A
647 large SNV heterogeneity was found between and among populations. This suggests
648 varying mutation rates among species or populations or differences in immigration
649 history. Newly arrived immigrants may exhibit more homogeneous populations as they
650 have had less time to undergo diversification. SNVs frequencies showed marked
651 changes over time in some populations. For example, in one population, most of the
652 gene and SNV diversity disappeared during the investigated period, suggesting an
653 ongoing genome-wide selective sweep [43]. In turn, another population displayed
654 large, SNV-free genomic regions that appear to have swept through the populations
655 before the investigated period without removing diversity from other genomic areas,
656 pointing to a gene-specific sweep [112].

657 The two previous studies exemplify the insights that can be obtained on
658 contemporary microbial evolution in the wild through metagenome-based population
659 genomics coupled with time series. As of now, this approach appears to remain
660 underexplored in the context of oceanic studies. The connectivity of the surface ocean
661 complicates the application of the approach, as it is difficult to disentangle mutations
662 that originate in one location from those arriving via immigration. Nonetheless,
663 temporal trends in SNV frequencies, as well as changes in both gene and SNV
664 diversity, can offer valuable insights into the effects of shifting selective pressures
665 induced by climate change on the ocean microbiome. This is of particular relevance in
666 locations such as the Mediterranean Sea, which has experienced during the last years

667 an increase in the frequency and intensity of marine heatwaves [113]. While these
668 heatwaves have induced mass mortality events among multicellular marine
669 organisms, their impact on the marine microbiome remains poorly understood.
670
671 **Microbial populations in a changing ocean**
672 The ocean microbiome currently faces multiple challenges derived from
673 anthropogenic-induced climate change, such as sea-surface warming, decreasing O_2
674 and increasing CO_2 levels, acidification, changes in water circulation, changes in
675 nutrient inputs and other biotic factors (such as new parasites or predators) [105]. Thus
676 far, relatively few studies have investigated the reaction of marine microbes to long-
677 term global change, even though the associated selective changes can have
678 significant consequences in their community structure, populations, evolution, and
679 ultimately, in the biogeochemical cycles they mediate [105]. As a response to the
680 changing oceanic conditions, microbes are anticipated to undergo shifts in their
681 geographic distributions, alterations in community structure, modifications in gene
682 expression —including epigenetic changes—, and adaptations to the new
683 environmental conditions [35,105,114]. However, the relative significance of these
684 mechanisms in shaping the overall response remains uncertain. Population genomics
685 has the potential to provide new insights into the relative relevance of these processes
686 in the reaction of microbes to a changing ocean.
687
688 **CONCLUSIONS**
689 Beginning in the 90s with the onset of the “molecular revolution” and continuing into
690 the 2000s with the advent of High-Throughput Sequencing technologies, omics
691 approaches have significantly advanced our understanding of the ocean microbiome,

revealing the various lineages it harbors, their distributions, and metabolisms. Specific markers, such as the rRNA gene, provided a clearer dimension of the diversity that is contained in the ocean microbiome. Yet, the rRNA gene normally underestimates or misses the dimension of diversity that is found within individual species **(Figure 4)**. So far, only a limited number of studies have delved into the population-level diversity of environmental microbes. Understanding the population diversity of microbes is fundamental for a better comprehension of ecosystem function and the adaptation of microbes to different niches. Isolating and culturing environmental strain has been one of the main obstacles in accessing the species-level diversity of microbes. Today, the use of metagenomics and metatranscriptomics allows us to investigate the diversity that is present within species, bypassing the need for culturing. Population-level studies have the potential to open a new chapter in environmental microbiology, deepening our understanding of the ocean microbiome's composition, configuration, and intricate relationships with ecosystem functioning. This new knowledge will also be pivotal in the context of global change as we seek to comprehend the ocean microbiome's resilience or vulnerability, as well as its potential impact on broader Earth system processes.

LIST OF ABBREVIATIONS

AMD: Acid Mine Drainage

FACS: Fluorescence Activated Cell Sorting

Fst: Fixation index

GSSS: Gene-Specific Selective Sweep

GWSS: Genome-Wide Selective Sweep

HGT: Horizontal Gene Transfer

HTS: High-Throughput Sequencing

MAG: Metagenome-Assembled Genome

Mb: Megabases

MVS: Metavariant species

N: Census population size

N_e: Effective population size

OTU: Operational Taxonomic Unit

RPKG: Reads Per Kilobase of genome and Gigabase of metagenomic data.

SAAV: Single Amino-Acid Variant

SAG: Single-Amplified Genome

SNV: Single Nucleotide Variant

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Not applicable

Competing interests

The author declares that he has no competing interests.

Funding

This work was supported by the project MINIME (PID2019-105775RB-I00, Agencia Estatal de Investigación, Spain).

Authors' contributions

R.L. wrote the manuscript and prepared the figures.

Acknowledgments

I thank Fran Latorre for his assistance with preparing figures 3 & 4.

Authors' information

R.L. is a molecular and computational ecologist who specializes in aquatic microbial ecosystems. He possesses expertise in molecular biology, multiomics, and bioinformatics. R.L. earned his Ph.D. from Lund University in Sweden and is presently a tenured principal investigator at the Institute of Marine Sciences (ICM; https://www.icm.csic.es/en), CSIC, in Barcelona, Spain. He is the head of the log-lab (https://www.log-lab.barcelona), whose research agenda focuses on understanding the structuring, evolution, and dynamics of natural microbial communities and populations. Additionally, the log-lab aims to disentangle the complex network of microbial interactions in ecosystems and to link genomic content—across individual genomes and communities—to ecological functionality and evolutionary processes.

REFERENCES

1. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive earth's biogeochemical cycles. Science. 2008;320:1034–9.

2. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A. 1998;95:6578–83.

3. Suttle CA. Viruses in the sea. Nature. 2005;437:356–61.

4. Gasol JM, Kirchman DL, editors. Microbial Ecology of the Ocean. Wiley-Blackwell; 2018.

5. Bar-On YM, Milo R. The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet. Cell. 2019;179:1451–4.

6. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A. 2016;113:5970–5.

7. Falkowski PG, de Vargas C. Genomics and evolution. Shotgun sequencing in the sea: a blast from the past? Science. 2004. p. 58–60.

8. Falkowski P. The power of plankton. Nature. 2012;483:S17-20.

9. Jardillier L, Zubkov MV, Pearman J, Scanlan DJ. Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. ISME J. 2010;4:1180–92.

10. Li WKW. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. Limnol Oceanogr. 1994;39:169–75.

11. Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science. 2015;347:1257594.

12. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: Integrating terrestrial and oceanic components. Science. 1998;281:237–40.

13. del Giorgio PA, Duarte CM. Respiration in the open ocean. Nature. 2002;420:379–84.

14. Massana R. Eukaryotic Picoplankton in Surface Oceans. Annu Rev Microbiol. 2011;65:91–110.

15. Massana R, Logares R. Eukaryotic versus prokaryotic marine picoplankton ecology. Environ Microbiol [Internet]. 2012; Available from: http://dx.doi.org/10.1111/1462-2920.12043

16. Pedrós-Alió C, Acinas SG, Logares R, Massana R. Marine microbial diversity as seen by high throughput sequencing. In: Gasol JM, Kirchman DL, editors. Microbial Ecology of the Oceans. Hoboken, New Jersey: Wiley-Blackwell; 2018. p. 592.

17. Pedrós-Alió C. Marine microbial diversity: can it be determined? Trends Microbiol. 2006;14:257–63.

18. Eguíluz VM, Salazar G, Fernández-Gracia J, Pearman JK, Gasol JM, Acinas SG, et al. Scaling of species distribution explains the vast potential marine prokaryote diversity. Sci Rep. 2019;9:1–8.

19. Amann R, Rosselló-Móra R. After All, Only Millions? [Internet]. MBio. 2016. Available from: http://dx.doi.org/10.1128/mBio.00999-16

20. Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. Status of the Archaeal and Bacterial Census: an Update. MBio. 2016;7:e00201-16.

21. Louca S, Mazel F, Doebeli M, Parfrey LW. A census-based estimate of Earth's bacterial and archaeal diversity. PLoS Biol. 2019;17:e3000106.

22. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Sequencing of the Sargasso Sea. Science. 2004;304:66–74.

23. Duarte CM. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. Limnol Oceanog Bull. 2015;24:11–4.

24. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, et al. A Holistic Approach to Marine Eco-Systems Biology. PLoS Biol. 2011;9:e1001177.

25. Larkin AA, Garcia CA, Garcia N, Brock ML, Lee JA, Ustick LJ, et al. High spatial resolution global ocean metagenomes from Bio-GO-SHIP repeat hydrography transects. Scientific Data [Internet]. 2021;8. Available from: http://dx.doi.org/10.1038/s41597-021-00889-9

26. Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, et al. Marine microbial metagenomes sampled across space and time. Sci Data. 2018;5:180176.

27. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015;348:1261605.

28. Logares R, Deutschmann IM, Junger PC, Giner CR, Krabberød AK, Schmidt TSB, et al. Disentangling the mechanisms shaping the surface ocean microbiota. Microbiome. 2020;8:55.

29. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, et al. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. Cell. 2019;179:1068-1083.e21.

30. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. Nat Commun. 2018;9:373.

31. Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun Biol. 2021;4:604.

32. Shapiro BJ, Polz MF. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol. 2014;22:235–47.

33. García-García N, Tamames J, Linz AM, Pedrós-Alió C, Puente-Sánchez F. Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. ISME J. 2019;13:2969–83.

34. Larkin AA, Martiny AC. Microdiversity shapes the traits, niche space, and biogeography of microbial taxa [Internet]. Environmental Microbiology Reports. Wiley-Blackwell; 2017. p. 55–70. Available from: http://dx.doi.org/10.1111/1758-2229.12523

35. Brennan GL, Logares R. Tracking contemporary microbial evolution in a changing ocean. Trends Microbiol. 2023;31:336–45.

36. Mould DL, Hogan DA. Intraspecies heterogeneity in microbial interactions. Curr Opin Microbiol. 2021;62:14–20.

37. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. 2005;3:711–21.

38. Polz MF, Rajora OP. Population genomics : microorganisms. Cham, Switzerland: Springer; 2019.

39. Shapiro BJ, Polz MF. Microbial Speciation. Cold Spring Harb Perspect Biol [Internet]. 2015;7. Available from: http://cshperspectives.cshlp.org/content/7/10/a018143.abstract

40. Shapiro BJ, David L a., Friedman J, Alm EJ. Looking for Darwin's footprints in the microbial world. Trends Microbiol. 2009;17:196–204.

41. Schluter D. Evidence for ecological speciation and its alternative. Science. 2009;323:737–41.

42. Mayr E. Systematics and the Origin of Species. New York: Columbia University Press; 1942.

43. Cohan FM. Bacterial species and speciation. Syst Biol. 2001;50:513–24.

44. Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 2006;361:1929–40.

868 45. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species
869 definition for prokaryotes. Proc Natl Acad Sci U S A. 2005;102:2567–72.

870 46. Arevalo P, VanInsberghe D, Polz MF. A Reverse Ecology Framework for
871 Bacteria and Archaea. Population Genomics: Microorganisms. 2019. p. 77–96.

872 47. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. A Reverse Ecology
873 Approach Based on a Biological Definition of Microbial Populations. Cell.
874 2019;178:820-834.e14.

875 48. Shapiro BJ, Levade I, Kovacikova G, Taylor RK, Almagro-Moreno S. Origins of
876 pandemic Vibrio cholerae from environmental gene pools. Nature Microbiology
877 [Internet]. 2016;2. Available from: http://dx.doi.org/10.1038/nmicrobiol.2016.240

878 49. Charlesworth B. Effective population size and patterns of molecular evolution
879 and variation. Nat Rev Genet. 2009;10:195–205.

880 50. Xu J. Microbial population genetics. Wymondham: Caister Academic; c2010.

881 51. Walk ST, Feng PCH. Population Genetics of Bacteria: a Tribute to Thomas S.
882 Whittam. Feng, Peter C. H. Whittam, Thomas S. Walk, Seth T, editor. American
883 Society for Microbiology Press; 2011.

884 52. Logares RE. Does the global microbiota consist of a few cosmopolitan species?
885 Ecol Austral. 2006;16:85–90.

886 53. Logares R. Population genetics: the next stop for microbial ecologists? Cent Eur
887 J Biol. 2011;6:887–92.

888 54. Louca S. The rates of global bacterial and archaeal dispersal. ISME J [Internet].
889 2021; Available from: http://www.nature.com/articles/s41396-021-01069-8

890 55. Finlay BJ. Global dispersal of free-living microbial eukaryote species. Science.
891 2002;296:1061–3.

892 56. Sul WJ, Oliver TA, Ducklow HW, Amaral-Zettler LA, Sogin ML. Marine bacteria
893 exhibit a bipolar distribution. Proc Natl Acad Sci U S A. 2013;110:2342–7.

894 57. Woolfit M. Effective population size and the rate and pattern of nucleotide
895 substitutions. Biol Lett. 2009;5:417–20.

896 58. Lynch M, Bobay L-M, Catania F, Gout J-F, Rho M. The Repatterning of
897 Eukaryotic Genomes by Random Genetic Drift. Annu Rev Genomics Hum Genet.
898 2011;12:347–66.

899 59. Lynch M, Conery JS. The origins of genome complexity. Science.
900 2003;302:1401–4.

60. Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Ann Rev Mar Sci. 2017;9:231–55.

61. Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 2015;112:15690–5.

62. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A. 2012;109:18488–92.

63. Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. Comparing effective population sizes of dominant marine alphaproteobacteria lineages. Environ Microbiol Rep. 2014;6:167–72.

64. Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N, et al. Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proc Natl Acad Sci U S A. 2013;110:9824–9.

65. Biller SJ, Berube PM, Berta-Thompson JW, Kelly L, Roggensack SE, Awad L, et al. Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. Sci Data. 2014;1:140034.

66. Chen Z, Wang X, Song Y, Zeng Q, Zhang Y, Luo H. Prochlorococcus have low global mutation rate and small effective population size. Nature Ecology & Evolution. 2021;6:183–94.

67. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol. 2020;18:491–506.

68. Segata N. On the Road to Strain-Resolved Comparative Metagenomics. mSystems. 2018;3:1–6.

69. Leimbach A, Hacker J, Dobrindt U. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Curr Top Microbiol Immunol. 2013;358:3–32.

70. Povolotsky TL, Hengge R. Genome-Based Comparison of Cyclic Di-GMP Signaling in Pathogenic and Commensal Escherichia coli Strains. J Bacteriol. 2016;198:111–26.

71. Jung A, Metzner M, Ryll M. Comparison of pathogenic and non-pathogenic Enterococcus cecorum strains from different animal species. BMC Microbiol. 2017;17:33.

72. Pierce JV, Bernstein HD. Genomic Diversity of Enterotoxigenic Strains of Bacteroides fragilis. PLoS One. 2016;11:e0158171.

73. Moore LR, Rocap G, Chisholm SW. Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. Nature. 1998;393:464–7.

74. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science. 2006;311:1737–40.

75. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science. 2014;344:416–20.

76. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, et al. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012;335:48–51.

77. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, et al. Patterns of gene flow define species of thermophilic Archaea. PLoS Biol [Internet]. 2012;10. Available from: http://dx.doi.org/10.1371/journal.pbio.1001265

78. Rappe MS, Giovannoni SJ. The uncultured microbial majority. Annu Rev Microbiol. 2003;57:369–94.

79. Stepanauskas R. Single cell genomics: an individual look at microbes. Curr Opin Microbiol. 2012;15:613–20.

80. Denef VJ. Peering into the Genetic Makeup of Natural Microbial Populations Using Metagenomics. In: Polz MF, Rajora OP, editors. Population Genomics: Microorganisms. Cham: Springer International Publishing; 2019. p. 49–75.

81. Sjöqvist C, Delgado LF, Alneberg J, Andersson AF. Ecologically coherent population structure of uncultivated bacterioplankton. ISME J. 2021;15:3034–49.

82. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 2016;26:1612–25.

83. Van Rossum T, Costea PI, Paoli L, Alves R, Thielemann R, Sunagawa S, et al. metaSNV v2: detection of SNVs and subspecies in prokaryotic metagenomes. Bioinformatics. 2022;38:1162–4.

84. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017;27:626–38.

85. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 2021;39:727–36.

86. Laso-Jadart R, O'Malley M, Sykulski AM, Ambroise C, Madoui M-A. Holistic view of the seascape dynamics and environment impact on macro-scale genetic connectivity of marine plankton populations. BMC Ecol Evol. 2023;23:46.

971 87. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies
972 microbial strains in metagenomic datasets. Nat Biotechnol. 2015;33:1045.

973 88. Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, et al.
974 DESMAN: A new tool for de novo extraction of strains from metagenomes. Genome
975 Biol. 2017;18:1–22.

976 89. Quince C, Nurk S, Raguideau S, James R, Soyer OS, Summers JK, et al.
977 STRONG: metagenomics strain resolution on assembly graphs. Genome Biol.
978 2021;22:1–34.

979 90. Tan C, Cui W, Cui X, Ning K. Strain-GeMS: optimized subspecies identification
980 from microbiome data based on accurate variant modeling. Bioinformatics.
981 2019;35:1789–91.

982 91. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al.
983 Structure and function of the global ocean microbiome. Science. 2015;348:1261359.

984 92. Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al.
985 SAR11 clade dominates ocean surface bacterioplankton communities. Nature.
986 2002;420:806–10.

987 93. Carlson CA, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K.
988 Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of
989 the northwestern Sargasso Sea. ISME J. 2009;3:283–95.

990 94. Haro-Moreno JM, Rodriguez-Valera F, Rosselli R, Martinez-Hernandez F, Roda-
991 Garcia JJ, Gomez ML, et al. Ecogenomics of the SAR11 clade. Environ Microbiol.
992 2020;22:1748–63.

993 95. Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappé MS, Giovannoni SJ. High
994 intraspecific recombination rate in a native population of Candidatus pelagibacter
995 ubique (SAR11). Environ Microbiol. 2007;9:2430–40.

996 96. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappé MS, et al. Single-amino
997 acid variants reveal evolutionary processes that shape the biogeography of a global
998 SAR11 subclade. Elife. 2019;8:1–26.

999 97. Leconte J, Timsit Y, Delmont TO, Lescot M, Piganeau G, Wincker P, et al.
1000 Equatorial to Polar genomic variability of the microalgae Bathycoccus prasinos
1001 [Internet]. bioRxiv. 2021 [cited 2023 Feb 5]. p. 2021.07.13.452163. Available from:
1002 https://www.biorxiv.org/content/10.1101/2021.07.13.452163v3

1003 98. Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, et al.
1004 Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular
1005 specializations at the base of the green lineage. Genome Biol. 2012;13:R74.

99. Vannier T, Leconte J, Seeleuthner Y, Mondy S, Pelletier E, Aury J-M, et al. Survey of the green picoalga Bathycoccus genomes in the global ocean. Sci Rep. 2016;6:37900.

100. Da Silva O, Ayata S-D, Ser-Giacomi E, Leconte J, Pelletier E, Fauvelot C, et al. Genomic differentiation of three pico-phytoplankton species in the Mediterranean Sea. Environ Microbiol. 2022;24:6086–99.

101. Laso-Jadart R, Ambroise C, Peterlongo P, Madoui M-A. metaVaR: Introducing metavariant species models for reference-free metagenomic-based population genomics. PLoS One. 2020;15:e0244637.

102. Arif M, Gauthier J, Sugier K, Iudicone D, Jaillon O, Wincker P, et al. Discovering millions of plankton genomic markers from the Atlantic Ocean and the Mediterranean Sea. Mol Ecol Resour. 2019;19:526–35.

103. Vernette C, Henry N, Lecubin J, de Vargas C, Hingamp P, Lescot M. The Ocean barcode atlas: A web service to explore the biodiversity and biogeography of marine organisms. Mol Ecol Resour. 2021;21:1347–58.

104. Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol. 2014;16:2659–71.

105. Hutchins DA, Fu F. Microorganisms and ocean global change. Nature Microbiology. 2017;2:17058.

106. Loewe L. Systems in Evolutionary Systems Biology. In: Kliman RM, editor. Encyclopedia of Evolutionary Biology. Oxford: Academic Press; 2016. p. 297–318.

107. Kimura M. Evolutionary rate at the molecular level. Nature. 1968;217:624–6.

108. Perfeito L, Fernandes L, Mota C, Gordo I. Adaptive mutations in bacteria: High rate and small effects. Science. 2007;317:813–5.

109. Hindre T, Knibbe C, Beslon G, Schneider D. New insights into bacterial adaptation through in vivo and in silico experimental evolution. Nat Rev Microbiol. 2012;10:352–65.

110. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of molecular evolution over 60,000 generations. Nature. 2017;551:45–50.

111. Denef VJ, Banfield JF. In Situ Evolutionary Rate Measurements Show Ecological Success of Recently Emerged Bacterial Hybrids. Science. 2012;336:462–6.

112. Bendall ML, Stevens SLR, Chan LK, Malfatti S, Schwientek P, Tremblay J, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10:1589–601.

113. Garrabou J, Gómez-Gras D, Medrano A, Cerrano C, Ponti M, Schlegel R, et al. Marine heatwaves drive recurrent mass mortalities in the Mediterranean Sea. Glob Chang Biol. 2022;28:5708–25.

114. Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, et al. Scientists' warning to humanity: microorganisms and climate change. Nat Rev Microbiol [Internet]. 2019; Available from: http://www.nature.com/articles/s41579-019-0222-5