Ocean currents promote rare species diversity in protists

Paula Villa Martín,[1] Ales Bucek,[1] Tom Bourguignon,[1]Simone Pigolotti[1*]

[1] Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan,
*To whom correspondence should be addressed; E-mail: simone.pigolotti@oist.jp .

Oceans host communities of plankton composed of relatively few abundant species and many rare species. The number of rare protists species in these communities, as estimated in metagenomic studies, decays as a steep power law of their abundance. The ecological factors at the origin of this pattern remain elusive. We propose that oceanic currents affect biodiversity patterns of rare species. To test this hypothesis, we introduce a spatially-explicit coalescence model able to reconstruct the species diversity in a sample of water. Our model predicts, in the presence of oceanic currents, a steeper power law decay of the species abundance distribution and a steeper increase of the number of observed species with sample size. A comparison of two metagenomic studies of planktonic protist communities in oceans and in lakes quantitatively confirms our prediction. Our results support that oceanic currents positively impact the diversity of rare aquatic microbes.

Introduction

Oceanic plankton can be transported across very large distances by currents. Many planktonic species are cosmopolitan, i.e. they are found virtually everywhere across the global ocean (*1, 2*). These observations suggest that, at first sight, the distribution of planktonic species is not limited by dispersal, and therefore that niche preference is the predominant factor determining species abundance (*3*). However, niche theory predicts species-poor planktonic communities for plankton thriving on a limited set of resources.

Observations contradict this prediction and show that planktonic communities are very diverse (*4–7*). This violation of the basic principles of niche theory (*8, 9*) has puzzled ecologists for decades (*10*) and has fostered numerous studies attempting to explain the diversity of plankton (*11*). One proposal is that variable environments offer more possibilities for specialization of ecological traits (*5, 12–18*). Another proposal is that oceanic currents can create barriers reducing competition among species, and therefore promoting species coexistence (*19, 20*). Quantitative analyses also suggest that oceanic currents play a significant role in organizing large-scale diversity patterns (*21, 22*), and that dispersal limitation contributes, alongside with niche specialization, to microbial biodiversity of oceans (*23–27*).

DNA metabarcoding has allowed rapid and extensive measurements of the diversity of aquatic microbial communities, providing new means to study the ecological forces shaping planktonic communities. Metabarcoding studies have revealed that, beside commonly observed species, planktonic communities are characterized by a vast range of rare species. This so-called "rare biosphere" (*28, 29*) makes up the majority of planktonic species (*25, 30*). The diversity of planktonic species can be quantified by the Species Abundance Distribution (SAD), defined as the frequency $P(n)$ of species with abundance n in a sample. SADs of rare marine protists are qualitatively different from those of abundant species (*31, 32*) and appears to follow a power law distribution

$$P(n) \propto 1/n^{\alpha}. \tag{1}$$

The exponent α varies significantly among samples, appears weakly correlated with environmental factors, and is significantly larger than 1 on average (*33*). Diversity patterns in other microbial communities, such as that of the human gut (*34*), are well described by a form of SAD following the Fisher log series, $P(n) \propto e^{-cn}/n$ (*35–37*), as predicted by Hubbell's neutral model (*36, 38, 39*). For large communities, the parameter c is very small, so that the distribution is close to a power law with $\alpha = 1$. Hubbell's neutral model is therefore unable to explain the decay of SADs in oceanic protist communities. Steep SADs, in agreement with the data, can be obtained with a modified neutral model that takes into account density-dependence of growth and death rates (*33, 40–42*). However, the ecological forces determining

this density-dependence in the oceans are unknown.

In this paper, we propose that the steep decay of SADs observed in the oceans is caused by the particular way oceanic currents limit dispersal. Although several studies have shown that currents can affect effective population size (*43*) and provoke counterintuitive effects on fixation times (*44–46*), these studies did not scrutinize the effect of currents on multi-species communities. To test our hypothesis, we introduce a model that takes into account the role of oceanic currents in determining the genealogy of microbes in a sample. Our model predicts that, in the presence of oceanic currents, SADs are characterized by larger values of the exponent α. We also predicts that currents cause a sharper increase of species diversity as a function of sample size. To test these predictions, we analyze 18S rRNA sequencing data generated from oceanic (*33*) and lake protist communities (*47*). The observations quantitatively match our predictions, supporting the idea that oceanic currents are responsible of the differences in biodiversity patterns between oceans and lakes.

Results

Coalescence model predicts the effect of oceanic currents on SADs. We introduce a computational model to assess the effect of oceanic currents on the protist species distribution of a water sample. In this model, we assign a tracer with spatial coordinates to each individual in the sample (see Fig. 1). Tracers are initially placed in a local area, representing the portion of water where the sample was collected. The coordinates of each tracer move backward in time, following the spatial trajectory of the ancestors of each individual (see Fig. 1). If two tracers are at a sufficiently close distance, they coalesce into a single tracer with a given probability. This new tracer represents the most recent common ancestor of the two individuals. Finally, tracers are assigned at a fixed rate μ to one species. These events represent immigration due to other causes than ocean currents; assigned tracers are eliminated from the system. At the end of the run, individuals in the original sample are considered conspecific if they have coalesced to a common ancestor before being eliminated (see Fig. 1 and Methods). This coalescence model can be

interpreted as the backward version of an individual-based community model which includes the effect of fluid flows (*44, 48*) (see Supplementary Fig. 1). The coalescent formulation has the advantage of describing the dynamics of one sample embedded in a larger ecosystem (*49, 50*).

We simulate the coalescence model with and without oceanic currents. In the latter case, movements of tracers are modeled as a simple diffusion process, taking into account individual movements and small-scale turbulence. In the former case, we superimpose to this diffusion process the effect of large scale oceanic currents. We model oceanic currents with a kinematic model of a meandering jet, which is a ubiquitous large-scale structure characterizing oceanic flows (*51–54*). Population sizes and parameters characterizing the flow are sampled in a physically realistic range (*53, 54*) (see Methods). All other parameters characterizing population dynamics are chosen identically in the two cases (see Methods).

SADs predicted by the model display a considerable variability depending on parameters and demographic stochasticity, both in the presence and absence of currents (see Fig. 2a and 2b). To characterize individual SAD curves, we fit them with a power law function $P(n) \propto 1/n^{\alpha}$ using maximum likelihood in an optimal range of abundances (see Methods). For comparison, we also fit an exponential $P(n) \propto e^{-cn}$, and a Fisher log series $P(n) \propto e^{-cn}/n$ in the same range. The power law provides a better fit than the exponential in most cases (74% and 77% of samples with and without currents, respectively) and than the Fisher log series (75% and 62% of samples with and without currents, respectively).

Introducing oceanic currents in the model increases, on average, the steepness of SADs (see Fig. 2a and 2b). We investigate the physical mechanism causing this effect. One effect of transport by currents is to enhance the effective diffusivity (*55*). We test whether this effect is responsible for the change of pattern in the SAD by running our model with the effective diffusivity of the kinematic model, but without currents (see Supplementary Fig. 2). We find that the SADs are weakly affected by the change in diffusivity. This means that the change in SADs must be due to the structures created by the oceanic currents, which can not be simplified into a diffusion process. We further run our model with a parameter choice yielding currents constant in time (see Supplementary Fig. 3). Neither in this case we observe the

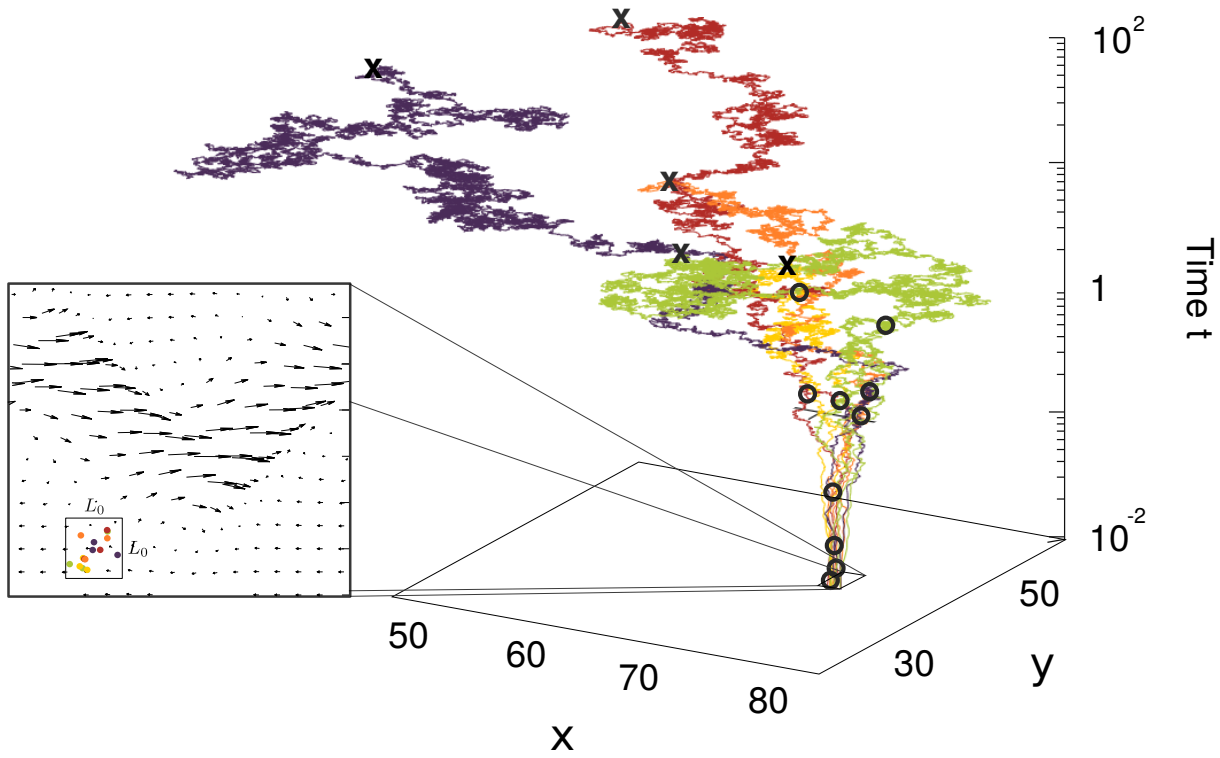

Figure 1: **Genealogy in oceanic currents.** (left) The coalescence model predicts the protist species composition of a sample of oceanic water of size $L_0 \times L_0$. Different colors represent different species. Arrows represent the velocity field caused by ocean currents. (right) Trajectories of the coalescence model with ocean currents. Individuals are represented by tracers, that are transported backward in time and can coalesce with other tracers if they reach a close distance. Coalescence events are marked by open circles; trajectories of individuals that have coalesced are shown in the same color. Tracers are removed from the population at an immigration rate μ (marked by crosses). See also Supplementary video.

steep SADs as in the presence of time-dependent currents. Taken together, these results suggest that the time-varying, chaotic nature of oceanic transport is responsible for the steepening of SAD curves.

Protist SADs are steeper in oceans than in freshwater. To test our predictions, we analyze DNA metabarcoding datasets from two studies of aquatic protists. The first dataset includes oceanic protist DNA sequences of 157 water samples from the TARA ocean expedition (*33*). The second dataset includes protist DNA sequences of 184 freshwater samples taken from lakes (*47*). In both cases, we use 97% sequence identity threshold to cluster protist sequences into operational taxonomic units (OTUs). We calculate SAD for each sample of both datasets using OTUs as proxies for species (see Methods).

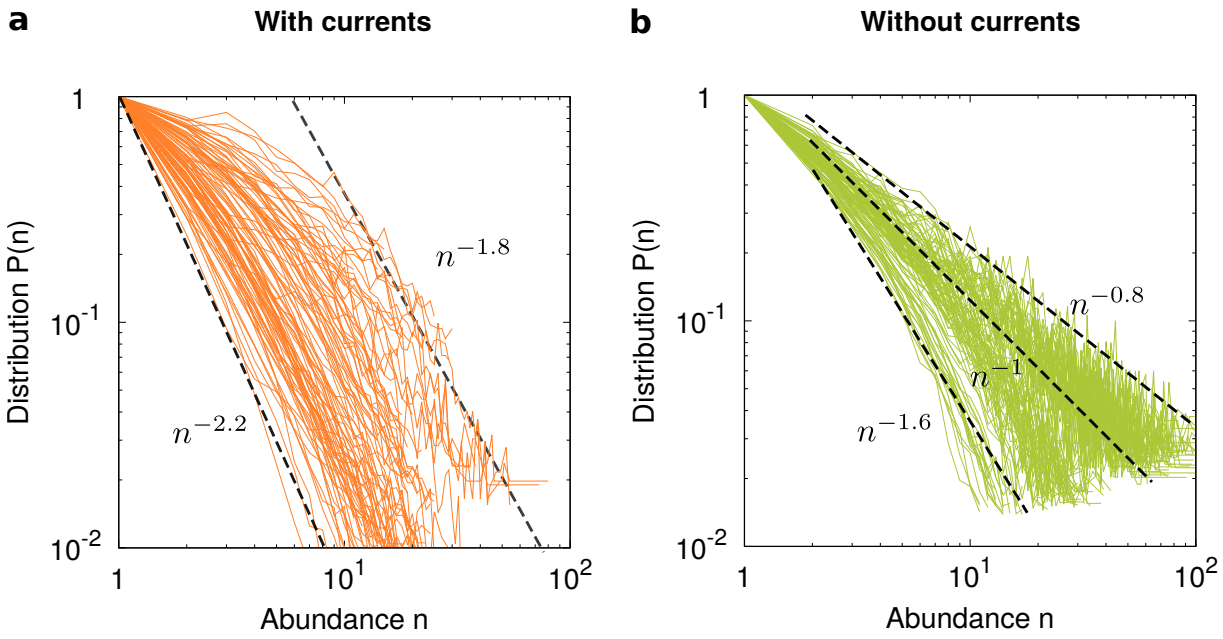

Figure 2: **Coalescence model predicts effect of currents on species abundance distribution.** We show SADs A) in presence (orange lines) and B) absence (green lines) of oceanic currents for the coalescence model. Here and in the following, SAD curves are rescaled so that $P(1) = 1$ to ease visualization. Model details and parameters are presented in Methods. Dashed lines are power laws to guide the eye (see also Fig. 3).

From now on we discard "abundant species", defined as those in abundance classes $P(n)$ including less than 4 species. The remaining "rare species" are the subject of our study and constitute 93% of all species in ocean samples and 78% of all species in lake samples.

As for the model, SAD curves display considerable sample-to-sample variability, both in ocean and in freshwater samples (see Fig. 3). The variability is possibly caused by differences in size, protist community composition, and ecological conditions among samples. Observed SAD curves are better fitted by a power law than exponential or Fisher log series in most cases. The exponential distribution provides a better fit than the power law in 13% of lake samples and 13% of oceanic samples, whereas the Fisher log series provides better fits than the power law in 39% of lake samples and 18% of oceanic samples. We obtain similar results with different OTU definition (95% instead of 97% similarity) and different thresholds separating abundant from rare species (see Supplementary Table 1 and 2).

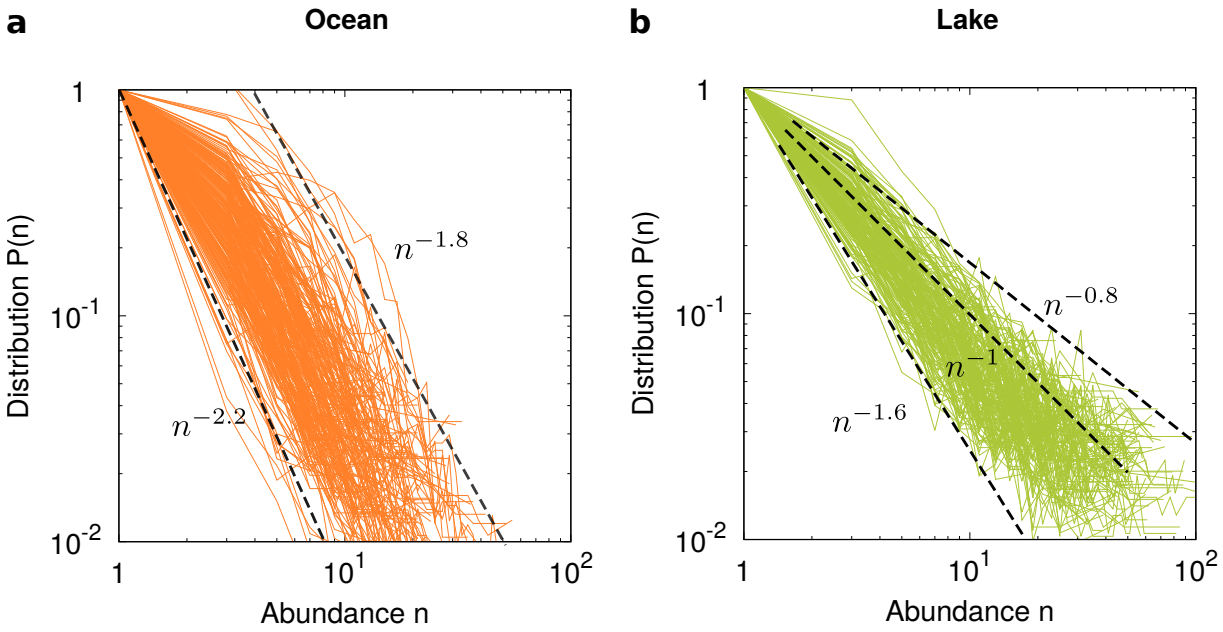

Figure 3: **Rare species abundance distributions present a steeper decay with abundance in oceans than in lakes.** Continuous lines represent SADs of protist communities from (a) 157 oceanic (*33*) and (b) 184 freshwater (*47*) samples. Total numbers of individuals in each sample are in the ranges of (a) $(10^3, 10^5)$ and (b) $(10^4, 10^6)$. In both panels, power laws (dashed lines) are shown to guide the eye.

Strikingly, the power law decay of SADs is on average steeper in oceans than in lakes (see Fig. 3), as predicted by our coalescence model.

Distribution of the SAD exponent is quantitatively predicted by the coalescence model. We quantify the agreement between model and data by analyzing the distribution of the power law exponent α in equation (1). In the presence of currents, the model predicts a value of the exponent significantly larger than one (average $\alpha = 1.70$, standard deviation $\sigma = 0.68$). In the absence of oceanic flows, the model predicts an average $\alpha = 1.26$, ($\sigma = 0.46$), a value compatible with the neutral prediction $\alpha = 1$ in well-mixed systems (*35–37*) and spatially-explicit neutral models (*50, 56*).

Observations in both oceans and lakes are in excellent agreement with the distributions of exponents predicted by the model (see Fig. 4a). Our analysis confirms that the average exponent α is significantly larger than 1 in the oceans (average $\alpha = 1.79$, $\sigma = 0.52$, see Fig. 4a and (*33*)). In the lakes, the average exponent is $\alpha = 1.15$ ($\sigma = 0.36$, see Fig. 4a). Adopting a different definition of OTUs (95% instead

of 97%) and/or different thresholds separating abundant from rare species leads to qualitatively similar result (see Supplementary Tables 1 and 2 and Supplementary Fig. 4). In particular, the average exponent α in the oceans is between 8% and 43% larger than in the lakes, depending on the threshold and the definition of OTUs.

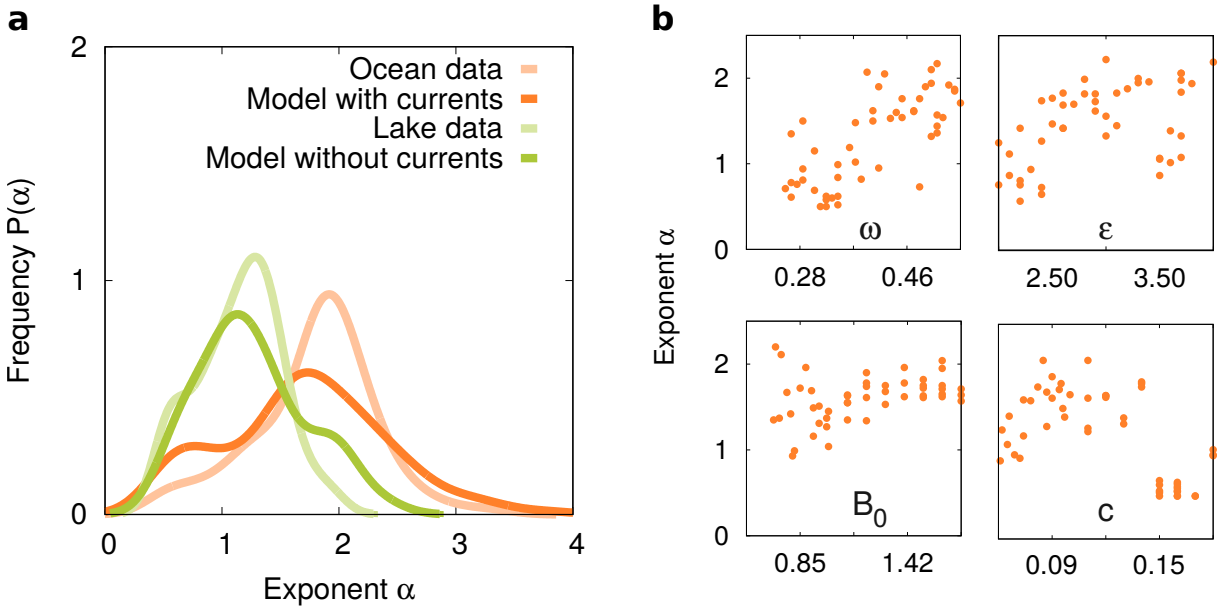

Figure 4: **Power law exponents of species abundance distributions in the global ocean.** We run our models for different population sizes and different values of flux parameters for ocean samples (see Methods). We select 157 oceanic samples and 184 freshwater samples as in Fig. 3. We fit the power law exponent α of the SADs to the model and to the data using maximum likelihood (*57–59*). (a) Continuous distributions of the exponent obtained by kernel density estimation (*60, 61*). (b) Dependence of the exponent on four main parameters of the oceanic flow. In each sub-panel, other parameters are kept constant (see Methods).

We find that four parameters characterizing the shape and the mixing level of the jet mostly affect α. However, the value of the exponent does not present a regular trend upon varying each parameter individually (see Fig. 4b and Supplementary Fig. 5).

Ocean currents lead to a steeper increase in number of species as a function of sample size. By simulating our model at varying sample size N with and without currents, we predict that currents should significantly increase the number of expected species in each sample (see Fig. 5a). This effect

is consistent with the increase of α in the presence of currents: increasing α suppresses very abundant species, and therefore makes the samples more diverse. This effect becomes more and more pronounced as N is increased. In the data, we find that samples from oceans contain more species than samples from lakes at similar sample size, consistently with our prediction (see Fig. 5a). The observed enrichment is even stronger than predicted by our model.

We now study the increase of number of species with sample size in oceanic and lake water samples individually. In the case of well-mixed populations, the species composition of a given sample is described by the Ewens sampling formula (*62*), which predicts that the expected number of species in the sample is

$$S = \sum_{j=0}^{N} \frac{\theta}{\theta + j - 1}. \tag{2}$$

where $\theta = 2N_{\text{eff}}\mu$ is the fundamental biodiversity number (*36*) and N_{eff} is the effective population size. Alternatively, sample species composition can be empirically described using a power law (*63, 64*)

$$S \propto N^{z}. \tag{3}$$

Our model predicts an increase in the number of species with the sample size largely as predicted by the Ewens sampling formula (see Fig. 5a and Fig. 5b). Both for ocean and freshwater samples, the power law model provides a better fit (see Fig. 5a and Fig. 5b) with a higher exponent for oceanic samples ($z = 0.73$) compared to lake samples ($z = 0.65$).

Discussion

Oceanic currents are known to largely affect plankton distribution at large scale (*19–21*). Here, we show that ocean currents have profound effects on plankton distribution and diversity of rare protist species even at the level of individual metagenomic samples. Our coalescence model bridges the gap between large-scale oceanic dynamics and ecological dynamics at the individual level and provides a versatile and

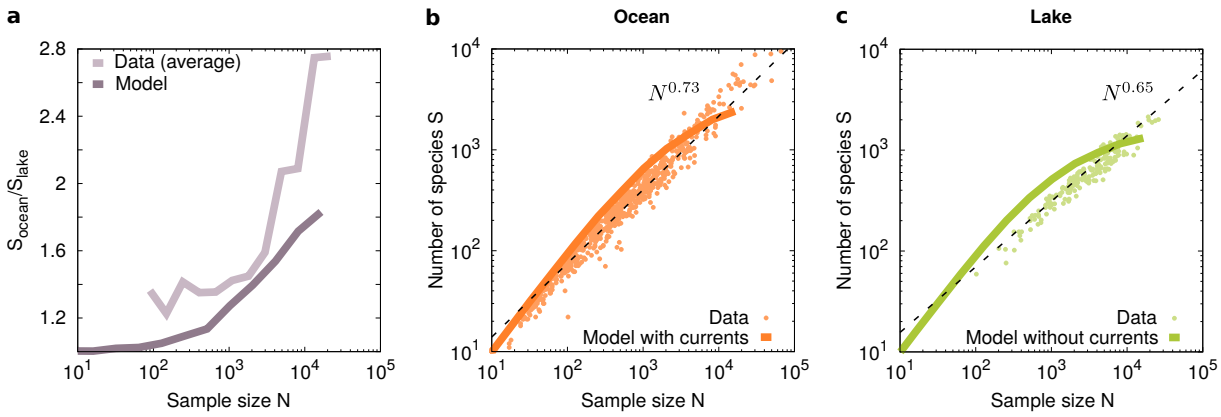

Figure 5: **Oceanic currents increase species number S in a water sample.** (a) Ratio $S_{\mathrm{ocean}}/S_{\mathrm{lake}}$ as a function of the sample size N for the model and the data. We simulate the model at increasing sample sizes N in powers of 2 and obtain continuous curves by interpolation. Other parameters are presented in Methods. Averaged data are obtained by binning for both oceans and lakes. (b,c) Number of species S in samples of N individuals in (b) oceans and (c) lakes. A power law (equation 3) fits the data better than Ewens sampling formula (equation 2) for both (b) oceanic (normalized log likelihood -19.39 versus -449.52) and (c) lake samples (normalized log likelihood -12.46 vs -111.94). Fitted exponents are $z = 0.73$ and $z = 0.65$ for oceans and lakes, respectively. The results of the coalescence model are shown with and without oceanic currents (orange and green line, respectively). Ewens sampling formula provides a better fit than the power law in both cases, (b) -420.95 vs -2131.43 and (c) -434.66 vs -1902.62.

powerful tool to validate individual-based ecological models using DNA metabarcoding data. Although, for simplicity, we focus on neutral dynamics and protists, our approach can be extended to more general ecological settings and to other plankton communities, including animals and prokaryotes. Such generalizations, combined with high-throughput sequencing data, has the potential to shed light on the main ecological forces determining plankton dynamics.

The coalescence model predicts that oceanic currents are responsible for steeper decay of SAD curves and steeper increase in the number of observed species with sample size. Both these predictions are in quantitative agreement with observations. The steep decay of SAD distributions in the oceans has been previously explained in terms of density-dependent effects (*33*). Although our study does not preclude this possibility, the comparison with freshwater ecosystems strongly suggests that oceanic currents effectively determine this density dependence. The steeper decay of SAD curves predicted by the coalescence

model depends on geophysical parameters characterizing mixing of oceanic currents. The irregular behavior of the SAD exponent as a function of these parameters (see Fig. 4b) potentially explains the difficulty of correlating observed values of α with other oceanographic measurements (*33*).

Oceanic flows, such as those considered here, act as barriers limiting transport. Our results support that these barriers reduces the pace of individual coalescence into species, and therefore limits the formation of abundant species. The comparison with a non-chaotic flow and with purely diffusive dynamics supports that the chaotic nature of oceanic flows is the main cause of the steep SAD exponents. A detailed physical theory of this phenomenon remains a challenge for future studies.

In summary, our study provides a mechanistic theoretical framework to analyze diversity of rare microbial species in aquatic environments at the individual level and paves a way to quantitatively understand how dispersal limitation by oceanic currents shapes the diversity of planktonic communities.

Methods

Coalescence model. We consider N microbial individuals in an aquatic environment and seek to reconstruct their ancestry. Each individual is associated to a tracer with two-dimensional spatial coordinates x, y. Initially, tracers are homogeneously distributed in a square $L_0 \times L_0$, representing the area where the sample was collected. The tracers move in space according to the stochastic differential equations

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}x &= v_x(x,y,t) + \sqrt{2D}\xi_x(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}y &= v_y(x,y,t) + \sqrt{2D}\xi_y(t)\,, \end{aligned} \tag{4}$$

where v_x, v_y is an advecting field representing the effect of oceanic currents, and the terms proportional to $\sqrt{2D}$ are diffusion terms modeling individual movement and small-scale turbulence. The quantities $\xi_x(t), \xi_y(t)$ are independent white noise sources satisfying $\langle \xi_i(t) \rangle = 0$, $\langle \xi_i(t)\xi_j(t') \rangle = \delta_{ij}\delta(t-t')$ where $\langle \dots \rangle$ denotes an average and $i, j \in (x, y)$. The advecting field v_x, v_y is discussed in the next

subsection. Since the coalescence model evolves backward in time, we integrate equation (4) from the final configuration with negative time increments.

Tracers at a short distance δ from each other can coalesce at a rate r. We implement immigration events by assigning species at a rate μ. At each time step $\mathrm{d}t$:

1. All tracers move from (x, y) to $(x + \Delta x, y + \Delta y)$ by numerically integrating equation (4).

2. Tracers are selected one by one and are removed with probability $\mu\,\mathrm{d}t$ (immigration event). Tracers coalesce with probability $r\,\mathrm{d}t$ when an individual j is in a area of size $\delta \times \delta$ centered at the selected individual i.

We set $r = 1$, $\mu = 10^{-4}$, and the diffusion constant to $D = 3 \cdot 10^{-9}$ as further discussed below. The interaction distance δ is chosen to satisfy $D = r\delta^2$, see (*48*). We consider sample areas of linear size on the order of the mean distance traveled by an individual in one generation, $L_0 = 5\ \mathrm{Km}$, estimating a protist life time of about one day (*65*), and protist movements of about $20\ \mathrm{Km}^2$ per day (*66*). Population size is randomly selected for each run in the range $N \in (10^3, 10^5)$ unless otherwise indicated. For Fig. 4b and Supplementary video we set $N = 8192$.

Kinematic model of the oceans. We model large-scale oceanic currents by means of a kinematic model of a meandering jet (*51–54*). The velocity field v_x, v_y is defined in terms of a stream function. In a fixed reference frame, the stream function reads

$$\psi(x', y', t') = \psi_0 \left\{ 1 - \tanh\left(\frac{y' - A\cos[\kappa(x' - c_x t')]}{\lambda\sqrt{1 + \kappa^2 A^2 \sin^2[\kappa(x' - c_x t)]}} \right) \right\}. \tag{5}$$

The stream function is more conveniently written in a dimensionless form

$$\phi(x, y, t) = -\tanh\left(\frac{y - B(t)\cos(kx)}{\sqrt{1 + k^2 B^2 \sin^2(kx)}} \right) + cy\ , \tag{6}$$

being $B(t) = B_0 + \epsilon\cos(\omega t + \Phi)$, $c = c_x L/\psi_0$ and $k = 2\pi/L$ with L the meander wave-length. The transformation between dimensional and dimensionless units is $x = x'/\lambda$, $y = y'/\lambda$ and $t = t'\psi_0/\lambda^2$ (*52*). Given the stream function, the components of the velocity field in dimensionless units are

$$\begin{aligned} v_x &= -\partial\phi/\partial y \\ v_y &= \partial\phi/\partial x. \end{aligned} \tag{7}$$

We run the simulations in a virtually infinite system. In the case without currents, this modeling choice is justified a posteriori by the fact that, based on our observations, lake SAD exponents do not present a significant dependence on lake area (see Supplementary Fig. 6). For the ocean simulations, results can be affected by the position of the local area. For this reason, the sample area $L_0 \times L_0$ is placed at random coordinates $x_0, y_0 \in (0, 8)$ for each run. For Fig. 5 we fix $x_0 = 7.5$ and $y_0 = 1$.

Parameters of the kinematic model. Realistic parameters of the dimensionless stream function, equation (6), are estimated as $L = 7.5, c = 0.12, B_0 = 1.2, \omega = 0.4, \epsilon = 0.3$ and , $\Phi = \pi/2$ (*53, 54*). We consider parameter ranges based on these values $c \in (0.06, 0.18), B_0 \in (0.7, 1.7), \omega \in (0.25, 0.55)$ and fix $L = 7.5, \Phi = \pi/2$. The value of ϵ has to be larger than a critical value depending on ω to prevent transported particles to remain trapped into long-lived eddies (*54*). To meet this condition while exploring a range of values of ω, we fix $\epsilon \in (2, 4)$. For Fig. 4b, Fig. 5 we set $c = 0.12, B_0 = 1.2, \omega = 0.5$, and $\epsilon = 3$.

To convert from dimensionless units to dimensional units, we use the spatial scale $\lambda = 40\,\mathrm{Km}$ (*51*) and the stream function scale $\psi_0 = 160\ \mathrm{Km}^2\ \mathrm{day}^{-1}$. With this choice, the time unit λ^2/ψ_0 is equal to one day. The parameter ψ_0/λ represents the maximum velocity in the center of the jet. With our choice of units, the velocity is equal to $40\ \mathrm{Km\ day}^{-1}$, slightly lower than the average velocity of large-scale oceanic currents (about $\psi_0/\lambda \approx 200\ \mathrm{Km\ day}^{-1}$ for the surface Gulf stream and $50\ \mathrm{Km\ day}^{-1}$ for the lower thermocline (*51*)).

In physical units, the coalescence rate is equal to $r = 1\,\mathrm{day}^{-1}$, i.e. about one generation time for protists (*65*). Our choice of the diffusion constant to $D = 3 \cdot 10^{-9}$ in dimensionless units corresponds to about $6 \cdot 10^{-5} m^2/s$ in physical units, which is consistent with observations (*55*).

Species abundance distribution (SAD). We compute the distribution $P(n)$ of the species abundances n for each sample. Species with low to intermediate abundance appear to follow different distribution than abundant species, as previously observed (*31–33*). For this reason, we filter out species in abundance classes below $P(n) = 4$. To avoid overfitting, we also discard samples with SAD composed of less than 10 points with different frequencies $P(n)$. After this selection, we are left with 157 samples for oceans and 184 for lakes. We compute the species abundance distribution $P(n)$ for the coalescent model with and without advection. Each sample is obtained for different flux parameters and population sizes (described above). The resulting distributions $P(n)$ are averaged over up to 10^2 realizations of the model and filtered in the same way as the data samples for consistency.

Data fits. To determine the exponent α, we fit the function

$$P(n) = C/n^{\alpha} \tag{8}$$

in a range of intermediate abundances $(n_{\min}, n_{\max})$. The exponent α, the proportionality constant C, and the values of $n_{\min}$ and $n_{\max}$ are simultaneously determined by maximizing the normalized log-likelihood $\ln L = (1/\mathcal{N}) \sum_i [n_i \ln P(n_i) - P(n_i) + \ln(n_i!)]$, where $\mathcal{N}$ is the number of non-zero abundance classes for n in the range $(n_{\min}, n_{\max})$ and we assumed Poissonian counts (*57–59*). We discard samples for which the range $(n_{\min}, n_{\max})$ includes less than 5 points. We also fit an exponential $P(n) = Ce^{-cn}$ and a Fisher log series $P(n) = Ce^{-cn}/n$ with the same method and in the same interval $[n_{\min}, n_{\max}]$ determined with the power law fit. Since all the distributions have the same number of free parameters, we always consider a better fit the distribution characterized by the largest normalized log-likelihood. The percentage of data samples for which a power law fits better than the exponential and Fisher log series and the corresponding exponents are presented in Supplementary Tables 1 and 2.

OTU Analysis. We analyze metabarcoding data from marine (*33*) and freshwater (*47*) protist planktonic communities.

We retrieve the dataset of oceanic samples from European Nucleotide Archive (accession id PRJEB16766). The dataset consists of assembled paired-end Illumina HiSeq2000 sequencing reads of PCR-

amplified V9 loop of protist 18S rRNA gene obtained from 121 seawater locations distributed worldwide. In the first analytical procedure, we trim the primer sites using USEARCH (v.11.0.667) (*67*). Primer sites include 15 and 20 nucleotide sites for the 5-end and 3-end, respectively. The trimmed sequences are quality filtered with USEARCH using the option –fastq_maxee 1.0, which discards sequences with > 1 total expected errors in the sequence. The sequences are dereplicated and singleton sequences (i.e. sequences with single occurence) removed using VSEARCH (v.2.10.1) (*68*). Chimeric sequences are detected and removed using UCHIME (implemented in VSEARCH) (*69*) and a combination of reference-based (with non-redundant SILVA SSU Ref database ver.132 used as reference) and de-novo methods. Sequences are then clustered into operational taxonomic units (OTUs) using VSEARCH and the –cluster_size option. We use sequence identity thresholds of 95% and 97% which provides different levels of taxonomic resolution.

We obtain the freshwater dataset, consisting of paired-end Illumina HiSeq2500 reads of amplified genomic region encompassing V9 loop of 18S rRNA gene and ITS1 gene for 217 European freshwater lakes, from Short Read Archive (Bioproject ID PRJNA414052). First, reads from PCR replicates and sequencing replicates are merged for each lake sample. Next, primer regions are trimmed with CUTADAPT (*70*), discarding reads missing one or both of the primer sites. Forward and reverse reads with minimal overlap of 70 base pairs and with maximum of 5 nucleotide differences in the overlapping region are merged with VSEARCH (command –fastq_mergepairs). Next, we extract from the amplified SSU V9 + ITS1 region the SSU V9 region using ITSx (v.1.1.1) (*71*). This step allows the taxonomic resolution of the clustered freshwater planktonic community OTUs to closely resemble the taxonomic community resolution of the marine planktonic community, which is based on sequenced V9 loop regions of 18S rRNA genes. The extracted reads are quality-filtered, dereplicated, singletons and chimeras are removed, and the quality-filtered reads are clustered into OTUs as described above for the marine dataset. The taxonomy is assigned against SILVA v123 eukaryotic 18S subset database. OTUs assigned to Fungi, Metazoa, or Embryophyta (i.e. non-protist eukaryotes) with at least BS > 0.8 support are

excluded from the final OTU tables.

References

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

2. T. Fenchel, B. J. Finlay, The ubiquity of small species: patterns of local and global diversity. *AIBS Bulletin* **54**, 777–784 (2004).

3. L. B. Becking, *Geobiologie of inleiding tot de milieukunde*, no. 18-19 (WP Van Stockum & Zoon, 1934).

4. J. A. Fuhrman, Microbial community structure and its functional implications. *Nature* **459**, 193 (2009).

5. M. Stomp, J. Huisman, G. G. Mittelbach, E. Litchman, C. A. Klausmeier, Large-scale biodiversity patterns in freshwater phytoplankton. *Ecology* **92**, 2096–2107 (2011).

6. S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, *et al.*, Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

7. C. De Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, *et al.*, Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).

8. G. Hardin, The competitive exclusion principle. *science* **131**, 1292–1297 (1960).

9. S. A. Levin, Community equilibria and stability, and an extension of the competitive exclusion principle. *The American Naturalist* **104**, 413–423 (1970).

10. G. E. Hutchinson, The paradox of the plankton. *The American Naturalist* **95**, 137–145 (1961).

11. M. Scheffer, S. Rinaldi, J. Huisman, F. J. Weissing, Why plankton communities have no equilibrium: solutions to the paradox. *Hydrobiologia* **491**, 9–18 (2003).

12. E. Litchman, C. A. Klausmeier, O. M. Schofield, P. G. Falkowski, The role of functional traits and trade-offs in structuring phytoplankton communities: scaling from cellular to ecosystem level. *Ecology letters* **10**, 1170–1181 (2007).

13. E. Litchman, C. A. Klausmeier, Trait-based community ecology of phytoplankton. *Annual review of ecology, evolution, and systematics* **39**, 615–639 (2008).

14. E. Litchman, P. de Tezanos Pinto, C. A. Klausmeier, M. K. Thomas, K. Yoshiyama, *Fifty years after the "Homage to Santa Rosalia": Old and new paradigms on biodiversity in aquatic ecosystems* (Springer, 2010), pp. 15–28.

15. J. A. Bonachela, M. Raghib, S. A. Levin, Dynamic model of flexible phytoplankton nutrient uptake. *Proceedings of the National Academy of Sciences* **108**, 20633–20638 (2011).

16. C. T. Kremer, C. A. Klausmeier, Coexistence in a variable environment: eco-evolutionary perspectives. *Journal of theoretical biology* **339**, 14–25 (2013).

17. J. Bonachela, S. Allison, A. Martiny, S. A. Levin, A model for variable phytoplankton stoichiometry based on cell protein regulation. *Biogeosciences* p. 4341–4356 (2013).

18. J. A. Bonachela, C. A. Klausmeier, K. F. Edwards, E. Litchman, S. A. Levin, The role of phytoplankton diversity in the emergent oceanic stoichiometry. *Journal of Plankton Research* **38**, 1021–1035 (2015).

19. A. Bracco, A. Provenzale, I. Scheuring, Mesoscale vortices and the paradox of the plankton. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **267**, 1795–1800 (2000).

20. G. Károlyi, Á. Péntek, I. Scheuring, T. Tél, Z. Toroczkai, Chaotic flow: the physics of species coexistence. *Proceedings of the National Academy of Sciences* **97**, 13661–13665 (2000).

21. F. d'Ovidio, S. De Monte, S. Alvain, Y. Dandonneau, M. Lévy, Fluid dynamical niches of phytoplankton types. *Proceedings of the National Academy of Sciences* **107**, 18366–18370 (2010).

22. D. J. McGillicuddy Jr, Mechanisms of physical-biological-biogeochemical interaction at the oceanic mesoscale. *Annual Review of Marine Science* **8**, 125–159 (2016).

23. J. B. H. Martiny, B. J. Bohannan, J. H. Brown, R. K. Colwell, J. A. Fuhrman, J. L. Green, M. C. Horner-Devine, M. Kane, J. A. Krumins, C. R. Kuske, *et al.*, Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology* **4**, 102 (2006).

24. M. C. Urban, M. A. Leibold, P. Amarasekare, L. De Meester, R. Gomulkiewicz, M. E. Hochberg, C. A. Klausmeier, N. Loeuille, C. De Mazancourt, J. Norberg, *et al.*, The evolutionary ecology of metacommunities. *Trends in ecology & evolution* **23**, 311–317 (2008).

25. P. E. Galand, E. O. Casamayor, D. L. Kirchman, C. Lovejoy, Ecology of the rare microbial biosphere of the arctic ocean. *Proceedings of the National Academy of Sciences* **106**, 22427–22432 (2009).

26. G. Chust, X. Irigoien, J. Chave, R. P. Harris, Latitudinal phytoplankton distribution and the neutral theory of biodiversity. *Global Ecology and Biogeography* **22**, 531–543 (2013).

27. D. Wilkins, E. Van Sebille, S. R. Rintoul, F. M. Lauro, R. Cavicchioli, Advection shapes southern ocean microbial assemblages independent of distance and environment effects. *Nature communications* **4**, 2457 (2013).

28. M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, G. J. Herndl, Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* **103**, 12115–12120 (2006).

29. C. Pedrós-Alió, The rare bacterial biosphere. *Annual review of marine science* **4**, 449–466 (2012).

30. M. D. Lynch, J. D. Neufeld, Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology* **13**, 217 (2015).

31. A. E. Magurran, P. A. Henderson, Explaining the excess of rare species in natural species abundance distributions. *Nature* **422**, 714 (2003).

32. W. Ulrich, M. Ollik, Frequent and occasional species and the shape of relative-abundance distributions. *Diversity and distributions* **10**, 263–269 (2004).

33. E. Ser-Giacomi, L. Zinger, S. Malviya, C. De Vargas, E. Karsenti, C. Bowler, S. De Monte, Ubiquitous abundance distribution of non-dominant plankton across the global ocean. *Nature ecology & evolution* p. 1 (2018).

34. P. Jeraldo, M. Sipos, N. Chia, J. M. Brulc, A. S. Dhillon, M. E. Konkel, C. L. Larson, K. E. Nelson, A. Qu, L. B. Schook, *et al.*, Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proceedings of the National Academy of Sciences* **109**, 9692–9698 (2012).

35. R. A. Fisher, A. S. Corbet, C. B. Williams, The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* pp. 42–58 (1943).

36. S. P. Hubbell, *The unified neutral theory of biodiversity and biogeography (MPB-32)* (Princeton University Press, 2001).

37. I. Volkov, J. R. Banavar, S. P. Hubbell, A. Maritan, Neutral theory and relative species abundance in ecology. *Nature* **424**, 1035 (2003).

38. D. Alonso, R. S. Etienne, A. J. McKane, The merits of neutral theory. *Trends in ecology & evolution* **21**, 451–457 (2006).

39. J. Rosindell, S. P. Hubbell, R. S. Etienne, The unified neutral theory of biodiversity and biogeography at age ten. *Trends in ecology & evolution* **26**, 340–348 (2011).

40. S. Pigolotti, A. Flammini, A. Maritan, Stochastic model for the species abundance problem in an ecological community. *Physical Review E* **70**, 011916 (2004).

41. F. He, Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Functional Ecology* **19**, 187–193 (2005).

42. S. Azaele, S. Pigolotti, J. R. Banavar, A. Maritan, Dynamical evolution of ecosystems. *Nature* **444**, 926 (2006).

43. J. P. Wares, J. M. Pringle, Drift by drift: effective population size is limited by advection. *BMC Evolutionary Biology* **8**, 235 (2008).

44. S. Pigolotti, R. Benzi, M. H. Jensen, D. R. Nelson, Population genetics in compressible flows. *Physical review letters* **108**, 128102 (2012).

45. F. Herrerías-Azcué, V. Pérez-Muñuzuri, T. Galla, Stirring does not make populations well mixed. *Scientific reports* **8**, 4068 (2018).

46. A. Plummer, R. Benzi, D. R. Nelson, F. Toschi, Fixation probabilities in weakly compressible fluid flows. *Proceedings of the National Academy of Sciences* **116**, 373–378 (2019).

47. J. Boenigk, S. Wodniok, C. Bock, D. Beisser, C. Hempel, L. Grossmann, A. Lange, M. Jensen, Geographic distance and mountain ranges structure freshwater protist communities on a european scale. *Metabarcoding and Metagenomics* **2**, e21519 (2018).

48. S. Pigolotti, R. Benzi, P. Perlekar, M. H. Jensen, F. Toschi, D. R. Nelson, Growth, competition and cooperation in spatial population genetics. *Theoretical population biology* **84**, 72–86 (2013).

49. J. Rosindell, Y. Wong, R. S. Etienne, A coalescence approach to spatial neutral ecology. *Ecological Informatics* **3**, 259–271 (2008).

50. S. Pigolotti, M. Cencini, D. Molina, M. A. Muñoz, Stochastic spatial models in ecology: a statistical physics approach. *Journal of Statistical Physics* **172**, 44–73 (2018).

51. A. S. Bower, A simple kinematic mechanism for mixing fluid parcels across a meandering jet. *Journal of Physical Oceanography* **21**, 173–180 (1991).

52. R. Samelson, Fluid exchange across a meandering jet. *Journal of physical oceanography* **22**, 431–444 (1992).

53. P. Castiglione, M. Cencini, A. Vulpiani, E. Zambianchi, Transport in finite size systems: an exit time approach. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **9**, 871–879 (1999).

54. M. Cencini, G. Lacorata, A. Vulpiani, E. Zambianchi, Mixing in a meandering jet: A markovian approximation. *Journal of physical oceanography* **29**, 2578–2594 (1999).

55. A. Martin, Phytoplankton patchiness: the role of lateral stirring and mixing. *Progress in oceanography* **57**, 125–174 (2003).

56. M. Danino, D. A. Kessler, N. M. Shnerb, Stability of two-species communities: drift, environmental stochasticity, storage effect and selection. *Theoretical population biology* **119**, 57–71 (2018).

57. A. Clauset, C. R. Shalizi, M. E. Newman, Power-law distributions in empirical data. *SIAM review* **51**, 661–703 (2009).

58. A. Klaus, S. Yu, D. Plenz, Statistical analyses support power law distributions found in neuronal avalanches. *PloS one* **6**, e19779 (2011).

59. J. Alstott, E. Bullmore, D. Plenz, powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one* **9**, e85777 (2014).

60. M. Rosenblatt, Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics* pp. 832–837 (1956).

61. E. Parzen, On estimation of a probability density function and mode. *The annals of mathematical statistics* **33**, 1065–1076 (1962).

62. H. Crane, *et al.*, The ubiquitous ewens sampling formula. *Statistical science* **31**, 1–19 (2016).

63. H. Morlon, D. W. Schwilk, J. A. Bryant, P. A. Marquet, A. G. Rebelo, C. Tauss, B. J. Bohannan, J. L. Green, Spatial patterns of phylogenetic diversity. *Ecology letters* **14**, 141–149 (2011).

64. K. J. Locey, J. T. Lennon, Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* **113**, 5970–5975 (2016).

65. A. J. Milligan, Oceanography: Plankton in an acidified ocean. *Nature Climate Change* **2**, 489 (2012).

66. H. Kontoyiannis, D. R. Watts, Observations on the variability of the gulf stream path between 74 w and 70 w. *Journal of physical oceanography* **24**, 1999–2013 (1994).

67. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

68. T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

69. R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, R. Knight, UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).

70. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

71. J. Bengtsson-Palme, M. Ryberg, M. Hartmann, S. Branco, Z. Wang, A. Godhe, P. De Wit, M. Sánchez-García, I. Ebersberger, F. de Sousa, A. S. Amend, A. Jumpponen, M. Unterseher, E. Kristiansson, K. Abarenkov, Y. J. K. Bertrand, K. Sanli, K. M. Eriksson, U. Vik, V. Veldre, R. H. Nilsson, Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution* **4**, 914–919 (2013).

Acknowledgements We thank M. Cencini, E. Economo, M.A. Muñoz, and L. Peliti for comments on a preliminary version of the manuscript.

Author Contributions P.V.M. and S.P. conceived the research; P.V.M. and S.P. wrote the code and performed the numerical simulations; A.B. and T.B. analyzed the metagenomic data; P.V.M. and S.P wrote the first draft, with subsequent input from A.B. and T.B.

Competing Interests The authors declare no competing interests