SCIENTIFIC REPORTS
nature research

OPEN

Scaling of species distribution explains the vast potential marine prokaryote diversity

Victor M. Eguíluz[1,2*], Guillem Salazar[3], Juan Fernández-Gracia[2], John K. Pearman[1], Josep M. Gasol[4], Silvia G. Acinas[4], Shinichi Sunagawa[3], Xabier Irigoien[5,6] & Carlos M. Duarte[1]

Global ocean expeditions have provided minimum estimates of ocean's prokaryote diversity, supported by apparent asymptotes in the number of prokaryotes with sampling effort, of about 40,000 species, representing <1% of the species cataloged in the Earth Microbiome Project, despite being the largest habitat in the biosphere. Here we demonstrate that the abundance of prokaryote OTUs follows a scaling that can be represented by a power-law distribution, and as a consequence, we demonstrate, mathematically and through simulations, that the asymptote of rarefaction curves is an apparent one, which is only reached with sample sizes approaching the entire ecosystem. We experimentally confirm these findings using exhaustive repeated sampling of a prokaryote community in the Red Sea and the exploration of global assessments of prokaryote diversity in the ocean. Our findings indicate that, far from having achieved a thorough sampling of prokaryote species abundance in the ocean, global expeditions provide just a start for this quest as the richness in the global ocean is much larger than estimated.

The ocean, the largest habitat in the biosphere, is a microbial-dominated ecosystem holding an estimated 10^{29} prokaryote cells[1]. Exploration of the ocean biodiversity associated with the huge prokaryote pool was prevented due to the limitations in the cultivation of marine prokaryotes[2]. This barrier was partially overcome by efficient sequencing approaches, typically targeting the genes that code for the 16S region of rDNA, which allows the definition and enumeration of the operational taxonomic units (OTUs) present in a sample, thereby providing a culture-free basis to assess biodiversity somewhat equivalent to that of species numbers[3]. In the past decade, global ocean expeditions and research based on them have utilized these technological developments in order to attempt to estimate the total number of prokaryote OTUs in the ocean[4–8]. For instance, the TARA Oceans Expedition explored prokaryote biodiversity in the upper ocean and described the detection of 35,650 prokaryote OTUs[5] in a set of globally distributed samples, with the exception of the Arctic, while the Malaspina Expedition gave a minimum estimate of the number of prokaryote OTUs in the deep ocean which is an order of magnitude lower, at around 3,700[4]. The TARA Expedition estimated the total richness to be 37,470 OTUs based on the Chao estimator, which defines a lower bound on species richness. This result should be interpreted to be at least 37.470 OTUs in the upper ocean.

The fraction of the total volume of the ocean sampled by any study is minimal and thus requires extreme extrapolation (over 20 orders of magnitude) from the number of species found in the samples to an estimate for the global ocean. The approach used is that of rarefaction curves, a development first introduced in 1943 by Fisher *et al*. to provide a basis to estimate the species richness of Malaysian butterflies[9], subsequently popularized by Sanders (1968)[10] to compare benthic invertebrate species richness from marine surveys with different sample sizes. Rarefaction curves use resampling techniques to develop a curve of the number of species against the number of samples collected[11]. Initially introduced to evaluate how comprehensive the assessment of species numbers was based on a sampling set, it was subsequently used to infer the total number of species in the ecosystem

[1]King Abdullah University of Science and Technology (KAUST), Red Sea Research Center (RSRC), Thuwal, 23955-6900, Saudi Arabia. [2]Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), E07122, Palma de Mallorca, Spain. [3]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zurich, Vladimir-Prelog-Weg 1-5/10, CH-8093, Zurich, Switzerland. [4]Departament de Biologia Marina i Oceanografia, Institut de Ciències del Mar-CSIC, Pg. Marítim de la Barceloneta 37-49, 08003, Barcelona, Spain. [5]AZTI - Marine Research, Herrera Kaia, Portualdea z/g, Pasaia (Gipuzkoa), 20110, Spain. [6]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain. *email: victor@ifisc.uib-csic.es

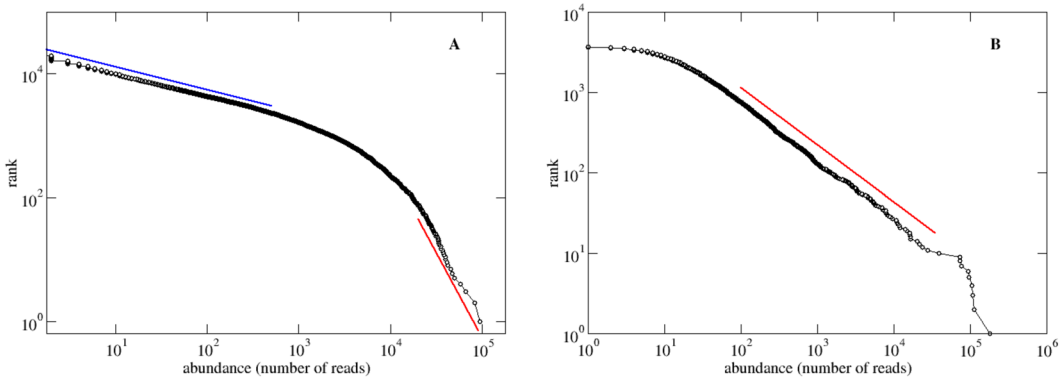

Figure 1. Abundance distribution of prokaryote OTUs in the upper and deep ocean. The rank *vs* abundance distribution for the (**A**) upper ocean and (**B**) deep ocean shows broad distributions with power-law tails. The abundance-rank distribution, $r \sim x^{-\alpha}$, where r is the rank of abundance x, has the same functional dependence (only the ranks have to be normalized between 0 and 1) as the complementary cumulative distribution CCD, $\mathrm{CCD}(x) = \sum_{i=x,\infty} P(i)$, where $P(i)$ is the abundance distribution. Thus, if the abundance rank distribution is given by $r \sim x^{-\alpha}$ the abundance distribution decays as $P(x) \sim x^{-1-\alpha}$. (**A**) For the upper ocean, the abundance distribution shows a double power-law decay separated at a characteristic scale of 2,313 reads: for abundances $x < 2{,}313$, the scaling exponent is 0.37 (blue line); for abundances $x > 2{,}313$, the scaling exponent is $\alpha = 1.57$ (see Materials and Methods). (**B**) For the deep ocean, the abundance-rank distribution is characterized by a power-law decay, $P(x) \sim x^{-1-\alpha}$, with an exponent of $\alpha = 0.89$ (red line).

investigated as that corresponding to the asymptote of the curve[12]. This approach was adopted to deliver estimates of the prokaryote species richness in the global ocean[4,5]. These estimates correspond mathematically to minimum estimates (e.g., Chao estimator)[13], yet their precision has not been assessed. Indeed, beyond the apparent asymptote in rarefaction curves, other estimators have been proposed to estimate species richness[13–16]. Marine prokaryote communities are characterized by the presence of a few abundant OTUs and a large number of rare OTUs[2], suggesting a much broader distribution of OTU abundance than that required to reliably apply rarefaction curves to estimate the global biodiversity of prokaryotes. Here we examine the scaling of prokaryote diversity in the ocean as a step to better understanding the extent that current assessments may underestimate prokaryote diversity in the global ocean. We do so using an array of novel approaches, including assessments across the global ocean coupled with experimental and *in silico* tests, to establish the scaling of ocean microbial diversity and explore its implications for the discovery of microbial diversity.

Results

Prokaryote diversity in the upper and deep ocean. The distribution of prokaryote OTUs in the upper ocean and deep ocean samples of the TARA Oceans[5] and Malaspina[4] Expeditions conform to broad distributions with power-law behavior, $P(x) \sim x^{-1-\alpha}$, where x represents the abundance measured in number of reads, and is characterized (the tail of the distribution) by a scaling exponent $\alpha = 1.57$ for the upper ocean, and $\alpha = 0.89$ for the deep ocean (Fig. 1), similar to the classic power-law describing the number of species per taxa of Willis and Yule (1922)[17]. A comparison to other broad distributions (lognormal, Weibull) shows that a distribution with a power-law tail (either pure power-law or truncated power-law) are most likely to be the best fitting (Table 1). This finding implies that the most abundant 1% OTUs account for 40% of the sequences while the least abundant 90% of sampled OTUs account for only 10% of the sequences in the upper ocean; while for the deep ocean, the most abundant 1% of OTUs account for more than 70% of the sequences while the least abundant 90% of sampled OTUs account for only 8% of the sequences.

Theoretical scaling. Prokaryote diversity and, in general, species diversity can be characterized by magnitudes like the Shannon and Simpson indices, which by giving greater weight to the larger, common species, provide estimators with less uncertainty[13] (Supplementary Table 1). However, the presence of rare species impacts the estimation of species richness. Species richness scales with sampling effort as a consequence of the power-law tail of the distribution of prokaryote abundance. Let us assume that the number of OTUs of abundance x, n_x, is given by $n_x = Ax^{-1-\alpha}$, where A is a normalizing constant, the scaling exponent α is larger than 0, $\alpha > 0$, and the abundances are in the range $n_x \in [1, N_{\max}]$. Thus, the total species richness, S, is given by $S = \sum_{x=1,N\max} n_x$. In the limit of large $N_{\max}$, the richness can be approximated as $S = A\zeta(1+\alpha)$, that is, $A = S/\zeta(1+\alpha)$, where $\zeta(\alpha)$ is the Riemann zeta function. The total number of reads N can be obtained by $N = \sum_{x=1,N\max} x\, n_x$. For $\alpha > 1$, we obtain

$$N = \frac{\zeta(a)}{\zeta(1+a)} \tag{1}$$

For $\alpha < 1$, in the continuous limit $N = A\int_1^{N_{\max}} x^{-a}dx = \frac{1}{(1-a)\zeta(1+a)} S(N_{max}^{1-a} - 1)$ and the assumption that $N_{max}^{1-\alpha} \gg 1$, we obtain

	ΔAIC PL	ΔAIC TPL	ΔAIC LN	ΔAIC Weibull	α	standard error (α)	β	λ
Upper ocean	0.47	**0**	0.74	0.77	1.57	0.09	1.34	0.000019
Deep ocean	0.03	**0**	2.01	15	0.89	0.02	0.73	0.000002
Mesocosm C1	23	**0**	8.44	5.84	0.52	0.02	0.41	0.000106
Mesocosm C2	**0**	2	2.01	3.00	0.52	0.31	0.52	0
Mesocosm C3	19	**0**	13	11	0.53	0.02	0.43	0.000088
Mesocosm C4	26	**0**	8.36	5.57	0.54	0.02	0.42	0.000119
Mesocosm C5	38	**0**	13	9.49	0.57	0.02	0.38	0.000178
Mesocosm C6	18	**0**	13	11	0.52	0.02	0.44	0.000080

Table 1. Comparing fitting models to the prokaryote abundance distribution. The delta Akaike Information Criterion (ΔAIC) indicates the most likely fit (value **0** in bold) and the difference to the most likely fit. For the six cases reported, the most likely fit is a distribution with a power-law decay (either pure or truncated). The parameters of a power law distribution $P(x) \sim x^{-1-\alpha}$ are the scaling exponent α; for the truncated power-law $P(x) \sim x^{-1-\beta} \exp(-\lambda x)$, are the scaling exponent β, and the characteristic abundance λ (λ = 0, for a pure power-law). ΔAIC PL: delta Akaike Information Criterion for power-law distribution fit; ΔAIC TPL: delta Akaike Information Criterion for truncated power-law distribution fit; ΔAIC LN: delta Akaike Information Criterion for log-normal distribution fit; ΔAIC W: delta Akaike Information Criterion for Weibull distribution fit. The standard error of the power-law scaling exponent (α) is also reported. For the upper ocean, the prokaryote abundance distribution shows a double power-law regime. A Maximum Likelihood Estimation for a double power-law model gives $P(x) \sim x^{-1-\delta}$, with exponent $\delta = 1.54$ for $x < 2{,}313$; and $P(x) \sim x^{-1-\alpha}$, with exponent $\alpha = 0.36$ for $x \geq 2313$ (see Materials and Methods).

$$N = \frac{1}{(1-a)\zeta(1+a)} S N_{max}^{1-a} \tag{2}$$

Finally, the abundance of the most abundant OTU can be evaluated as the value N_{max} at which there is only one group with abundance larger or equal than N_{max}, that is, in the continuous limit $\int_{N_{max}}^{\infty} n_x dx = 1$. This leads to $S N_{max}^{\alpha}$ (a detailed calculation can be found in ref. [18]).

Combining the previous expressions, we obtain the following scaling laws: $S \propto N_{max}^{\alpha}$ and for $\alpha < 1$

$$S \propto N_{max}^{a} \propto N^{\alpha} \tag{3}$$

while for $\alpha > 1$

$$S \propto N_{max}^{a} \propto N. \tag{4}$$

The same scaling laws are obtained in the Yule model[19] (which can also be mapped to the Simon model[20,21]), where the scaling exponent α is related to the ratio between speciation rate g and group growth s, $\alpha = g/s$. Systems showing distributions with power-law tails are ubiquitous: several methodologies have been described to fit and compare different functional forms as well as mechanisms to explain their origin[18,22–24].

Empirical and *in silico* scaling. The scaling of species richness and the distribution of species abundances are two sides of the same coin. The power-law distribution of prokaryote species abundance implies that species richness (S) scales with sampling effort (N, number of samples) as $S \sim N^{\gamma}$, where (i) γ equals the exponent of the rank-abundance power-law (i.e., $\gamma = \alpha$), when this exponent is $\alpha < 1$, as observed in the deep ocean (Malaspina Oceans Expedition, Fig. 2), and (ii) S is proportional to sampling effort (i.e., $\gamma = 1$) for larger exponents $\alpha > 1$, such as observed for the upper ocean (TARA Expedition, Fig. 2). Indeed, the power-law scaling of species richness with sampling effort implicit in the power-law distribution of the prokaryote species abundance distribution (Fig. 1) implies that the asymptote of rarefaction curves is artifactual and that indeed, the number of species does not approach any asymptote at the sampling effort this far deployed by global expeditions (Fig. 2). This expectation was confirmed by producing an *in silico* global ocean microbiome with an underlying distribution of prokaryote species abundance with the same shape and exponent as those empirically derived for the upper and deep ocean (dotted lines in Fig. 2). The *in silico* data was obtained, first, by expanding the empirically fitted data to larger populations and, second, by randomly generating abundance OTUs from the expanded distributions (see Materials and Methods). These simulations showed that increasing sampling effort, expressed as the total number of 16S reads sequenced, about 30 to 50 times relative to that applied to the upper and deep ocean by the TARA Oceans (3.3×10^6 reads, ref. [5]) and Malaspina Expedition (1.8×10^6 reads, ref. [4]) respectively would lead to estimates of prokaryote species abundance 4.2 and 1.2 times greater than inferred on the basis of rarefaction curves for the upper and deep ocean respectively (Fig. 2 and Supplementary Fig. 1). The estimators are calculated for a global population of 10^8 reads, which corresponds to 1 liter of upper ocean water (10^5 prokaryote cells/ml) and 10 liters of deep ocean water (10^4 prokaryote cells/ml) (Supplementary Table 1).

Mesocosm experiment. We challenged the mathematically-derived predictions, tested and confirmed by the *in silico* experiment, by enclosing a plankton community of the Central Red Sea in duplicate, and sampling

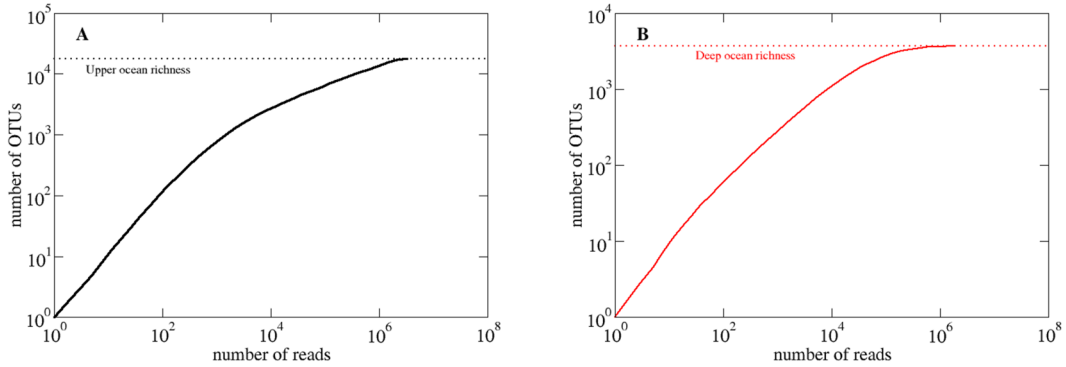

Figure 2. Number of species as a function of the number of reads. The expected number of OTUs in a random sampling of the total population grows sublinearly with sampling size, $S \sim N^{\gamma}$. (**A**) In the upper ocean (continuous black line), we can identify a first quasi-linear regime with $\gamma = 0.90$ (confidence interval 95% <0.01) and a second regimen with $\gamma = 0.33$ (confidence interval <0.01), while (**B**) in the deep ocean (continuous red line) the exponent $\gamma = 0.62$ (confidence interval <0.01). The number of OTUs in the upper ocean (horizontal dotted black line) is estimated at 35,650 OTUs[5] and in the deep ocean (horizontal dotted red line) the maximum number of OTUs found is 3,695[4].

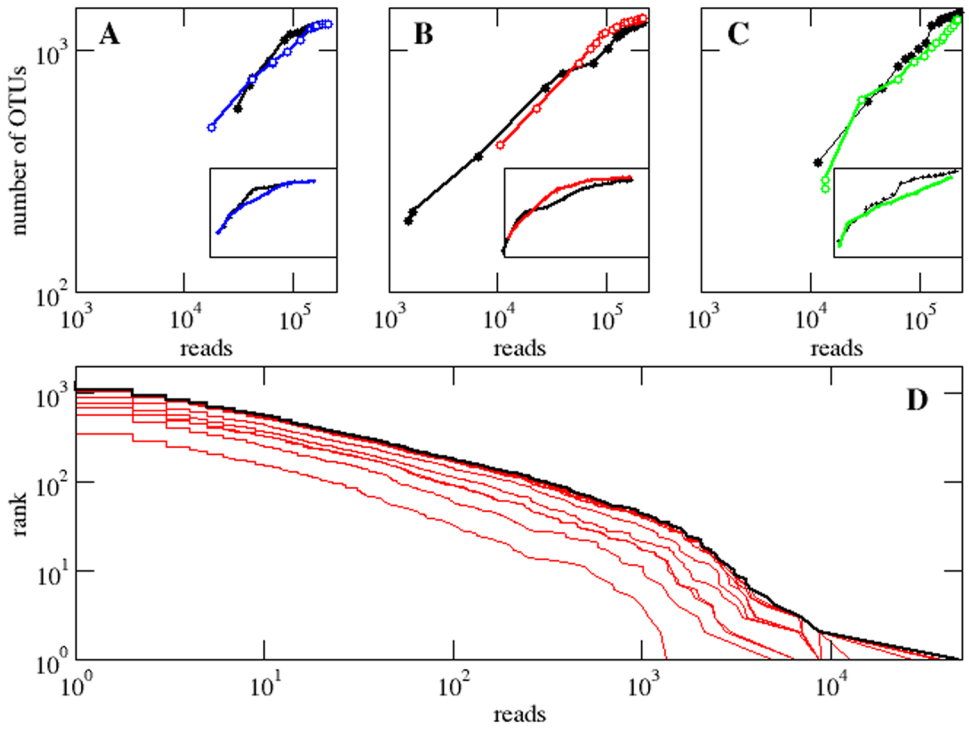

Figure 3. Scaling of the number of OTUs with the number of reads in an experiment. The number of prokaryote OTUs as a function of the number of reads is plotted, in a log-log scale, every two days as the experiment runs for 20 days in different conditions (**A**) control (Mesocosm C1 and C2), (**B**) single dose nitrate phosphate addition (NP) (Mesocosm C3 and C4), and (**C**) single dose nitrate phosphate sulfate addition (NPS) (Mesocosm C5 and C6). For all the conditions, we plot two replicates. The number of OTUs, S, scales with the number of reads, N, as $S \sim N^{\gamma}$, with $\gamma = 0.44$, 0.40 (control), 0.38, 0.40 (NP), 0.48, 0.52 (NPS). The insets show the same data in linear scale (same ranges as main plots) where an apparent saturation asymptote is observed. (**D**) Abundance vs rank plot for one of the controls for successive days from bottom to top. The exponent of a power-law distribution fit, $P(x) \sim x^{-1-\alpha}$, for the aggregated data after 20 days (black line) is $\alpha = 0.52$.

and sequencing it every day during 20 days[25] (c.f. Materials and Methods). The abundance distribution of prokaryote OTUs in the sampled Central Red Sea community continued to increase with additional sampling effort (Fig. 3), according to a power-law distribution with an average exponent of $\alpha = 0.53$, comparable to that obtained for the deep ocean ($\alpha = 0.89$) and for the less abundant of the upper ocean ($\alpha = 0.36$) (Fig. 3D). In line

	Scaling exponent γ	Confidence Interval (95%)	Days of observation
Mesocosm C1	0.44	0.026	14
Mesocosm C2	0.70	0.089	18
Mesocosm C3	0.44	0.039	19
Mesocosm C4	0.29	0.043	17
Mesocosm C5	0.36	0.062	19
Mesocosm C6	0.54	0.076	20

Table 2. Scaling exponents and confidence interval for the mesocosm experiment. For each condition and for each replica of the mesocosm experiment, the number of prokaryote species is fitted with the number of reads $S \sim N^{\gamma}$, with a least square method and the confidence intervals are calculated according to the number of days of observations in each condition.

with the upper and deep ocean cases, a comparative analysis performed for all the samples of the mesocosm experiment in three experimental conditions (control, single dose Nitrate-Phosphate addition and single dose Nitrate-Phosphate-Silicate addition) shows that a distribution with a power-law decay (either as a pure power-law or a truncated power-law) is the most likely fit (Supplementary Tables 2–7). The results confirmed the expectation that the number of OTUs retrieved in this community increased, on average, with the power 0.46 of the cumulative number of 16S reads sequenced without a clear asymptotic behavior despite exhaustive sampling (Fig. 3A–C and Tables 1 and 2).

Discussion

The results presented show that the abundance of different prokaryotic species in the ocean is described by a power-law distribution that implies that the total number of OTUs continues to increase, with a power given by that of the rank-abundance power-law, with increasing sampling effort. The dependence of the estimated richness on sampling effort is not an exclusive property of a power-law distribution and it has also been reported for lognormal distributions both theoretically[26] and empirically[7,23]. We expect that the effort-dependence of the species richness applies to distributions with sufficient long tails and thus characterized by the presence of many rare species (OTUs). Thus, in the presence of a rare biosphere[2], the effort-dependence of richness estimates is the expected outcome. Hence, the estimates that the upper and deep ocean contain ca. 37,000 and 3,700 prokaryote OTUs[4,5], respectively, derived from rarefaction curves is an underestimate (Fig. 2). The estimation of the diversity based on sampling effort (both the number of samples collected and the sequencing depth applied to each sample) still represents a challenge and requires broad extrapolations. We have addressed the estimation of prokaryote diversity with the parsimonious assumption that the sampled distribution represents the population distribution, furthermore supported by the relatively conserved shape of this abundance distributions when sampling is replicated as in our mesocosm experiment (see Supplementary Tables 2–7). Thus, we have explored the estimation of prokaryote diversity derived from fitting different underlying distributions to the upper and deep ocean, and the mesocosm experiment. Future research increasing sampling effort, both for individual communities and locations across the ocean, are likely to yield OTU counts much higher than these estimates. The power-law distribution of species richness is not a new observation in ecology[27–31] but is rooted in the seminal work of Willis and Yules showing a power-law distribution of species membership within taxa[17]. Indeed, a recent estimate of oceanic prokaryote species richness derived by extrapolating across more than 20 orders of magnitude the relationship between species numbers and number of cells sampled to match the 10^{29} prokaryote cells estimated in the global ocean, led to an estimate of 10^{10} different OTUs for this ecosystem[7]. Whereas the estimate derived from such wild extrapolation rests on a number of assumptions and does not necessarily reflect the shape of species abundance distribution of oceanic prokaryotes, it supports our empirical, mathematical, modeling and experimental results that indicate that the number of prokaryote OTUs in the ocean is far larger than currently estimated. A much-enhanced sampling effort is, therefore, required to unveil the prokaryote diversity concealed within the rare biosphere. Enhanced sampling efforts should be deployed both to retrieve the least abundant components of anyone community and also to benefit from the dynamics of microbial populations, which can bring otherwise rare components of the microbial biosphere to a level of abundance where they may be retrieved in sequencing projects (e.g., ref. [32]). Efforts to achieve an inventory of prokaryotic OTUs in the ocean will require a far more exhaustive sampling than deployed to date combined with sound extrapolation approaches rooted in the observed abundance distributions of prokaryotic OTUs.

Materials and Methods

Data and experimental design. We have analyzed three datasets. The three empirical datasets are: from the TARA expedition we collected the abundance of 18,022 OTUs from the surface water and deep chlorophyll maximum layers in 63 and 46 sites, respectively, containing 3,323,839 reads[5] (available at http://ocean-microbiome.embl.de/companion.html). From the Malaspina expedition, we collected the abundance of 3,695 free-living and particle-attached OTUs from 30 globally distributed sites in the bathypelagic ocean[4] (available at https://github.com/GuillemSalazar/MolEcol_2015). The experimental data reported the OTU abundance every day for a period of 20 days in three experimental conditions: (a) control (referred as Mesocosm C1 and C2), (b) single dose Nitrate-Phosphate addition (referred as C3 and C4), and (c) single dose Nitrate-Phosphate-Silicate addition (referred as C5 and C6) (Nitrate = 2 μM, Phosphate = 0.12 μM, Silicate = 3.75 μM)[25]. Samples range from an average of 11,126 ± 5,400 (SD) reads leading to 337 ± 100 (SD) OTUs the first day to an aggregated number of

212,761 ± 22,000 (SD) reads and 1,331 ± 56 (SD) OTUs after completion of the experiment. Raw reads, which the OTUs counts were based on, have been deposited in the NCBI Sequence Read Archive under the accession number SRP051855.

Statistical analysis. *Abundance distribution.* The model fittings of the power-law distributions, the truncated power-law distributions, lognormal distributions, and the stretched exponential distributions ware obtained with the Maximum Likelihood Estimation applied to the empirical data[33]. For the upper ocean, we have fitted also a double power-law distribution.

In silico prokaryote diversity: upper ocean. We proposed a distribution with two power-law regimes, with the parameter values (scaling exponents and transition point) obtained as described below: $P(x) = Ax^{-1-\delta}$, for abundances $x \leq x_c$, and $P(x) = Bx^{-1-\alpha}$, for $x > x_c$. The condition that the distribution is continuous at x_c ($P(x_c) = Ax_c^{-1-\delta} = Bx_c^{-1-\alpha}$) and the normalization ($\Sigma P(x) = 1$), lead to the values $A = \delta + (\delta - \alpha)\, x_c^{-\alpha}$, and $B = A x_c^{(\delta-\alpha)}$. We assigned to the exponents α and δ, and to the transition point x_c the values obtained from the Maximum Likelihood $\alpha = 1.54$, $\delta = 0.36$, and $x_c = 2{,}313$.

In silico prokaryote diversity: deep ocean. We proposed a shifted power-law to capture the power-law tail and the deviation at the head of the distribution: $P(x) = \alpha\, ((x + x_0)/(1 + x_0))^{-1-\alpha}$. The parameters α and x_0 can be obtained by the Maximum Likelihood Estimation: $\alpha = N_{OTU}\, \Sigma \log\, ((x_0 + x_i)/(1 + x_0))$, and $(x_0 + 1)\, \Sigma 1/(1 + x_i) = N_{OTU}\, \alpha\, /\, (1 - \alpha)$. To solve these implicit equations, we proposed x_0 and α, evaluate the previous expressions, and obtained new values x_0' and α'. We repeated these steps until we reached the condition $|x_0' - x_0| < T$, for some convergence value T. For $T = 10^{-6}$, the values we obtained are $\alpha = 0.89$, and $x_0 = 20.34$.

Akaike Information Criterion (AIC). The Akaike Information Criterion is defined as $AIC = -2 \log L + 2V$, where L is the maximum likelihood of a fit model, and V is the number of free parameters. The delta Akaike Information Criterion is calculated as $\Delta AIC = AIC\text{-}AIC_{min}$, where AIC_{min} corresponds to the minimum value of all the candidate models, and AIC the value of the candidate model. The weight AIC

$$w_i(AIC) = \frac{exp\left(\frac{-1}{2}\Delta_i \mathrm{AIC}\right)}{\sum_{K=1}^{M} exp\left(\frac{-1}{2}\Delta_k \mathrm{AIC}\right)}$$

can be interpreted as the probability that the model is the best model (in the AIC sense, that it minimizes the Kullback–Leibler discrepancy), given the data and the set
of candidate models (e.g., Burnham & Anderson, 2001).

Extrapolation of abundance distributions for larger number of samples. For the upper Ocean, the abundance distribution is fitted to a double power-law defined as $P(x) = Ax^{-1-\delta}$ for $x < x_c$ and $P(x) = Bx^{-1-\alpha}$ for $x_c < x$. A continuity condition ($Ax_c^{-1-\delta} = Bx_c^{-1-\alpha}$) and the normalization condition ($1 = \int_1^{\infty} P(x)dx$) gives the values for the constants A and B as $A = \alpha\delta(\alpha + (\delta - \alpha)x_c^{-\delta})^{-1}$ and $B = A\, x_c^{\alpha-\delta}$. In order to fit this distribution, we have to obtain estimates for the two exponents δ and α and for the cutoff x_c. We use first the maximum likelihood method implemented in ref. [30] which fits the exponent for the tail α and the value of the cutoff x_c. Then we adjust the value of the exponent for the range $[1, x_c]$ by using the same method, only fixing the minimum value to 1 and disregarding any data over the cutoff value x_c. In order to extract the behavior of the parameters for an increasingly large ecosystem, we used increasingly randomly aggregated samples from the TARA Oceans Expedition (139 samples in total). The average parameters for aggregations of samples of similar total number of reads are shown in the left column of Supplementary Fig. 2 in black and the error bars reflect their standard deviation. Next, in order to extrapolate these parameters to larger number of reads we fitted the estimated parameters to some simple curves (shown in red in Supplementary Fig. 2). The results were $x_c = 0.0002 \cdot N_{reads}^{1.1} + 52.6$, $\delta = 0.32\ (1 + 0.71\ \exp(-N_{reads}/570007))$ and $\alpha = 1.42\ (1 - 0.2\ \exp(-N_{reads}/110185))$. Note that the values of the scaling exponent of the tail of the distribution α are in agreement with recently reported estimates[34]. For the *in-vitro* generation of larger samples we extrapolated the parameter values to the value corresponding to the desired number of reads and generated random numbers from the corresponding distribution up to the desired number of reads, using the method of the inversion of the cumulative distribution.

For the deep Ocean, the abundance distribution is fitted to a shifted power-law $P(x) = A(x + x_0)^{-1-\alpha}$ with a maximum possible value for the abundance x_{max}. The value of A is given by the normalization condition ($1 = \int_1^{Xmax} P(x)dx$) and is $A = \alpha((1 + x_0)^{-\alpha} - (x_{max} + x_0)^{-\alpha})^{-1}$. In this case, we need to estimate again three parameters to fit the distribution. In order to estimate the parameters, we first fitted the exponent α and the shifting parameter x_0 by solving iteratively the equations from maximum likelihood:

$$a = S\left(\sum_{i=1}^{S} \log \frac{(x_i + x_0)}{1 + x_0}\right)^{-1}$$

$$x_0 = aS\left((1 + a)\sum_{i=1}^{s} \frac{1}{x_i + x_0}\right)^{-1},$$

where S stands for the number of data points. With those estimated parameters we estimated the maximum abundance $x_{\max}$ through the average abundance $<x>$ found in the data by solving the implicit equation $<x> = \int_1^{Xmax} xP(x)dx$:

$$\langle x \rangle = \frac{a}{1-a} \frac{(x_{max}+x_0)^{1-a} - (1+x_0)^{1-a}}{(1+x_0)^{-a} - (x_{max}+x_0)^{-a}} - x_0$$

The parameters are shown in the right column of Supplementary Fig. 2 and again in black are average estimates with standard deviations shown with error bars, and in red the simple fitted curves used for the extrapolation. In this case the simple curves fitted were $x_0 = 0.000003\ N_{reads}^{1.1} - 1$, $\alpha = 0.88\ (1 - 0.45 \exp(-N_{reads}/363263))$ and $<x> = 0.00042\ N_{reads}^{0.97} + 23.6$.

The estimation for a larger number of reads was performed as for the upper ocean but using the proper shifted power-law distribution as given by the extrapolated parameters.

Data availability

The TARA expedition dataset is available at http://ocean-microbiome.embl.de/companion.html; the Malaspina expedition dataset is available at https://github.com/GuillemSalazar/MolEcol_2015; and the experimental data have been deposited in the NCBI Sequence Read Archive under the accession number SRP051855.

Received: 1 August 2019; Accepted: 21 November 2019;
Published online: 10 December 2019

References

1. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences* **95**, 6578–6583 (1998).
2. Pedrós-Alió, C. & Manrubia, S. The vast unknown microbial biosphere. *Proceedings of the National Academy of Sciences* **113**, 6585–6587, https://doi.org/10.1073/pnas.1606105113 (2016).
3. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. *FEMS Microbiology Reviews* **25**, 39–67, https://doi.org/10.1111/j.1574-6976.2001.tb00571.x (2001).
4. Salazar, G. *et al*. Global diversity and biogeography of deep-sea pelagic prokaryotes. *ISME J* **10**, 596–608, https://doi.org/10.1038/ismej.2015.137 (2016).
5. Sunagawa, S. *et al*. Structure and function of the global ocean microbiome. *Science* **348**, https://doi.org/10.1126/science.1261359 (2015).
6. Zinger, L. *et al*. Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems. *PLOS ONE* **6**, e24570, https://doi.org/10.1371/journal.pone.0024570 (2011).
7. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* **113**, 5970–5975, 10.1073/pnas.1521291113 (2016).
8. Yooseph, S. *et al*. Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**, 60–66, http://www.nature.com/nature/journal/v468/n7320/abs/nature09530.html#supplementary-information (2010).
9. Fisher, R. A., Corbet, A. S. & Williams, C. B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology* **12**, 42–58, https://doi.org/10.2307/1411 (1943).
10. Sanders, H. L. Marine Benthic Diversity: A Comparative Study. *The American Naturalist* **102**, 243–282, https://doi.org/10.1086/282541 (1968).
11. Gart, J. J., Siegel, A. F. & German, R. Z. Rarefaction and Taxonomic Diversity. *Biometrics* **38**, 235–241, https://doi.org/10.2307/2530306 (1982).
12. Gotelli, N. J. & Colwell, R. K. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* **4**, 379–391, https://doi.org/10.1046/j.1461-0248.2001.00230.x (2001).
13. Haegeman, B. *et al*. Robust estimation of microbial diversity in theory and in practice. *The Isme Journal* **7**, 1092, https://doi.org/10.1038/ismej.2013.10, https://www.nature.com/articles/ismej201310#supplementary-information (2013).
14. Gotelli, N. J. & Colwell, R. K. Estimating species richness. *Biological diversity: frontiers in measurement and assessment* **12**, 39–54 (2011).
15. Gotelli, N. J. & Chao, A. In Encyclopedia of Biodiversity (Second Edition) 195–211 (Academic Press, 2013).
16. Chao, A., Colwell, R. K., Lin, C.-W. & Gotelli, N. J. Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* **90**, 1125–1133, https://doi.org/10.1890/07-2147.1 (2009).
17. Willis, J. C. & Yule, G. U. Some Statistics of Evolution and Geographical Distribution in Plants and Animals, and their Significance. *Nature* **109**, 177–179, https://doi.org/10.1038/109177a0 (1922).
18. Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* **46**, 323–351, https://doi.org/10.1080/00107510500052444 (2005).
19. Yule, G. U. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character **213**, 21–87, https://doi.org/10.1098/rstb.1925.0002 (1925).
20. Simon, H. A. On a Class of Skew Distribution Functions. *Biometrika* **42**, https://doi.org/10.1093/biomet/42.3-4.425 (1955).
21. Simkin, M. V. & Roychowdhury, V. P. Re-inventing Willis. *Physics Reports* **502**, 1–35, https://doi.org/10.1016/j.physrep.2010.12.004 (2011).
22. Perc, M. The Matthew effect in empirical data. *J. R. Soc. Interface* **11**, 20140378 (2014).
23. Deluca, A. & Corral, A. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophysica* **61**, 1351–1394 (2013).
24. Voitalov, I., van der Hoorn, P., van der Hofstad, R. & Krioukov, D. Scale-free networks well done. Phys. *Rev. Research* **1**, 033034 (2019).
25. Pearman, J. K., Casas, L., Merle, T., Michell, C. & Irigoien, X. Bacterial and protist community changes during a phytoplankton bloom. *Limnology and Oceanography* **61**, 198–213, https://doi.org/10.1002/lno.10212 (2016).
26. Curtis, T. P., Sloan, W. T. & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences* **99**, 10494–10499, https://doi.org/10.1073/pnas.142680199 (2002).
27. Hoffmann, K. H. *et al*. Power law rank–abundance models for marine phage communities. *FEMS Microbiology Letters* **273**, 224–228, https://doi.org/10.1111/j.1574-6968.2007.00790.x (2007).
28. Datta, S., Delius, G. W., Law, R. & Plank, M. J. A stability analysis of the power-law steady state of marine size spectra. *Journal of Mathematical Biology* **63**, 779–799, https://doi.org/10.1007/s00285-010-0387-z (2011).

29. Rozenfeld, A. F. *et al*. Network analysis identifies weak and strong links in a metapopulation system. *Proceedings of the National Academy of Sciences* **105**, 18824–18829, https://doi.org/10.1073/pnas.0805571105 (2008).
30. Edwards, R. A. & Rohwer, F. Viral metagenomics. *Nature Reviews Microbiology* **3**, 504, https://doi.org/10.1038/nrmicro1163 (2005).
31. Angly, F. *et al*. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**, 41, https://doi.org/10.1186/1471-2105-6-41 (2005).
32. Hugoni, M. *et al*. Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. *Proceedings of the National Academy of Sciences* **110**, 6004–6009, https://doi.org/10.1073/pnas.1216863110 (2013).
33. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-Law Distributions in Empirical Data. *SIAM Rev* **51**, 661–703, https://doi.org/10.1137/070710111 (2009).
34. Ser-Giacomi, E. *et al*. Ubiquitous abundance distribution of non-dominant plankton across the global ocean. *Nature Ecology & Evolution* **2**, 1243–1249, https://doi.org/10.1038/s41559-018-0587-2 (2018).

Acknowledgements

This research was funded by the Malaspina Circumnavigation Expedition supported by the Spanish Ministry of Science and Innovation through project Consolider-Ingenio Malaspina 2010 (CSD2008-00077) as well as King Abdullah University of Science and Technology (KAUST) through baseline funding to C.M. Duarte and X. Irigoien; by Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional (FEDER) through projects SPASIMM FIS2016-80067-P (AEI/FEDER, UE) and REMEI (CTM2015-70340-R); and by the Spanish State Research Agency through the María de Maeztu Program for Units of Excellence in R&D (MDM-2017-0711 to IFISC). S.S. is supported by the ETH and Helmut Horten Foundation. We thank Craig Michel for sequencing library preparation and Laura Casas for laboratory work. Further, we thank Naroa Aldanondo, Susana Carvalho, Amr Gusti, Karie Holtermann, Ioannis Georgakakis, Nazia Mojib and Tane Sinclair-Taylor as well as the personnel of the Coastal & Marine Resources core laboratory (CMOR) for their help in undertaking the sampling.

Author contributions

V.M.E. and C.M.D. conceived the idea; V.M.E., C.M.D. and J.F.-G. performed the analysis; V.M.E., G.S., J.F.-G. J.K.P., J.M.G., S.A., S.S., X.I. and C.M.D. contributed to the discussion, and writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-54936-y.

Correspondence and requests for materials should be addressed to V.M.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.