Selective constraints on global plankton dispersal

B.A. Ward[1,*], B.B. Cael[2], S. Collins[3], and C.R. Young[2]

[1]Ocean and Earth Science, University of Southampton, Southampton, UK
[2]National Oceanography Centre, European Way, Southampton, UK
[3]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, UK
[*]Corresponding author: b.a.ward@soton.ac.uk

Abstract

Introduction

Marine microbial communities are highly interconnected assemblages of organisms shaped by ecological drift, natural selection and dispersal. The relative strength of these forces determines how ecosystems respond to environmental gradients, how much diversity is resident in a community or population at any given time, and how populations reorganise and evolve in response to environmental perturbations (**?**). Over evolutionary timescales, the outcome of these processes will affect community assembly through diversification, ecosystem function through biogeochemical cycling, and ultimately the resilience of marine ecosystems to environmental change (**?**). Therefore, understanding the mechanisms that lead to niche diversification and biogeographic structure in microbial communities is a fundamental pursuit of marine microbial research (**?**). A central question is to what degree are biogeographic patterns attributable to local selection based on contemporary environmental factors, or to independent stochastic processes occurring in geographic-isolated regions (Figure **??** and **?**).

Recent analysis of metagenomic data (Figure **??** and **?**), has shown that large-scale trends in community composition are correlated both with environmental

variables and community composition, with distinct clusters emerging along environmental gradients and among the most closely connected sites. When sample sites are clustered based on Jaccard dissimilarity of operational taxonomic units, there is discernible ecological similarity among sites within the the same ocean basins (Figure **??**a), although we also see sites within the same ocean basin clustered far apart and sites from geographically remote locations clustered together (Figure **??**b). The broad patterns appear to reflect geographic proximity, and perhaps to a lesser extent, environmental selection (**?**, and Figures **??** and **??**). Nonetheless, the mechanisms underlying the observed biogeography remain uncertain. The roles of selection and dispersal have both been examined using global-scale models, but typically with one in isolation from the other. On one hand, population dynamic models have focused on the role of selection from among a universal background of candidate species (**?**). This is the view framed in the **?** hypothesis: 'everything is everywhere, but the environment selects'. On the other hand, a number of studies have addressed the question of global gene flow in oceanic microbial communities, using particle tracking models to assess connectivity through the surface waters (**??**), but these have typically assumed ecological neutrality (**?**), and have thus ignored the role of selection. While some studies find that the ocean surface is very rapidly connected on timescales of decades or less (**?**), others suggest that current rates of passive dispersal are insufficient to overcome biogeographic differences created by chance mutations occurring in geographically isolated regions of the ocean (**?**).

I would argue that the "everything is everywhere and the environment selects" needs to be broken down into first, can everything get everywhere, which is a (relatively) simple matter of connectivity that treats cells as passive particles. The second bit is given that a genotype arrives at a given location, can it grow? This requires including selection, which breaks down into first, is that patch already occupied by someone with higher fitness, and second, if not, is the fitness advantage,population size, and invasion frequency of the invader high enough to invade from rare - is the invading population likely to escape stochastic loss while relatively rare? (I think we need to spell out all the moving parts in the "everything is everywhere and the environment selects" idea, because it is not as simple as it sounds).

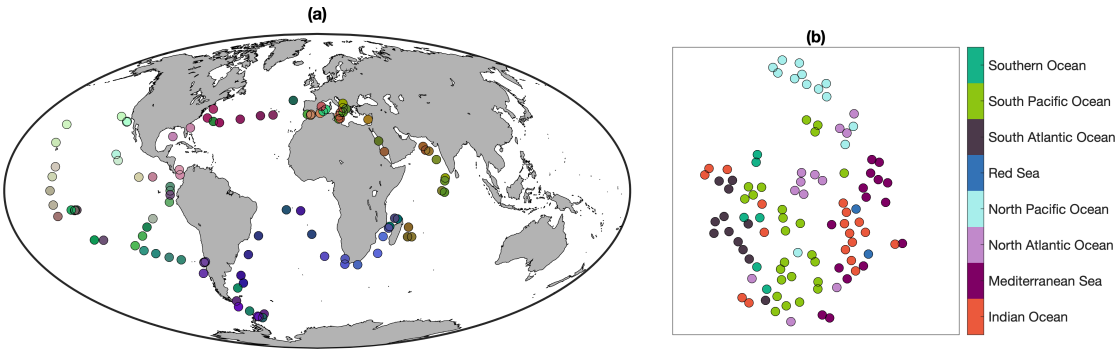

Figure 1: Taxonomic community similarity clusters in the 0.8 - 5 μm size fraction across *Tara Oceans* sites (replotted using data from **?**). (**a** and **b**) Community similarity is shown with colours by projecting the Taxonomic Jaccard dissimilarity matrix into the 'rgb' colour space using the t-SNE dimension-reduction algorithm (**?**). (**b**) Site x and y coordinates (dimensionless) are based on a 2-dimensional projection using the same algorithm.

In order to distinguish between the biogeographic effects of selection and dispersal, we need a framework that accounts for both processes together. In a case where global dispersal is fast relative to competitive exclusion, we should expect community composition to always reflect the local environmental conditions and for local composition to be largely predictable from environmental conditions in the absence of strong barriers to invasion imposed by local adaptation. (Figure **??**). However, the relative rates of dispersal and competitive exclusion will depend on physical dispersal rates in the first instance and differences fitness between local populations and invaders in the second, with larger fitness differences (stronger local adaptation) decreasing both the probability and rate of invader success for a given amount of physical dispersal.In the alternative case with limited connectivity between ocean regions, we should expect a more unpredictable community structure in any given location regardless of the ability of populations to adapt to their local conditions, with geographically-isolated but otherwise similar environments displaying significant differences in their taxonomic composition.

In this paper, we develop a population genetic model representing a single clonally-reproducing plankton population, embedded within an empirically-constrained representation of ocean circulation. In contrast to previous studies, our model accounts for population size, stochastic demography, natural selection, adaptation

and transport through the ocean interior. We find that estimates based on dispersal pathways restricted to the ocean surface markedly underestimate global connectivity. With a more realistic transport term accounting for dispersal at all depths, we find that varying the degree of selection and adaptation leads to very different model outcomes in terms of community biogeography and global connectivity. We show that accounting for the ability of populations to adapt to their thermal environment allows selection based on thermal niches to act as a major constraint on dispersal by favoring locally adapted populations even in the face of invaders being introduced by mixing, with the clear effects on biogeographic organisation at the global scale.

Simulations

To assess the rate of planktonic dispersal across the global ocean, we developed a model that tracks the relative abundances of adapting subpopulations in a globally distributed metapopulation, with spatially varying carrying capacity, $\mathbf{N}$. At the beginning of each simulation, a resident subpopulation is assumed to have population frequency of 1 throughout the global ocean. However, at each of 94 'seed locations' distributed more or less evenly around the ocean (dots in Figure **??**), the resident subpopulation is replaced with a taxonomically distinct (but ecologically identical) local subpopulation. From this initial condition, the model is integrated for 100 years in discrete time. Every six hours, plankton cells are dispersed by the ocean circulation. Every 24 hours the population is replaced with a new generation of $\mathbf{N}$ cells, drawn at random from the existing generation, with probabilities given by the relative abundance of each subpopulation and, where appropriate, a temperature-dependent selection coefficient, s (**?**). In regions where a subpopulation is present in high abundance, this stochastic process has no significant effect on the relative abundance, but it introduces a meaningful chance of local extinction wherever abundances are low (such as at edges of a subpopulation's range).

Our main set of simulations tracks the dispersal of a globally-abundant *Prochlorococcus* population with a cellular diameter of approximately 0.6 microns, setting $\mathbf{N}$ to the depth-integrated cellular abundance within each grid box (**?**). The results presented below are derived from from simulations based on a single repeating

year with time-invariant environmental temperatures and population carrying capacities. We also performed simulations where these variables followed a seasonal cycle, finding that the results were not overly sensitive to the change (Figure **??**).

Ecologically neutral dispersal by surface transport. We initially considered a scenario where cells are transported exclusively within the surface layer, with all subpopulations equally well adapted at all temperatures (i.e. ecologically neutral; **?**). The dark blue lines in Figure **??**a show the timescales over which the 94 *Prochlorococcus* seed subpopulations reach the rest of the ocean. Largely in agreement with **?**, almost 90% of the surface ocean is connected within a decade.

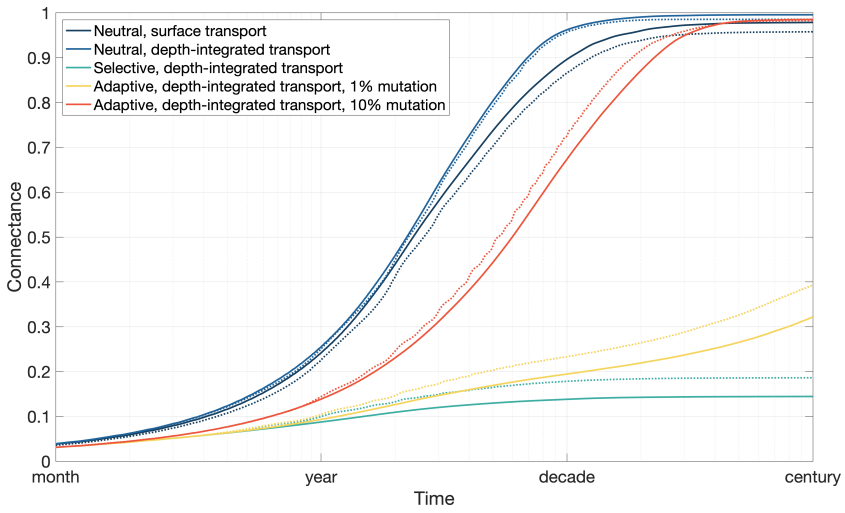

Figure 2: Fraction of connections between the 94 seed locations and the rest of the ocean through time. Solid lines show the results of simulations with time-invariant temperatures and carrying capacities. Dashed lines show the results of simulations with seasonally varying temperature and carrying capacity.

The global dispersal of the ecologically neutral subpopulations is broken down further in Figure **??**a. Here immigration times (background colours) suggest that temperate latitudes are generally more easily invaded than the equatorial regions. Conversely, emigration times (coloured dots) suggest that subpopulations initialised at lower latitudes are more rapidly dispersed throughout the ocean than those from higher latitudes. Plotting immigration times against emigration times for the 'surface-only' simulation (Figure **??**) confirms that low latitudes act as

source regions, with slow immigration and fast emigration, while high latitudes act as sink regions, with fast immigration and slow emigration.

These regional differences in immigration and emigration timescales are explained by the surface circulation patterns shown in Figure **??**a. The two-dimensional surface transport vectors are highly divergent in equatorial upwelling regions, driving a consistent efflux of cells that must be topped up to the carrying capacity by reproduction of the local resident population. These regions thus export cells to the rest of the ocean, while remaining resistant to immigration. The sub-tropical gyres, meanwhile, are characterised by convergent flow, with a consistent influx of cells diluting the local resident populations. These regions are thus easily invaded and are slower to export cells to the rest of the ocean.

Depth-integrated transport. The assumption that horizontal dispersal of plankton occurs only in the surface layer ignores the potential role of subsurface connectivity. To test the sensitivity of our results to this pathway, we calculated the depth-integrated horizontal transport of cells across the entire water column, weighting transport fluxes at each depth by the local population abundance. After this adjustment to the transport component, we repeated our initial experiment in the same way. Allowing for sub-surface transport slightly accelerates global plankton dispersal (pale blue line in Figure **??**), with more than 97% of the ocean connected in less than a decade. Spatially, the global distribution of immigration and emigration times has decreased everywhere (Figure **??**), with the the differences in timescales between source and sink regions largely obliterated in the less divergent depth-integrated flow field (black dots in Figure **??**).

Selection between types. The previous experiments have assumed that all subpopulations are equally well-adapted to conditions throughout the entire ocean, but we know that changing conditions select for different phenotypes along environmental gradients (**??**), and that dispersing populations will be selected against as they stray beyond their optimal environments. One component of natural selection is genotype sorting, where more fit types can replace less-fit ones.

To test the influence of these effects replace "these effects" with "genotype sorting based on standing variation", we focussed on a single trait, assigning thermal

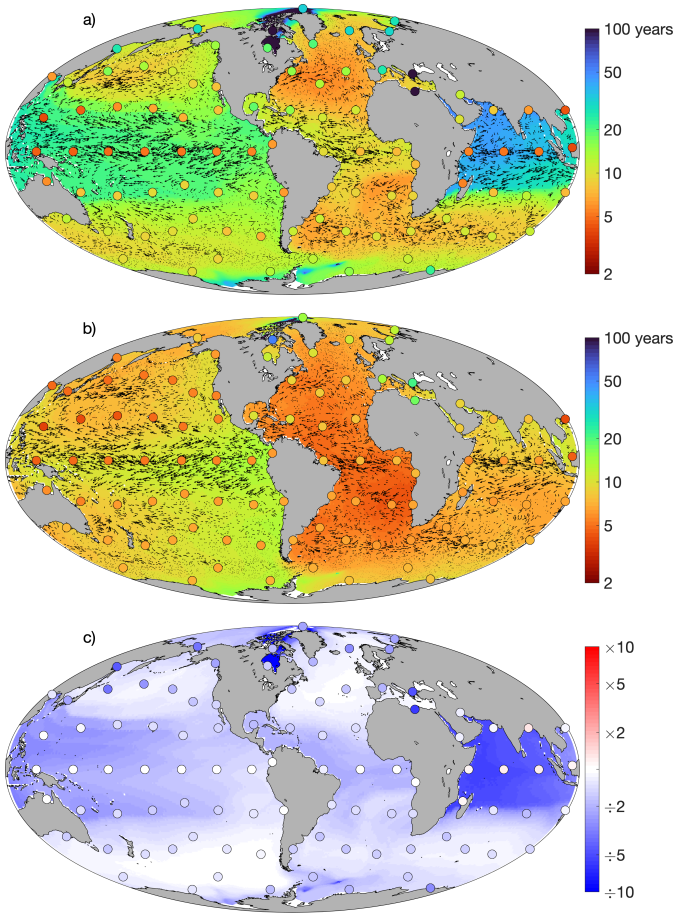

Figure 3: Immigration and emigration timescales (years) for ecologically neutral *Prochlorococcus* subpopulations, given (a) surface only transport, and (b) depth-integrated transport. Taxonomically distinct subpopulations were seeded in each of the 94 locations marked with dots. Emigration times, represented by the coloured dots, are defined as the time taken for each seed subpopulation to disperse to 95% of all locations). Immigration times, represented by the background colours, are defined as the time taken for 95% of all seed subpopulations to arrive in each location. Planktonic transport velocities are shown as vectors. (c) Relative changes in global immigration and emigration times when switching from surface-only to depth-integrated transport (b÷a).

tolerance curves such that populations are preferentially selected when ambient temperatures align with their thermal optima (Equation **??**). Each seed population is assigned a thermal optimum matching the average temperature at its initial location. At the same time, the global resident population is divided into 77 subpopulations, each with thermal optima matching the average temperature at its initial location which is consistent with the existence of locally adapted resident populations, but ignores the ability of genotypes to themselves evolve over time (see next section) - might be worth noting that genotypes sorting without mutation is how most ESMs represent evolution???. The model was then evaluated with the depth-integrated circulation scheme.

The global dispersal of the 94 seed populations is severely restricted by temperature-based selection (green line in Figure **??**), with global connectivity not rising above 15% in the 100 year simulation.

Adaptation within types. Temperature-related selection places a strong constraint on the dispersal of thermally-adapted populations. If populations are to overcome this restriction, they must adapt dynamically to their environments by generating heritable phenotypic changes over time (**??**). We included this capacity in the model by allowing all subpopulations to produce a small fraction of mutants I would replace "mutants" with "offspring" and add the following, possibly in SI: this incorporates a large range of molecular mechanisms, including plastic and heritable plastic responses, standard mutations, sex, and horizontal gene transfer. with different thermal optima (see SI-Methods). We initialised the experiment as before, with each subpopulation optimally adapted to its local temperature, but allowed for a small mutational flux maybe call it trait diffusion rate or flux? between adjacent phenotypes (**??**). In line with previous validated studies,We performed simulations with mutation rates of 1% and 10%. I think that Beckmann does a good job of justifying why trait diffusion mutation rates look so high relative to point mutation rates or per-site substitution rates, so we should just cite them and move on... but we could call it a "trait diffusion rate" rather than a "mutation rate", which I think is a more apt description of what we're actually representing. Still, 10% is high, and we could couch it as a proof of principle showing that you still get local adaptation hindering population homogenization, and that the

actual effect is likely to be more severe, in line with the 1% trait diffusion rate

Even with an extremely high mutation rate of 10%, global dispersal is markedly restricted by selection effects, with 95% connectivity only achieved after approximately 30 years (orange line in Figure **??**). When the mutation rate is set to 1%, just over 30% of the ocean has been connected within 100 years (yellow line in Figure **??**).

The global distributions of an illustrative seed population, 100 years after it was initialised in the central North Atlantic (35°N, 46°W) is shown in four illustrative cases in Figure **??**. In the neutral model the highlighted population has complete global coverage, with highest concentrations in the Atlantic subtropical gyres (Figure **??**a). Without selection, all seed populations are globally dispersed after 100 years, and community similarity appears to change smoothly as a function of geographic distance across the *Tara Oceans* sites. Sites within each ocean basin often cluster together, but there is little discrete separation between sites in adjacent basins.

With temperature-based selection enabled, but without adaptation, the distribution of the seed population is restricted to a relatively small area within the North Atlantic subtropical gyre (Figure **??**d), in waters between 5 and 28°C. The population is unable to disperse beyond its original North Atlantic habitat, excluded from thermally suitable environments in other ocean basins by their inability to succesfully traverse warmer or colder regions. In this case we see highly distinct clusters of sites even within ocean basins, indicative of the strong niche separation by temperature. Despite the presence of similar temperature niches in multiple ocean basins, only three clusters include sites drawn from different basins (Red Sea with Indian, N. Pacific with S. Pacific and S. pacific with S. Atlantic).

The distribution of the same seed population when it is allowed to adapt with a mutation rate of 1% is shown in Figure **??**g. After 100 years the lineage has dispersed much further into the South Atlantic, but the majority of its descendants remain trapped within the North Atlantic subtropical gyre. With mutation enabled we see fewer and slightly larger clusters, with more cross-basin clustering.

Only when the mutation rate is increased to 10% does the seed population attain similar global dispersal to the neutral case after 100 years, and even then its distribution is centred more strongly on its original Atlantic habitat Figure **??**j.

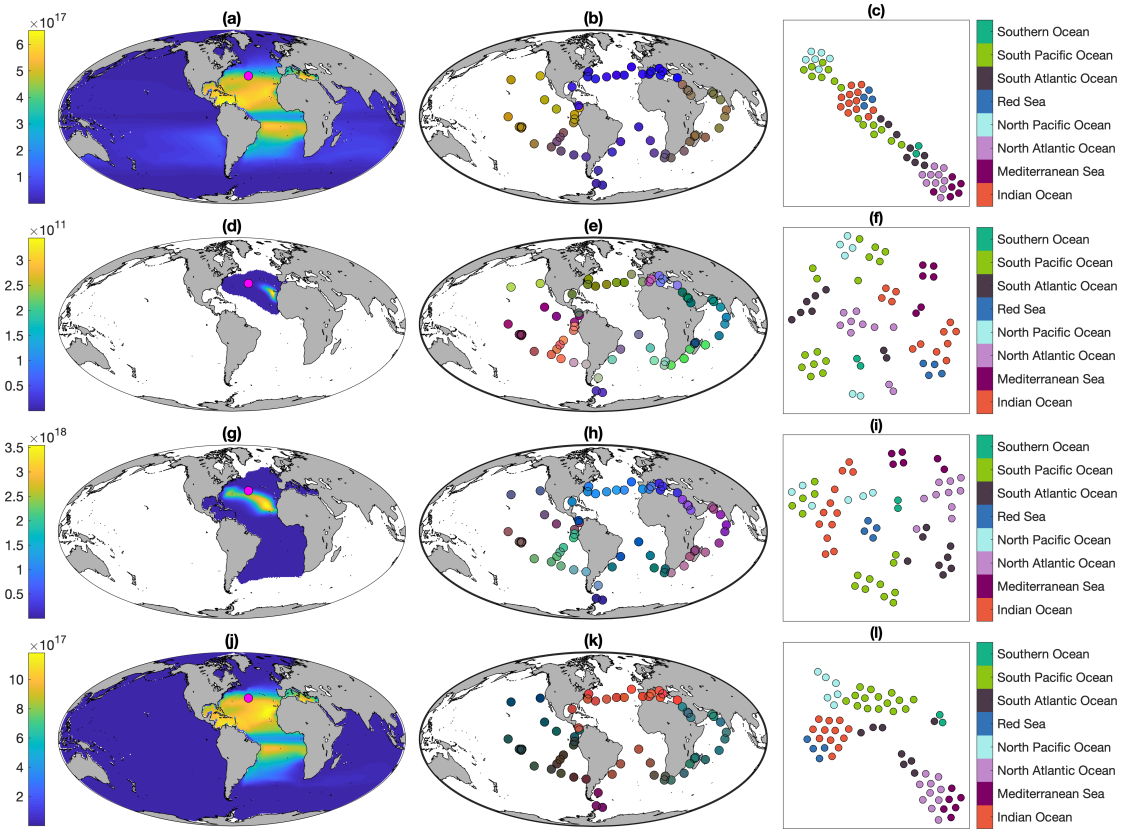

Figure 4: Global dispersal and taxonomic clustering of modelled subpopulations. Each row represents a different experiment: Row 1, neutral case; Row 2, selective case; Row 3, adaptive case (1% mutation rate); Row 4, adaptive case (10% mutation rate). The left-hand column maps the global abundance distribution of individual seed populations, initialised at the site indicated by the pink dot. The right-hand column shows population similarity clusters at the same *Tara Oceans* sites as Figure **??**. All panels show results after 100 years of dispersal.

With this extremely high rate of adaptation, the global community clusters strongly by ocean basin, but having some degree of selection leads to more distinct separation of sites in different ocean basins, relative to the neutral case. I really like this case study. I find it really clear. Hopefully others do as well?

Discussion

Plankton circulating within the global ocean are not dispersed as inert tracers. With their growth and relative fitness affected by the changing physical, chemical and biotic environment, populations are continually under selection and thus adapting as a function of their environmental setting. In environments outside their optimal habitat, dispersing populations are likely to be outcompeted by better adapted local populations, with an increasing risk of local extinction as their abundances decline there is a whole literature on this in the freshwater and even more so in the terrestrial ecology world, so I'll go find some nice reviews on invasions and local adaptation. This selective process has the capacity to place very strong constraints on the global dispersal of individual populations, and hence on the flow of genetic information from one ocean region to another. In our experiments, global connectivity only seems to be assured – on timescales of decades to centuries – when invading? subpopulations are able to rapidly adapt to changing conditions as they are dispersed.

Our simulations imply that even while the marine plankton are rapidly dispersed by the ocean circulation, significant barriers to establishment after dispersal exist – even for highly abundant and rapidly evolving microbial taxa. This has important implications for the study of plankton biogeography and community assembly, and for the interpretation of a growing archive of bioinformatic information. In particular, to what extent might local community assembly in any one part of the ocean be constrained by its isolation from other ocean regions? In other words, is 'everything really everywhere' as **?** suggests, or is a species' global distribution limited by dispersal?

So, my take on the above, in light of what the model shows is that dispersal is somewhat limiting - it takes time, but those times are short, especially when you take sub-surface routes into account. It's establishment that is limiting, and how limiting it is depends on how fast invaders can adapt relative to residents. Maybe we frame it as the effects of selection and adaptation on establishment of migrants? I think that without making it clear what the selection/adaptation are influencing in terms of biology, it gets a bit hard to follow.

With physical rates of dispersal in the model well constrained (**?**), the bal-

ance between selection, dispersal and adaptation on the establishment of migrant subpopulations over their dispersal range appears to have a pronounced affect on the global biogeography of phytoplankton (Figure **??**). In the neutral case, we find that abundant populations are rapidly distributed throughout the global ocean with gradual changes in community structure across distance characteristic of passive mixing. Enabling temperature-based selection places strong constraints on the establishment of immigrant subpopulations during global dispersal, with distinct communities emerging in environmentally dissimilar regions of the same ocean basins (Figure **??**e,f). At the same time, very different communities can emerge in otherwise environmentally similar regions, especially at lower latitudes, for which all interconnecting pathways must pass via the polar oceans where warm-adapted types are rapidly attenuated. Here, we assume that selection acts over the entire trajectory of subpopulations; the ability to survive transport through environments where reproduction is not possible, and resume growth in a more suitable environment, may present an additional complication. The ability to sustain both within basin and across basin community differences are both diminished with increasing rates of phenotypic adaptation, although the latter appears more robust over the 100 year timescales examined here.

Faster rates of adaptation allow populations to adjust their traits as they are dispersed across environmental gradients, thus achieving global distributions much wider than their original habitat. Nonetheless, even with very rapid rates of adaptation, dispersing model populations remain extremely rare in regions that are far from their original seed locations. This is consistent with rank-abundance distributions characterised by a long-tail of rare species in marine microbial communities (**?**), with community structure maintained as the homogenising effects of ocean mixing are counterbalanced by local selectionmaintaing local adaptation een in the face of constant mixing. Even when immigrant populations can adapt to local conditions, the necessity to compete with similarly adapted but much more abundant residents means that the incoming populations remain scarce. will hunt down some nice references for this in freshwater or experimental systems

Ultimately, the degree to which the dispersing populations are selected against in non-optimal environments, and the degree to which they are able to adapt, will determine the ubiquity, or otherwise, of marine microbial species. We have shown

that geographic proximity can be a strong correlate of microbial community structure even in an ecologically neutral model (**?**). However, the existence of distinct community clusters both within and across ocean basins (Figure **??**) is perhaps indicative of a system where everything is not everywhere, *because* the environment selects. Correctly accounting for selection, speciation and limited dispersal therefore appears critical to understanding community structure and biogeography in the ocean. Alongside global metagenomic surveys, our results suggest that when developing models of marine microbial biogeography and ecology we need to go beyond the assumption that everything is everywhere and to consider the selective limitations to dispersal and the adaptive means by which these are overcome. We have shown that a number of factors influence connectivity, including subsurface circulation, population abundance, generation time and selection. We note that while we could demonstrate temperature-based selection is likely to constrain dispersal, we have only considered one of many factors that are known to affect plankton fitness. We expect that the need to simultaneously adapt multiple traits along environmental gradients would likely decrease the rate of effective adaptation, further increasing selective constraints on dispersal. On the other hand, we have also neglected a number of factors that may serve to increase connectivity, with the existence of dormant stages (with low growth and mortality) likely to play a key role for some major groups, such as diatoms.

SI Supplementary Material

Methods

The Evolutionary Plankton Metacommunity Dynamics (EPMD) model considers the global distribution of an arbitrary number of planktonic subpopulations distributed across a two-dimensional (latitude and longitude) ocean grid. The probability of survival for each subpopulation in each generation is a function of its relative abundance and (optionally) its thermal tolerance to the local environmental temperature (**?**). Plankton cells are circulated in physical space according to a realistic ocean circulation model (**??**).

Passive dispersal by the ocean circulation. Plankton cells are transported between grid boxes using a $[J \times J]$ oceanic 'transport matrix' $\mathbf{A}$ that describes the transport of K populations of neutrally buoyant cells between J points in the ocean grid (**?**). This transport can be written as

$$\mathbf{X}_{t+1} = \mathbf{A}\mathbf{X}_t \tag{SI.1}$$

Here $\mathbf{X}_t$ is the $[J \times K]$ matrix of population abundances in each grid box of the GCM. Each element of the transport matrix $\mathbf{A}$ describes the transport of cells between source boxes (columns) and recipient boxes (rows). The transport matrix represents the annual mean transport during a single year of the "Estimating the Circulation and Climate of the Ocean" (ECCO) version 4 ocean model (**??**). It represents physical transport attributable to advection, diffusion and parameterised sub-grid-scale processes in the ocean model with 6-hourly resolution. Results in the main text use annual average circulation, temperature and acrrying capacities. We also performed simulations using monthly resolved temperatures and carrying cappacities, finding that our results were not overly sensitive to the change (Figure **??**).

Stochastic demography. We used a stochastic population model to estimate the global abundance of 94 ecologically neutral subpopulations, each initialised

with population abundance equal to the carrying capacity, $\mathbf{N}$, at unique "seed locations" that were distributed approximately evenly around the surface ocean. In addition to the seed populations, we included one additional tracer representing a globally resident species, with a population abundance of $\mathbf{X} = 0$ at all seed locations, and $\mathbf{X} = \mathbf{N}$ throughout the rest of the surface grid.

Under the assumption that all species have equal fitness, the number of individuals $\mathbf{X}$ surviving at each generation is drawn randomly from the local population (after oceanic transport and mutation) with probability $\mathbf{p}$ equal to the local population frequency ($\mathbf{x} = \mathbf{X}\mathbf{N}^{-1}$). Under these assumptions, the expected population size in each generation is given by the multinomial distribution,

$$\mathbf{X} \sim \mathcal{M}(\mathbf{N}, \mathbf{p}) \tag{SI.2}$$

For large values of $\mathbf{N}$ considered here, equation **??** is reasonably approximated by a normal distribution.

$$\mathbf{X} \approx \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \tag{SI.3}$$

with $\mu = Np$ and $\sigma = \sqrt{Np(1-p)}$ for each element in $\mathbf{X}$.

Selection. Selection can be further incorporated through the selection vector $\mathbf{s}$, that assigns each population in $\mathbf{X}$ a relative fitness of $\mathbf{s}$. With a local water temperature of T, a phytoplankton population with thermal optimum T_{opt} and thermal niche breadth w will have a selection coefficient of

$$s = \exp\left[-\left(\frac{T_{env} - T_{opt}}{w}\right)^2 \right] \tag{SI.4}$$

This is incorporated into the probability of selection such that the sum of all probabilities remains equal to one. The probability of selection for population j is thus

$$p_j = x_j s_j \Big[\sum_{k=1}^{K} x_k s_k \Big]^{-1} \tag{SI.5}$$

Adaptation. Adaptive evolution is enabled by diverting a small fraction of successfully reproducing individuals to populations with higher or lower thermal op-

tima. In practice, this is achieved by multiplying the population matrix [**X**] by the $K \times K$ mutational matrix **M** after each reproductive cycle (**??**).

$$\mathbf{X}_{t+1} = \mathbf{M}\mathbf{X}_t \tag{SI.6}$$

The mutation matrix itself is defined by the parameterised mutation rate (here 1 or 10%). This is the fraction of daughter cells in each population that, as mutants, are diverted to the neighbouring phenotypic class in each generation.

Simulations. In each case the model was integrated for 100 years. The transport matrix was applied every six hours, with selection and adaptation applied every 24 hours (or every 120 hours in the simulations with slower generation times).

Seed locations. Seed locations were identified by iteratively projecting a subdivided rectangular grid onto the surface of a sphere. Surface ocean coordinates from the GCM grid were then mapped onto these points (by shortest euclidean distance in cartesian coordinates). Finally, the 94 unique coordinates were mapped back to their nearest point on the GCM grid. The model was initialised with one tracer for each of these points, with a genotype frequency of $\mathbf{p} = 1$ at the seed location, and zero everywhere else.

Community dissimilarity.

Geographic and Environmental distance.

Depth-integrated transport. *Describe how depth-integrated transport matrix was derived.*

Mass conservation correction. Numerical constraints within the ocean general circulation model mean that some off-diagonal elements may be negative. To remove these artefacts, negative fluxes out of each grid box are converted to positive fluxes into the box. If we first define the matrix of negative off-diagonal elements as follows,

$$N_{i,j} = \begin{cases} 0 & \text{if } i = j \text{ or } A_{i,j} \geq 0 \\ A_{i,j} & \text{if } i \neq j \text{ and } A_{i,j} < 0 \end{cases} \tag{SI.7}$$

We first move the off-diagonal negatives to their transpose, changing the sign.

$$\mathbf{B} = \mathbf{A} - \mathbf{N} + (-\mathbf{N}^{\top}) \tag{SI.8}$$

To conserve mass, changes in the off-diagonals must be compensated by an equivalent change on the diagonals. This is achieved by adding both the row and column sums of the negative off-diagonals to the diagonal.

$$C_{ii} = B_{ii} + \sum_{j=1}^{J} N_{i,j} + \sum_{j=1}^{J} N_{j,i} \text{ (for all } i) \tag{SI.9}$$

The resultant matrix is positive-definite on the off-diagonals, and conserves mass.

Finally, the transport matrix is converted from a volume flux per time ($m^3\ d^{-1}$) to a unitless operator; first dividing through by the volume of the recipient cells ($\mathbf{v}$), then multiplying by the selected time step (Δt), and adding the identity matrix ($\mathbf{I}$).

$$\mathbf{D} = \frac{\mathbf{C}}{\mathbf{v}} \times \Delta t + \mathbf{I} \tag{SI.10}$$

Code availability

Model code and analysis tools can be downloaded from `github.com/geebes/EPMD`

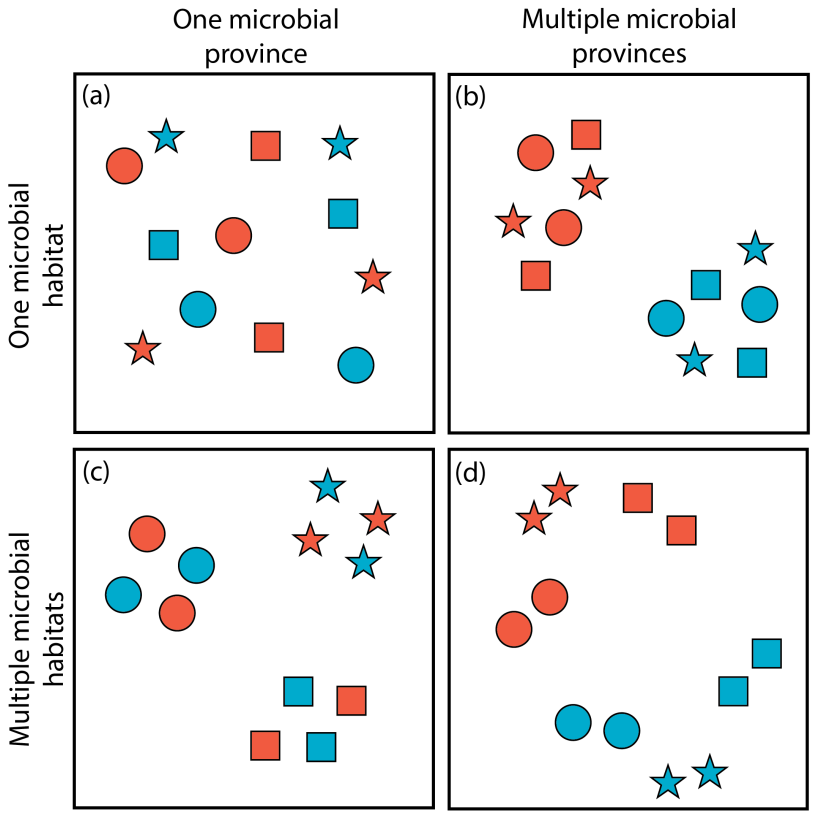

Figure SI.1: **?** propose four hypotheses partitioning the influence of different factors on microbial biogeography. Sampled communities, represented by shapes, are plotted on two-dimensional axes with similar communities clustered together. Different shapes indicate different habitat types that represent a combination of biotic and abiotic variables, while different colours represent different geographic locations. The null hypothesis **(a)** supposes that taxa are randomly distributed; there is a single habitat and a single province. Hypothesis **(b)** corresponds to a single habitat but multiple provinces, with all biogeographic structure due to historical events such as past isolation. Hypothesis **(c)** posits that 'everything is everywhere, but the environment selects' (**?**) - strong dispersal homogenises the species pool within a single provinces, but environmental differences between habitats maintain biogeographic patterns. In **(d)** microbial community assembly is influenced by stochastic events in geographically isolated provinces and selection based on current habitat types.

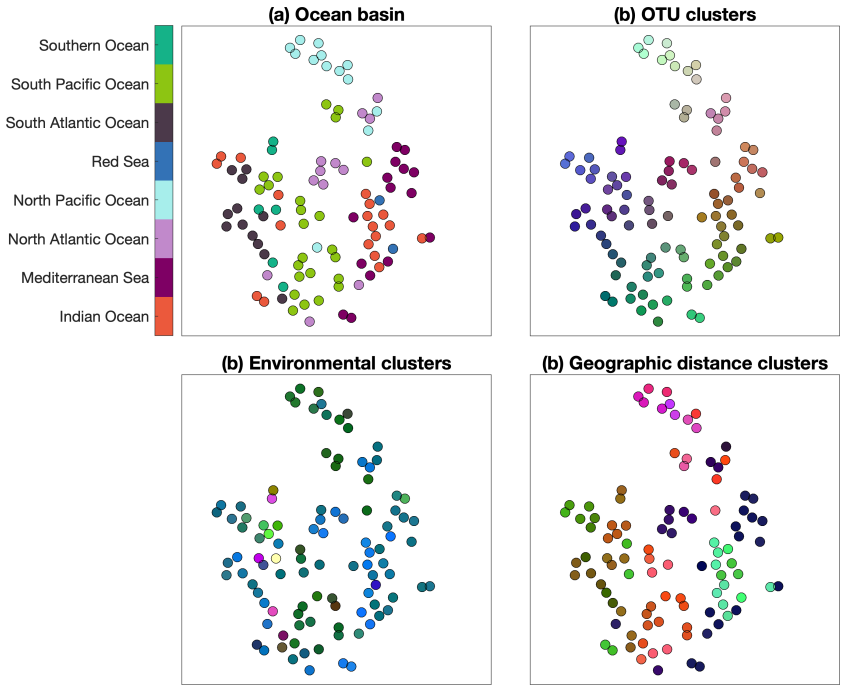
(a) Ocean basin
(b) OTU clusters
Southern Ocean
South Pacific Ocean
South Atlantic Ocean
Red Sea
North Pacific Ocean
North Atlantic Ocean
Mediterranean Sea
Indian Ocean
(b) Environmental clusters
(b) Geographic distance clusters

Figure SI.2

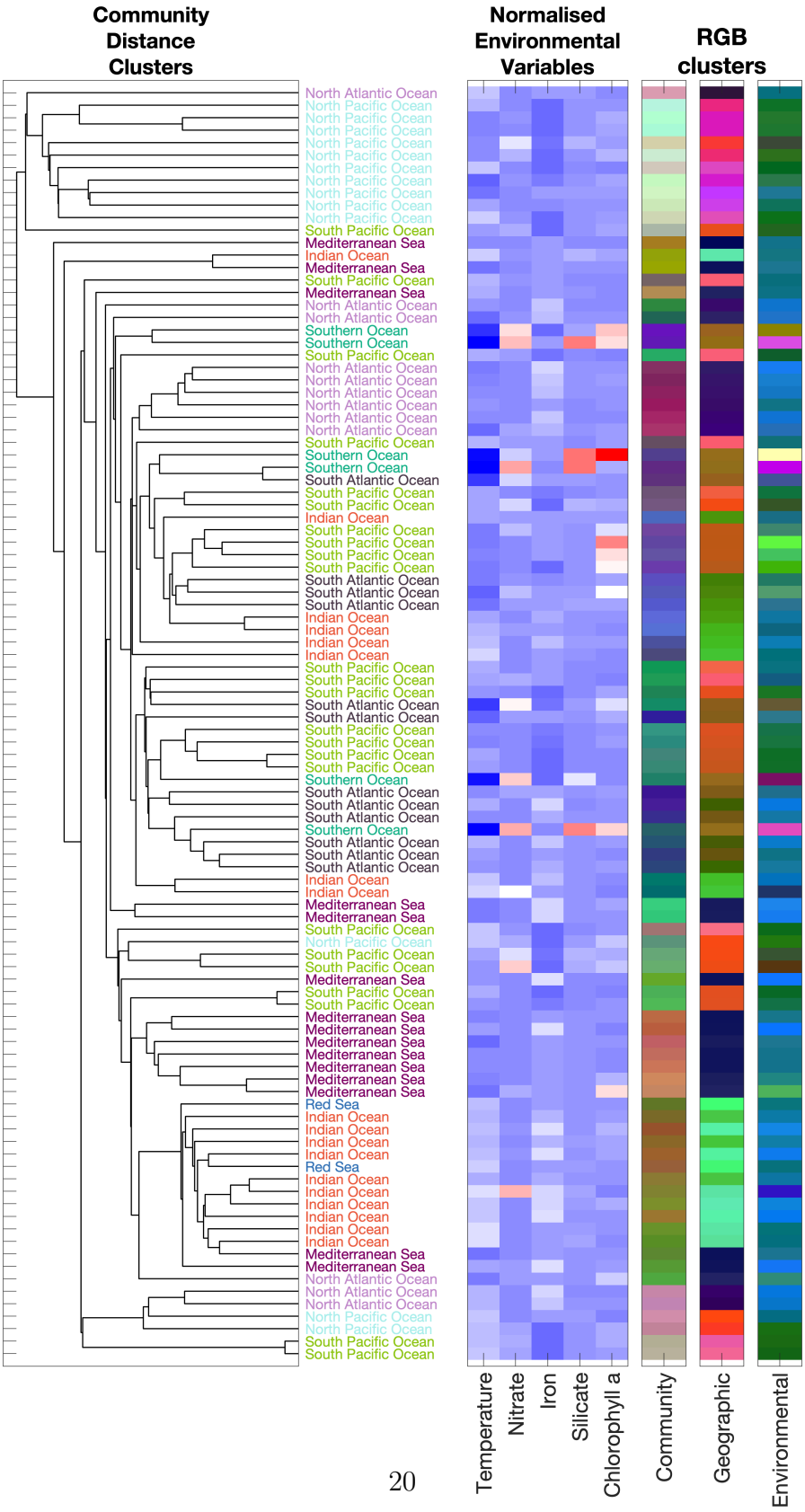
Community Distance Clusters
Normalised Environmental Variables
RGB clusters
North Atlantic Ocean
North Pacific Ocean
South Pacific Ocean
Mediterranean Sea
Indian Ocean
Southern Ocean
South Atlantic Ocean
Red Sea
Temperature
Nitrate
Iron
Silicate
Chlorophyll a
Community
Geographic
Environmental

Figure SI.3

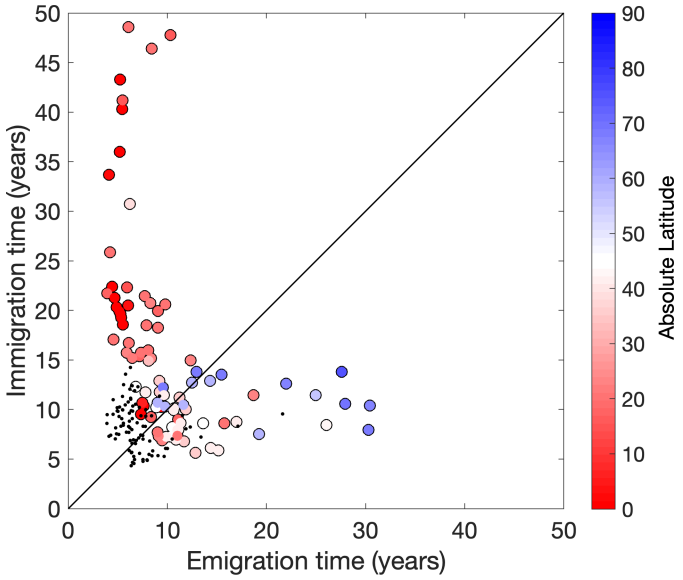

Figure SI.4: Immigration vs emigration times (years) at the 94 seed locations. The coloured circles show times from the surface-only case. The colour scale indicates absolute latitude, with low latitude regions clearly characterised by fast emigration and slow immigration, with the opposite true at higher latitudes. The black dots show the same time scales in the depth-integrated case.

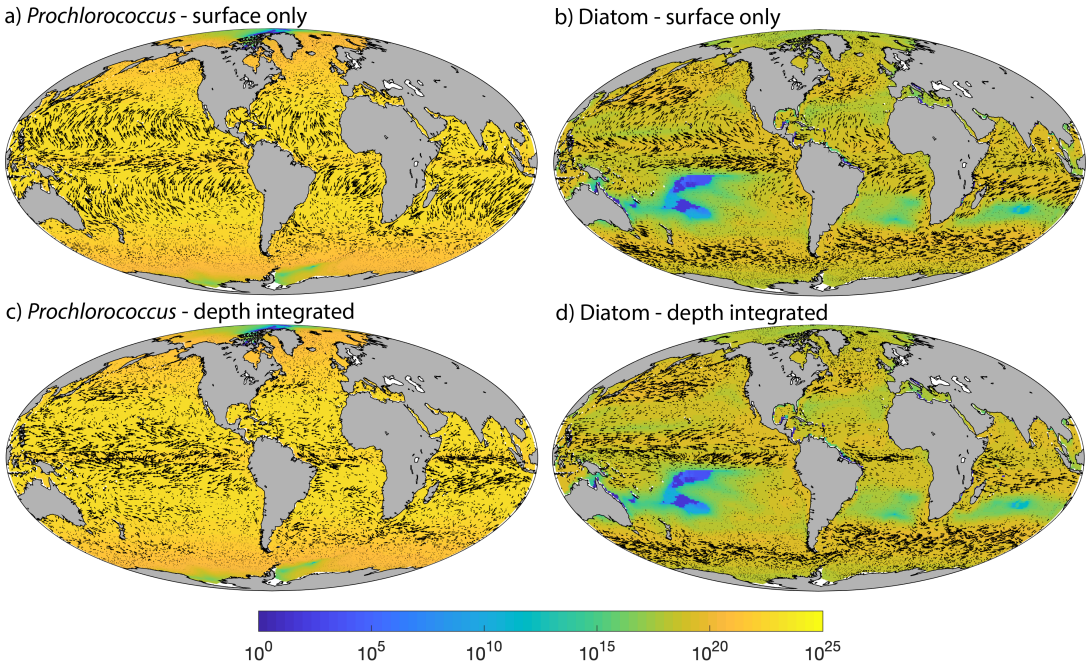

Figure SI.5: Abundances and cell transport vectors for *Prochlorococcus* (a and c) and diatoms (b and d). The upper panels show transport vectors through the surface layer. The lower panels show depth-integrated transport vectors. We used depth-integrated abundances in all cases.

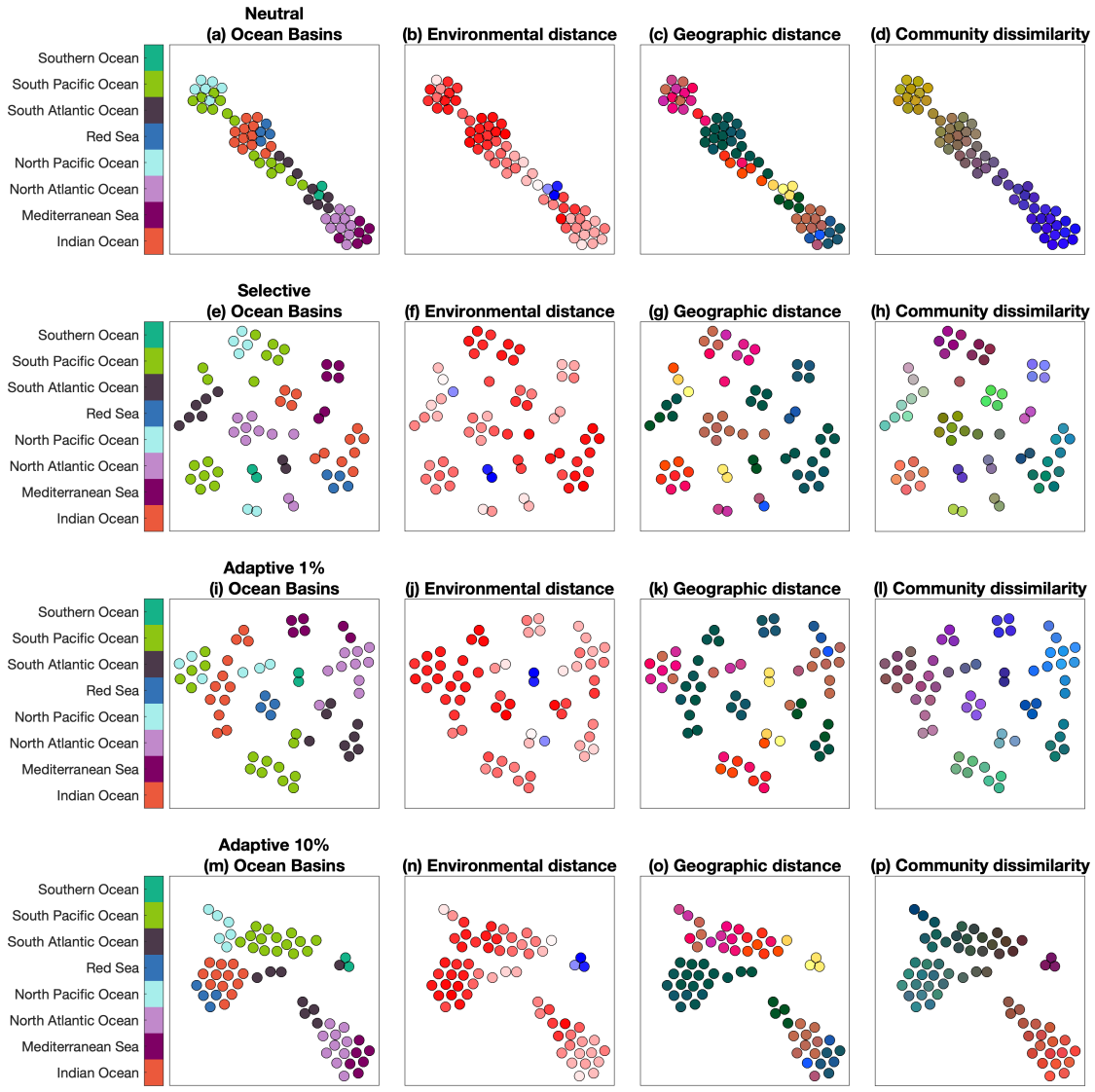

Figure SI.6: Community clustering in four simulations. For each experiment (rows), the simmilarity of model communities among sites is indicated by the spatial clustering, as derived using the t-SNE algorithm (**?**). In the first column, sample points are labelled with colours indicating their ocean basin. In the second column, the sample points are coloured according to their environmental similarity, with colder locations coloured blue, and warmer locations coloured red. The third column shows sites coloured by geographic distance, with geographically closer sites having similar colours. The final column shows sites coloured by the same community similarity used to derive their position on the two dimensional axes, but in this case the community similarity was projected into three dimensional space, which was then translated into colour by assigning the three dimensions to red, green and blue colour channels.