Journal of
Plankton Research

plankt.oxfordjournals.org

Featured Article

*J. Plankton Res.* (2014) 36(2): 334–343. First published online November 27, 2013 doi:10.1093/plankt/fbt115



# Conventional sampling methods severely underestimate phytoplankton species richness

**TAMARA RODRÍGUEZ-RAMOS[1]*, MARÍA DORNELAS[2], EMILIO MARAÑÓN[1] AND PEDRO CERMEÑO[3]**

[1]DEPARTAMENTO DE ECOLOXÍA E BIOLOXÍA ANIMAL, UNIVERSIDADE DE VIGO, CAMPUS AS LAGOAS-MARCOSENDE, VIGO 36310, SPAIN, [2]SCOTTISH OCEANS INSTITUTE AND CENTRE FOR BIOLOGICAL DIVERSITY, SCHOOL OF BIOLOGY, UNIVERSITY OF ST ANDREWS, EAST SANDS, FIFE KY16 8LB, UK AND [3]INSTITUTO DE CIENCIAS DEL MAR, CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS, PASSEIG MARÍTIM DE LA BARCELONETA, 37-49, BARCELONA E-08003, SPAIN

*CORRESPONDING AUTHOR: tamara@uvigo.es

Received April 26, 2013; accepted November 3, 2013

Corresponding editor: John Dolan

Conventional methods for the estimation of marine phytoplankton diversity include the collection of a small volume of seawater which is analysed under the microscope. We sampled natural communities and also synthetic communities generated under a neutral community model configuration and demonstrate that traditional sampling methods underestimate the species richness of marine phytoplankton communities. In our model, a synthetic community represents an ensemble of individuals enclosed in a parcel of seawater wherein the dynamics of each population is controlled by demographic stochasticity and dispersal. By sampling these synthetic communities, we found that roughly 20–45% of the species is missed by conventional, small volume samples. Consistent with the simulations, field data showed that the number of species increases with sampling effort by up to ~1.5-fold, revealing that these microbial communities might be more diverse than previously estimated. We suggest that conventional sampling methods have limited our ability to delineate the patterns of marine phytoplankton diversity and identify the underlying mechanisms. Improved sampling methods are proposed to obtain more accurate estimates of marine phytoplankton diversity.

KEYWORDS: phytoplankton; sample size; diversity; rare species; neutral models

available online at www.plankt.oxfordjournals.org

## INTRODUCTION

The statistical concept of a 'sample' refers to an estimate of a large body of information (Sournia, 1978), such as a group of organisms that make up an ecological community. The main objective of any sampling design is to obtain the maximum amount of information about the original source while minimizing the sampling effort. To do so, ecologists have developed analytical methods, such as species-accumulation curves and rarefaction analyses, which permit inference of the minimum sample size needed to obtain accurate estimates of community diversity. However, these methods have been rarely applied by microplankton ecologists. Conventional sampling protocols for the characterization of marine phytoplankton communities consist of collecting a small volume of seawater which is analysed under the microscope for species identification and cell counting (Lund *et al.*, 1958; Sournia, 1978). For example, the settling technique defined by Utermöhl in the late fifties allows details of phytoplankton communities to be recorded by settling a subsample of 5–100 mL for 24–48 h (Utermöhl, 1958). This technique is routinely used to characterize the taxonomic structure and biomass of phytoplankton communities. However, the extent to which these sampling volumes are sufficient to characterize the diversity of microbial communities remains largely untested.

Recent observations that microbial plankton communities include a large pool of species with low population abundances are changing our view of how these microbial species assemble into communities (Hughes *et al.*, 2001; Pedrós-Alió, 2006; Sogin *et al.*, 2006; Fuhrman, 2009). These studies suggest that a few dominant taxa form the bulk of microbial community biomass and energy flow, whereas many rare species account for high local diversities. This is a general pattern in species abundance distribution (SAD) for almost every type of biological community (McGill *et al.*, 2007). Determining the total number of rare species within microbial plankton communities is essential to obtain accurate estimates of species richness. However, owing to their low population abundances, their occurrence might have been systematically overlooked with traditional sampling protocols.

A major problem with our current estimates of microbial diversity is that sample sizes are much smaller than the size of the communities (Woodcock *et al.*, 2006). In nature, each individual organism experiences the environment on its own unique suite of spatial and temporal scales (Levin, 1992). As drifting organisms, planktonic microorganisms live in constant motion, which limits our ability to define the spatial boundaries of their communities and the extent to which these boundaries change in response to environmental perturbations. This is particularly important to microbial plankton communities, which are composed of populations of individuals with enormous dispersal ranges (Finlay, 2002; Caron and Countway, 2009; Cermeño and Falkowski, 2009). Furthermore, phytoplankton growth rates are in the range 0.1–1 per day (Marañón, 2005). As a result of the short generation times, as well as the high frequency of environmental reset (Sommer, 1985; Dolan, 2005), the entire community turns over in the order of days to weeks, giving rise to highly dynamic communities in terms of individual abundance, species richness and taxonomic composition.

The meta-community concept, a network of assemblages mutually interconnected through individuals' dispersal, provides an excellent framework within which to consider these questions (Gilpin and Hanski, 1991; Wilson, 1992; Vergnon *et al.*, 2009). By sampling a synthetic pool of species, we can explore aspects of the sampling process itself, such as the effect of increasing the sample size on the estimates of species richness, the 'unveiling' of rare species (Preston, 1948) or the effect of the spatial distribution of individuals in the detection of species. To the best of our knowledge, only a few studies have used this conceptual framework to study the dynamics and functioning of marine phytoplankton communities (Vergnon *et al.*, 2009; Segura *et al.*, 2011; Chust *et al.*, 2012). However, none of these tested the effect of sample size on the observed species richness.

Our objective was to quantify the extent to which traditional sampling methods underestimate the number of species of marine phytoplankton communities. To do so, we used two different approaches. First, we sampled natural phytoplankton communities and constructed species-accumulation curves by increasing the number of individuals sampled. Typically, the result is a curve that increases steeply at first, and then gradually levels off. The point at which the curve becomes asymptotic defines the so-called optimal sample size, an essential parameter to adequately characterize a community and thus to obtain meaningful estimates of species richness. Second, we constructed synthetic communities (i.e. a suite of assemblages enclosed in a parcel of seawater and connected through the dispersal of individuals) and tested the effect of sample size on the number of species observed under different dispersal scenarios. Changes in the dispersal parameter provide a means to explore the extent to which specific features of the spatial distribution of individuals, such as cell aggregation, patchiness (Montagnes *et al.*, 1999) or the presence of chain-forming species, influence the estimates of species richness. For instance, oceanic turbulence tends to homogenize microbial plankton populations in space and through time and hence, for a given number of individuals sampled, the number of species encountered is expected to increase

(Margalef, 1974). Both approaches indicated that marine phytoplankton communities are severely undersampled by applying conventional sampling protocols. We show that the main cause of underestimation is associated with a systematic undersampling of rare species, which indeed contribute to a large fraction of the species richness in marine phytoplankton communities.

## METHOD

### Sampling and microscopy analyses

Sampling of natural communities was conducted on board R/V Mytilus at a fixed station located in the Ría de Vigo (NW Iberian Peninsula, 42°14.09′N, 8°47.18′W) on 14 February 2012 (S1) and 30 March 2012 (S2). Subsurface seawater samples were collected using metal-clean Niskin bottles (12 L). Then, 2-L samples were fixed with 1% final concentration Lugol's iodine solution. Finally, 10 replicate subsamples of 50- and 5-mL each from S1 and S2, respectively, were examined under the microscope following the Utermöhl technique. The volume of seawater for sedimentation was chosen according to the chlorophyll *a* concentration of the sample (see Supplementary Data, section SD1), following conventional criteria (Lund *et al.*, 1958; Sournia, 1978). We counted and identified cells to species level using an inverted microscope (Nikon Eclipse TE2000-S). For *rare* species (low population abundances), we examined the entire slide to count and identify every single cell. For *common* and *dominant* species (high population abundances), we selected one transect at random on the slide to count individuals and identify species. Then, assuming a spatially homogeneous distribution of organisms in the sedimentation chamber, knowing the dimensions of the slide and the total volume settled, we extrapolated to obtain the total number of individuals per species in the total volume settled. Unknown species were classified according to morphological descriptions, for example '*Medium-sized, dark, thecate dinoflagellate*'. This nomenclature was consistent throughout the study. All the replicate subsamples were collected from the same sampling device (Niskin bottle). Hence, the total volume of sample inspected was 500 and 50 mL for S1 and S2 (pooled samples), respectively. Species were labelled according to their relative numerical abundance as *rare* (<1% of the total abundance), *common* (10–50%) and *dominant* (>50%).

### Model description and parameterization

A neutral community model based on previous configurations (Bell, 2000; Kadmon and Benjamini, 2006) was implemented to test the effect of sampling effort on the estimates of species richness of marine phytoplankton communities. The main assumptions and features of the model are listed in Supplementary Data, section SD2. In our model, a *synthetic community* represents an ensemble of individuals enclosed in a parcel of seawater, wherein the dynamics of each population is regulated by demographic stochasticity and dispersal. Each synthetic community is defined by a total number of individuals, $\mathcal{J}_M$, and a total number of species, $S_M$, and is divided into a suite of *compartments* mutually interconnected through individuals' dispersal (Fig. 1). Each compartment contains a given number of individuals, $\mathcal{J}_L$, and species, $S_L$. A Monte Carlo stochastic computer program iterates the probabilities of reproduction, $r$ (cell division), death ($d$) and dispersal (immigration, $m$, and emigration, $e$) to simulate the dynamics of the populations within each compartment. The model is set up with a maximum number of individuals in the compartment or carrying capacity, $K$, and the total number of species in the source pool, $S_M$. At each time step, the loss processes give rise to loss of individuals randomly selected. The vacant sites are filled in with new individuals entering the compartment via reproduction and immigration. After iterating all these processes, if $\mathcal{J}_L$ exceeds the carrying capacity, $K$, then the model eliminates at random the excess of individuals. Every individual has identical probabilities of reproduction, death and dispersal irrespective of its taxonomic affiliation.

Each synthetic community is defined by the parameters $\mathcal{J}_M$ and $S_M$. From empirical data (pooled samples), we estimated the expected number of species in

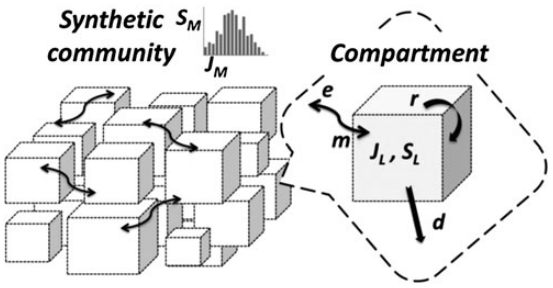


**Fig. 1.** Schematic representation of the neutral community model. The dynamics of populations is explained by demographic stochasticity and dispersal processes. The synthetic community is divided into several compartments interconnected through the dispersal of individuals. $\mathcal{J}_M$ is the total number of individuals and $S_M$ is the total number of species in the synthetic pool. $S_L$ and $\mathcal{J}_L$ are the number of species and the number of individuals within a given compartment, respectively. Death ($d$), reproduction ($r$) and dispersal (immigration, $m$, and emigration, $e$) probabilities control the dynamics of the populations. Neutrality means that all the species are ecologically equivalent, such that these probabilities apply randomly to all the individuals regardless of their taxonomic affiliation.

a well sampled community, $S_M$, in the range 90−105, by using species-accumulation curves and non-parametric estimators of species richness. Then, we generated pools of $10^6$−$10^7$ individuals ($\mathcal{J}_M$), according to a Poisson log-normal SAD built upon field data (see Supplementary Data, section SD2 for further details). Assuming a mean abundance of microphytoplankton in the Ría de Vigo of ∼100 individuals per mL (Cermeño and Figueiras, 2008), our synthetic communities would be enclosed within a 10−100 L water parcel.

The assemblage of individuals contained within a given compartment was simulated by drawing $\mathcal{J}_L$ individuals from the synthetic community described above. The taxonomic composition of each compartment is the result of the stochastic iteration of four probabilities. Reproduction ($r$) and mortality ($d$) probabilities were kept constant because they are expected to remain invariable in the short temporal scale (one time-step is 1 h) of the model (see also Supplementary Data, section SD2). Hence, for a given sample size, the differences in the species composition of the compartments were mainly due to differences in dispersal limitation and therefore, in the spatial distribution of the individuals. Different dispersal scenarios were simulated with the aim of determining the potential effect of patchiness and cell aggregation on the detection of species. We set up different dispersal probabilities (0.0005, 0.005, 0.01, 0.25 and 1). For instance, an immigration probability of 1 means no dispersal limitation, such that every individual in the synthetic community can arrive into any compartment per time step, which simulates a spatially homogeneous distribution of individuals. Increased sampling effort was simulated by increasing the numbers of individuals sampled, $\mathcal{J}_L$.

The model was written in R (R Development Core Team, 2012) and is available from the authors upon request.

## Data analysis

*Species-accumulation curves*

We constructed species-accumulation curves using data obtained from samples collected in the field and by sampling synthetic communities. Species-accumulation curves show the cumulative number of species as sampling effort increases (Gotelli and Colwell, 2001). The first point in the curve is defined as the mean number of individuals (axis $x$) and species (axis $y$) encountered when only one sample is considered. The second point results from averaging the number of individuals and species when the number of samples taken is 2 (i.e. every possible pair of subsamples pooled together), and so on.

The resulting data of number of individuals and species richness were then represented on a semi-log scale (Ugland *et al.*, 2003; Gray *et al.*, 2004),

$$S = c + z^* \log \mathcal{N}$$

where $S$ is the number of species, $\mathcal{N}$ is the sampling effort in terms of number of individuals; $c$ is a coefficient (intercept); and $z$ (slope), the scaling parameter, measures the rate at which we find new species with increasing sampling effort. The data were fitted to a linear regression model, using reduced major axis regression analysis (model II) given that both variables, $x$ and $y$, were measured systematically with error.

*Similarity indices*

Undersampling of rare species may lead to differences in community composition among replicates (subsamples). To explore this possibility, we calculated the dissimilarity among subsamples as a proxy of *precision*. Furthermore, the dissimilarity between each individual subsample and the corresponding pooled sample was used as a measure of the *accuracy* of the method. We computed three different indices of community (dis)similarity. The Jaccard index, a measure of the similarity between samples $j$ and $k$, is defined as (Jaccard, 1901; Chao *et al.*, 2006),

$$\mathcal{J}_{jk} = \frac{a}{(a+b+c)},$$

where $a$ is the number of species present in both samples $j$ and $k$, $b$ is the number of species present in sample $j$ but absent in $k$ and $c$ is the number of species present in sample $k$ but absent from $j$. This index emphasizes compositional changes. The Sørensen index, based on the Jaccard index, is computed as (Chao *et al.*, 2005),

$$S\phi_{jk} = \frac{2^*a}{(2^*a+b+c)},$$

and gives double weight to shared species among samples (Sørensen, 1948; Chao *et al.*, 2006). We represented their reciprocal ($1-\mathcal{J}_{jk}$ and $1-S\phi_{jk}$) to obtain dissimilarity values. Finally, the Bray−Curtis index of dissimilarity, which takes the relative abundance of species into account, is defined as (Bray and Curtis, 1957),

$$BC_{jk} = \frac{\sum |n_{ij} - n_{ik}|}{\sum (n_{ij} + n_{ik})},$$

where $n_{ij}$ and $n_{ik}$ are the numbers of individuals of the species $i$ in the samples $j$ and $k$, respectively. In all cases, a dissimilarity value equal to 1 means that no species are

shared between the pair of samples considered. These indices were computed using the functions *vegdist* in the package *vegan* (Oksanen *et al.*, 2012; R Development Core Team, 2012).

*Rarefaction analyses*

Rarefaction generates the expected number of species in a small collection of $n$ individuals drawn randomly from a large pool of $N$ individuals (Gotelli and Colwel, 2001). Hence, it allows the comparison of samples in terms of species richness after standardizing by the sampling effort applied. We distinguished three different functional groups: diatoms, dinoflagellates and *others* (mainly ciliates). Then, we performed rarefaction analyses using the function *rarefy* in the package *vegan* (Oksanen *et al.*, 2012; R Development Core Team, 2012), by re-sampling individuals (each subsample = 20 individuals) without replacement.

## RESULTS

A full list of the species identified and the main features of the phytoplankton communities analysed are provided in Supplementary data, Appendix.

The analysis of the hydrographic features (data not shown) and the vertical profiles of chlorophyll *a* concentration in the water column (Supplementary Data, section SD1) suggest that S1 and S2 were influenced by winter-mixing and spring-bloom conditions, respectively. On average, we estimated a total of 3049 ± 739 cells per 50 mL of seawater in S1 (~61 cells $mL^{-1}$), and 7002 ± 969 cells per 5 mL in S2 (~1400 cells $mL^{-1}$) ($n = 10$ subsamples, each). The mean number of species per subsample was 48 ± 5 and 39 ± 2 in S1 and S2, respectively. Only 19 species, out of a total of 91, were common to all the subsamples analysed from S1. The corresponding figures for S2 were 17, out of 73 species. In terms of relative abundance, most of the species (62.3 ± 11.7% of the total species richness) from S1 were on average *rare* (Table I), and only a few species (5.8 ± 1.6%) were in the range of *common* taxa. No *dominant* species were found. Similarly, in S2 most of the species fell within the category of *rare* (65.7 ± 5.7% on average in subsamples) (Table I), whereas a small fraction were in the range of *common* (9.8 ± 0.8%) and *dominant* (1.7 ± 1.2%). In both instances, the total number of species was significantly correlated with the number of rare taxa (Pearson's correlation coefficient was $r = 0.851$, $P = 0.002$ for S1; and $r = 0.905$, $P < 0.001$ for S2). After pooling all the subsamples together, we found 91 species (30 487 individuals sampled) in S1 and 73 species (70 022 individuals) in S2. In S1, 75.8% of the species were *rare* (Table I), and only 2 species, representing 2.2% of the total species richness, fell within the category of *common*. The percentage of *rare* species in S2 increased to 86.3% (Table I). Conversely, the fraction of *common* and *dominant* species decreased to 2.7% and 1.4% of total species richness, respectively. On average, conventional samples missed 48 and 46% of the total number of species observed in S1 and S2, respectively.

Figure 2 shows the species-accumulation curves computed from field data, and from model simulations at different levels of dispersal. At low sampling effort, the number of species observed in the field was close to that predicted by the model at low dispersal levels, suggesting that the individuals in both natural communities were heterogeneously distributed in space. The number of species increased with sampling effort either by adding subsamples (field data) or alternatively, by increasing the

*Table I: Total number of species, percentage of rare species and community dissimilarity indices computed from the field data*

| | S1 | | | | | S2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample | Species | % rare | Jac | *Sø* | B–C | Species | % rare | Jac | *Sø* | B–C |
| 1 | 48 | 54.17 | 0.57 | 0.73 | 0.45 | 42 | 71.43 | 0.31 | 0.53 | 0.25 |
| 2 | 54 | 62.96 | 0.55 | 0.77 | 0.42 | 36 | 72.22 | 0.29 | 0.63 | 0.24 |
| 3 | 48 | 54.17 | 0.53 | 0.79 | 0.40 | 36 | 72.22 | 0.32 | 0.53 | 0.26 |
| 4 | 51 | 72.55 | 0.55 | 0.72 | 0.42 | 39 | 66.67 | 0.28 | 0.56 | 0.23 |
| 5 | 55 | 72.73 | 0.52 | 0.74 | 0.40 | 38 | 71.05 | 0.34 | 0.52 | 0.27 |
| 6 | 51 | 62.75 | 0.56 | 0.70 | 0.43 | 36 | 66.67 | 0.31 | 0.55 | 0.25 |
| 7 | 42 | 59.52 | 0.53 | 0.72 | 0.41 | 40 | 70.00 | 0.35 | 0.49 | 0.27 |
| 8 | 42 | 64.29 | 0.63 | 0.71 | 0.51 | 42 | 71.43 | 0.33 | 0.53 | 0.26 |
| 9 | 45 | 60.00 | 0.59 | 0.73 | 0.47 | 42 | 76.19 | 0.42 | 0.50 | 0.32 |
| 10 | 41 | 60.98 | 0.55 | 0.75 | 0.43 | 39 | 69.23 | 0.33 | 0.52 | 0.26 |
| Pooled | 91 | 75.82 | 0.90 | 0.90 | 0.82 | 73 | 86.30 | 0.90 | 0.90 | 0.82 |

*Jac* (Jaccard index), *Sø* (Sørensen index) and B–C (Bray–Curtis index) dissimilarity represent the mean of every pair-wise comparison. *Pooled* values refer to the mean of the comparison between each subsample and the corresponding pooled sample.

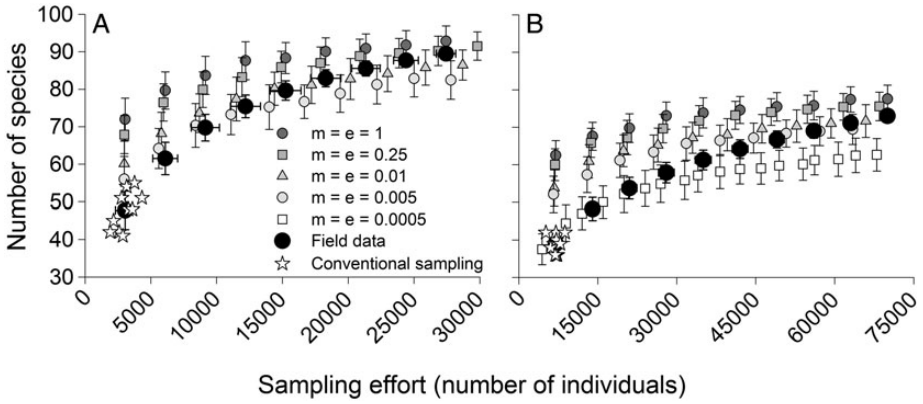


**Fig. 2.** Species-accumulation curves for (**A**) S1 and (**B**) S2 computed from field data (black dots) and model simulations at different levels of dispersal. Vertical and horizontal bars (sometimes not visible due to scale) represent the standard deviation of the number of species and the number of individuals identified from field data ($n = 10$) and model simulations ($n = 100$ simulations per dispersal level and sample size). The number of species obtained from conventional sample volumes (subsamples) is also depicted (white stars).

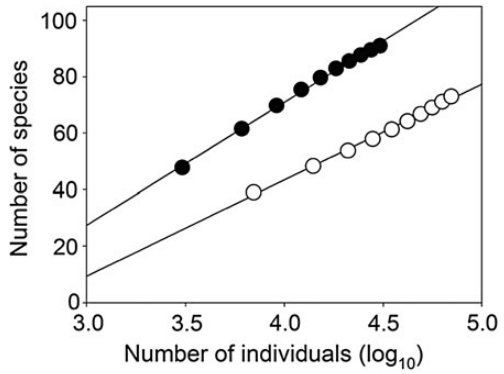


**Fig. 3.** Semi-log relationships between the cumulative species richness (number of species) and sampling effort (number of individuals) for S1 (black dots) and S2 (white dots).

number of individuals sampled from our synthetic communities. These curves fitted to a power function such that the number of species increased steeply at first, and then gradually levelled off. Our results show that 20–50% of the total number of species was missed by using conventional sampling methods (Fig. 2). The species-accumulation curves generated from field data tended to converge with the curves obtained from less crumpled synthetic communities (higher dispersal levels) as sampling effort increased.

The species-accumulation curves constructed from field data were represented on a semilog scale (Fig. 3). Then, we fitted a linear regression model (model II) to the data. The regression slopes were $43.62 \pm 0.71$ ($R^2 = 0.99$, $n = 10$) for S1, and $34.08 \pm 0.68$ ($R^2 = 0.99$, $n = 10$) for S2. The rate at which we found new species with increasing sampling effort was significantly higher in S1 (Student's *t*-test following Clarke's method (Clarke, 1980), $P < 0.001$).

We further explored the degree of variability of taxonomic composition among subsamples by calculating pair-wise dissimilarity indices (Table I). All the indices pointed to relatively high levels of dissimilarity. The Sørensen index showed a high degree of dissimilarity in both S1 and S2 (~0.74 and ~0.53 on average, respectively). The Jaccard and Bray–Curtis indices showed slightly lower dissimilarities (~0.56 and ~0.43 for S1 and ~0.33 and ~0.26 for S2, for the Jaccard and Bray–Curtis, respectively). After comparing the subsamples with the corresponding pooled sample (Table I), all the indices showed very high levels of dissimilarity (>0.8).

Figure 4 shows the rarefaction curves for diatoms, dinoflagellates and *others*. For a given number of individuals sampled, *others* and dinoflagellates showed the lowest and highest number of species, respectively (Fig. 4, upper panels). The rarefied number of species of diatoms and others were higher during winter mixing (Fig. 4, bottom panels). In contrast, dinoflagellates exhibited similar number of species regardless of the sampling time, but the curve was less saturated in S2. This suggests that the number of species of dinoflagellates is more severely underestimated during events of enhanced biomass concentration.

## DISCUSSION

Species richness is the simplest and most commonly used index of ecological diversity. For decades, terrestrial ecologists have been aware of the importance of performing species-accumulation curves and rarefaction analyses in order to define the minimum sample size needed to obtain accurate and comparable estimates of species richness (Gotelli and Colwell, 2001; Colwell *et al.*, 2004; Chao *et al.*, 2005). Only recently have microbial ecologists

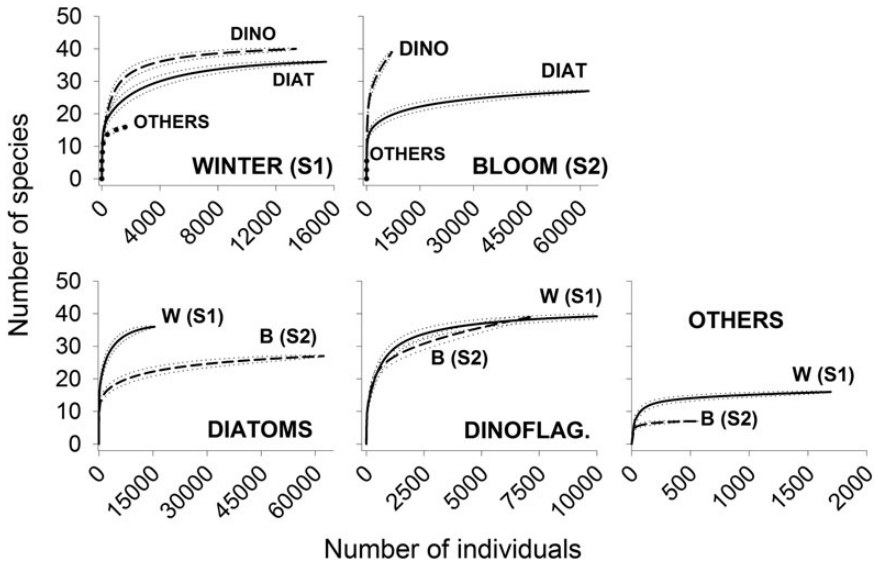


**Fig. 4.** Rarefaction curves for different functional groups (DIAT, diatoms; DINO, dinoflagellates, and OTHERS, mainly ciliates). Comparison among functional groups (upper panels) and between winter mixing (W) and spring-bloom (B) conditions (bottom panels). Thin-dotted lines represent the standard error.

begun to apply these analytical methods to explore the limits of microbial diversity (Hughes *et al.*, 2001; Pedrós-Alió, 2006; Sogin *et al.*, 2006; Dolan and Stoeck, 2011). These studies suggest that a large pool of species with low population abundances has been overlooked systematically by conventional sampling methods, stressing the need to test their performance and revisit the patterns of diversity.

By sampling natural communities and synthetic pools of species, we have shown that conventional sampling methods considerably underestimate the species richness of marine phytoplankton communities. Our analyses indicate that the main cause of underestimation is related to the failure of traditional methods to sample species with low population abundances. Indeed, despite increasing the sample size by 10-fold, our species-accumulation curves did not reach any clear saturation level, indicating that the rarest species still remained to be sampled. Based on non-parametric estimators of species richness such as Jackniffe or Chao 1, we found that roughly 90% of the species were recovered by increasing the sample size, whereas only a modest ~50% resulted from applying conventional sampling methods.

Rare species have been estimated to comprise >80% of the total pool of species in ciliate communities (Foissner *et al.*, 2002). Their occurrence is even greater in bacterial assemblages wherein rare species, defined according to specific genetic markers, might account for >90% of the total diversity (Pedrós-Alió, 2006; Sogin *et al.*, 2006). The strong positive correlation between total and rare species observed in our study bears out the importance of rare taxa in phytoplankton communities. The systematic undersampling of these rare taxa might have biased the understanding of important descriptors of marine phytoplankton communities, such as the species richness, size structure or species-abundance distributions. For instance, our results highlight differences in the degree of species underestimation attributable to differences in taxonomic affiliation and cell size (Fig. 4). Large-sized phytoplankton cells are commonly encountered at low population abundances (Huete-Ortega *et al.*, 2010), which could bias their detection in previous studies of phytoplankton diversity.

A key point in the definition of any methodological protocol is the accuracy of the measurement. In his review on the microdistribution of plankton Cassie (Cassie, 1963) poses: 'since it is hoped that every plankton sample is representative of the population of a very much larger volume of water, the range of variation in samples taken within this volume is of importance in evaluating the differences between more widely spaced samples'. The subsamples analysed here showed a significant variability in species composition despite being collected from the same sampling device. Less than 25% of the species were common to every subsample, questioning the validity of local ($\alpha$-) diversity estimates in studies of community dynamics and diversity–productivity relationships. The enormous uncertainty of previous estimates of species richness further suggests that the patterns of community turnover, $\beta$-diversity, which rely

on local estimates, should be interpreted cautiously. Several studies have reported increases in community dissimilarity with geographic distance for aquatic microorganisms (Hillebrand and Azovsky, 2001; Chust *et al.*, 2012), which could be in part a result of methodological biases in the detection of rare species. For instance, dominant taxa in a given location may be rare elsewhere and thus undetectable by conventional sampling methods. Besides, species that occur in low abundance are less likely to be detected at a given location, even if their true distributions are, in fact, widespread (Amend *et al.*, 2013). These issues might have resulted in artificially steep distance–decay relationships. Increased sampling effort is expected to increase the probability of sampling rare species thereby reducing the dissimilarities among distant communities.

Species–area relationships depict the total number of species plotted against the area of the system. On a log–log scale, the slope of this relationship ($z$) provides information on important aspects of species biogeography and evolution. Low values of $z$ are indicative of broad dispersal ranges and low habitat specificity, whereas steeper slopes imply barriers to dispersal and mosaic-like distributions (Smith *et al.*, 2005). Azovsky (Azovsky, 2002) reported $z$ values of 0.066 for benthic Arctic diatoms, Smith *et al.* (Smith *et al.*, 2005) found $z$ values of 0.134 for a compilation of phytoplankton data from experimental and natural aquatic ecosystems, whereas Hillebrand *et al.* (Hillebrand *et al.*, 2001) reported a wide range of $z$ values ($\sim$0.1 to $\sim$0.7) for unicellular microalgae, similar to those found for multicellular organisms. Assuming a given regional/global diversity, an increase in local species richness (e.g. increasing sampling effort) would imply lower values of $z$, given that a larger fraction of the regional species pool would be represented locally. This agrees with the idea that free-living microbial eukaryotes exhibit high local diversities, yet their regional/global species numbers are relatively low (Finlay, 2002; Fenchel and Finlay, 2004). Alternatively, Woodcock *et al.* (Woodcock *et al.*, 2006) have suggested the possibility that the effect of undersampling increases with the area surveyed. Species–area relationships are built up by pooling data obtained from discrete samples. They argue that our ability to detect more rare species decreases with increasing the area of the system, which might have artificially flattened the species–area relationships reported for microbial communities.

Low dispersal model configurations allow simulation of processes such as particle aggregation and patchiness, which limit individuals' dispersal and hence lead to heterogeneous distributions of species. Our results show that at low sampling effort, the number of species observed in the field was lower than those predicted by the model at high dispersal levels (homogeneous distribution of organisms), suggesting that the individuals in these natural communities were heterogeneously distributed in space. This is further supported by the high dissimilarity observed among samples (Table I). The results of the model serve to illustrate how processes that generate heterogeneous distributions of organisms could bias the detection of species and hence the estimates of species richness. Our analysis shows that this issue is largely solved by increasing the volume of sample analysed.

The hydrodynamic conditions of the study site influence the detection of species and thus the shape of the species-accumulation curves. The rate at which new species were detected with increasing sample size was significantly higher in S1 (Fig. 3), which was characterized by intense vertical mixing. We propose three potential explanations that, combined or individually, might explain this result. First, the arrival of immigrants and thus the number of rare species in the system increases with oceanic mixing. Second, turbulent regimes homogenize the distribution of individuals in space and hence, for a given number of individuals counted, the probability of finding new species increases (Margalef, 1978). Third, the higher evenness in the community species-abundance distribution, characteristic of winter mixing conditions, might further increase the likelihood of detecting rare species. This possibility has been proposed by Caporaso *et al.* (Caporaso *et al.*, 2012) for microbial plankton communities in the Western English Channel. On the other hand, in stratified systems, phytoplankton populations are typically segregated across space and through time and thus, to adequately estimate the diversity of these microbial communities, it is necessary to increase the sampling effort. Although more data are needed to confirm these results, our analyses suggest that *a priori* the probability of detecting new species is a positive function of the turbulence of the system, linking the patterns of community structure to the hydrodynamics of the water column (Margalef, 1974).

To adequately characterize the diversity of marine phytoplankton communities, traditional sampling methods must be reconsidered. First, the observation that a large pool of rare species account for a substantial amount of the species richness in marine microbial plankton communities implies the need for building species-accumulation curves. Rare species comprise a critical component of these communities and their detection requires improved sampling designs, including replicate subsamples, and larger sample volumes that can be obtained by using for example plankton nets. Second, in many instances the total number of individuals differs significantly among communities inhabiting different sites. Hence, estimates of species richness must be standardized by a quantitative measure of sampling effort to make meaningful

comparison of community diversity (Gotelli and Colwell, 2001).

We have shown that conventional sampling volumes miss a large fraction of the species richness in marine phytoplankton communities. By using small volumes of seawater, systematic undersampling has hampered the search for diversity patterns and might have underrated fundamental mechanisms of community assembly such as the power of oceanic turbulence and dispersal. Our results suggest that marine phytoplankton communities might be more diverse than previously estimated. Species-accumulation curves and rarefaction analyses are thus essential analytical tools to implement within traditional protocols and hence obtain more accurate estimates of phytoplankton community diversity.

## SUPPLEMENTARY DATA

Supplementary data can be found online at http://plankt.oxfordjournals.org.

## ACKNOWLEDGEMENTS

We thank all people who contributed to the collection of data. We especially thank M. P. Lorenzo for her help in sampling and sample processing, and I. G. Teixeira and F. G. Figueiras for microscopy analyses and identification of species.

## FUNDING

This work was supported by the Spanish Ministerio de Economía y Competitividad (MINECO), through projects DISTRAL (CTM2011-25035) and PERSEO (CTM2008-03699). T. R.-R. acknowledges the receipt of a FPI predoctoral fellowship from MINECO.

## REFERENCES

Amend, A. S., Oliver, T. A., Amaral-Zettler, L. A. *et al.* (2013) Macroecological patterns of marine bacteria on a global scale. J. Biogeogr., **40**, 800–811.

Azovsky, A. I. (2002) Size-dependent species-area relationships in benthos: Is the world more diverse for microbes? *Ecography*, **25**, 273–282.

Bell, G. (2000) The distribution of abundance in neutral communities. *Am. Nat.*, **155**, 606–617.

Bray, J. R. and Curtis, J. T. (1957) An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr., **27**, 325–349.

Caporaso, J. G., Paszkiewicz, K., Field, D. *et al.* (2012) The Western English Channel contains a persistent microbial seed bank. *ISME J.*, **6**, 1089–1093.

Caron, D. A. and Countway, P. D. (2009) Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquat. Microb. Ecol.*, **57**, 227–238.

Cassie, R. M. (1963) Microdistribution of plankton. *Oceanogr. Mar. Biol. Annu. Rev.*, **1**, 223–252.

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

Cermeño, P. and Figueiras, F. G. (2008) Species richness and cell-size distributions: size structure of phytoplankton communities. Mar. Ecol. Prog. Ser., **357**, 79–85.

Chao, A., Chazdon, R. L., Colwell, R. K. *et al.* (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol. Lett.*, **8**, 148–159.

Chao, A., Chazdon, R. L., Colwell, R. K. *et al.* (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, **62**, 361–371.

Chust, G., Irigoien, X., Chave, J. *et al.* (2012) Latitudinal phytoplankton distribution and the neutral theory of biodiversity. *Global Ecol. Biogeogr.*, **22**, 531–534.

Clarke, M. (1980) The reduced major axis of a bivariate sample. *Biometrika*, **67**, 441–446.

Colwell, R. K., Mao, C. X. and Chang, J. (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, **85**, 2717–2727.

Dolan, J. R. (2005) Biogeography of aquatic microbes. Aquat. Microb. Ecol., **41**, 39–48.

Dolan, J. R. and Stoeck, T. (2011) Repeated sampling reveals differential variability in measures of species richness and community composition in planktonic protists. *Environ. Microbiol. Rep.*, **3**, 661–666.

Fenchel, T. and Finlay, B. (2004) The ubiquity of small species: Patterns of local and global diversity. *Bioscience*, **54**, 777–784.

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

Foissner, W., Agatha, S. and Berger, H. (2002) Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa), with emphasis on two contrasting environments, the Etosha Region and the Namib Desert. *Denisia*, **5**, 1–1459.

Fuhrman, J. A. (2009) Microbial community structure and its functional implications. *Nature*, **459**, 193–199.

Gilpin, M. E. and Hanski, I. (1991) Metapopulation Dynamics: Empirical and Theoretical Investigations. Academic Press, London, UK.

Gotelli, N. J. and Colwell, R. K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.*, **4**, 379–391.

Gray, J. S., Ugland, K. I. and Lambshead, J. (2004) Species accumulation and species area curves—a comment on Scheiner (2003). *Global Ecol. Biogeogr.*, **13**, 473–476.

Hillebrand, H. and Azovsky, A. I. (2001) Body size determines the strength of the latitudinal diversity gradient. *Ecography*, **24**, 251–256.

Hillebrand, H., Watermann, F., Karez, R. *et al.* (2001) Differences in species richness patterns between unicellular and multicellular organisms. *Oecologia*, **126**, 114–124.

Huete-Ortega, M., Marañón, E., Varela, M. *et al.* (2010) General patterns in the size scaling of phytoplankton abundance in coastal waters during a 10-year time series. *J. Plankton Res.*, **32**, 1–14.

Hughes, J. B., Hellmann, J. J., Ricketts, T. H. *et al.* (2001) Counting the uncountable: statistical approaches to estimating microabial diversity. Appl. Environ. Microbiol., **67**, 4399–4406.

Jaccard, P. (1901) Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. Vaud. Sci. Nat.*, **37**, 547–579.

Kadmon, R. and Benjamini, Y. (2006) Effects of productivity and disturbance on species richness: a neutral model. *Am. Nat.*, **167**, 939–946.

Levin, S. A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.

Lund, J. W. G., Kipling, C. and Le Cren, E. D. (1958) The inverted microscope method of estimating algal numbers and the statistical basis of estimations by counting. *Hydrobiologia*, **11**, 143–170.

Marañón, E. (2005) Phytoplankton growth rates in the Atlantic subtropical gyres. Limnol. Oceanogr., **50**, 299–310.

Margalef, R. (1974) *Ecologia*. Omega, Barcelona, Spain.

Margalef, R. (1978) Life-forms of phytoplankton as survival alternatives in an unstable environment. *Oceanol. Acta*, **1**, 493–509.

McGill, B. J., Etienne, R. S., Gray, J. S. *et al.* (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.*, **10**, 995–1015.

Montagnes, D., Poulton, A. and Shammon, T. (1999) Mesoscale, finescale and microscale distribution of micro-and nanoplankton in the Irish Sea, with emphasis on ciliates and their prey. Mar. Biol., **134**, 167–179.

Oksanen, J., Blanchet, F. G., Kindt, R. *et al.* (2012) *Vegan: Community Ecology Package*. R package version 2.0-5, http://CRAN.R-project.org/package=vegan.

Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? *Trends Microbiol.*, **14**, 257–263.

Preston, F. W. (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN:3-900051-07-0, URL: http://www.R-project.org/.

Segura, A. M., Calliari, D., Kruk, C. *et al.* (2011) Emergent neutrality drives phytoplankton species coexistence. *Proc. R. Soc. B.*, **278**, 2355–2361.

Smith, V. H., Foster, B. L., Grover, J. P. *et al.* (2005) Phytoplankton species richness scales consistently from laboratory microcosms to the world's oceans. *Proc. Natl Acad. Sci. USA*, **102**, 4393–4396.

Sogin, M. L., Morrison, H. G., Huber, J. A. *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA*, **103**, 12115–12120.

Sommer, U. (1985) Comparison between steady state and non-steady state competition: experiments with natural phytoplankton. *Limnol. Oceanogr.*, **30**, 335–346.

Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, **5**, 1–34.

Sournia, A. (1978) *Phytoplankton Manual*. Monographs on Oceanographic Methodology, vol. 6. UNESCO, Paris.

Ugland, K. I., Gray, J. S. and Ellingsen, K. E. (2003) The species-accumulation curve and estimation of species richness. *J. Anim. Ecol.*, **72**, 888–897.

Utermöhl, H. (1958) Zur vervoll kommung der quantitativen phytoplankton-methodik. *Mitt. Int. Ver. Theor. Angew. Limnol.*, **9**, 1–38.

Vergnon, R., Dulvy, N. K. and Freckleton, R. P. (2009) Niche versus neutrality: uncovering the drivers of diversity in a species-rich community. *Ecol. Lett.*, **12**, 1079–1090.

Wilson, D. S. (1992) Complex interactions in metacommunities, with implications for biodiversity and higher levels of selection. *Ecology*, **73**, 1984–2000.

Woodcock, S., Curtis, T. P., Head, I. M. *et al.* (2006) Taxa-area relationships for microbes: the unsampled and the unseen. Ecol. Lett., **9**, 805–812.