COMMENTARY

The vast unknown microbial biosphere

Carlos Pedrós-Alió[a,1] and Susanna Manrubia[a,b]

The desire to know how many living beings there are on the planet has been a permanent obsession of biologists. Many inventories have been done, especially with large animals and plants, since naturalists began modern taxonomy in the 18th century. But there is the feeling that we are missing a large part of the diversity of microorganisms. How large this ignorance is, however, is a matter of debate. In PNAS, Locey and Lennon (1) provide one such estimate that, if true, will have microbiologists struggling with the task of describing this diversity for decades, maybe centuries.

Fig. 1*A* shows the distribution of primary producers on the planet, both on land and in the oceans. Despite the known difficulties of determining species and avoiding synonyms, we have a reasonable knowledge of where the ∼300,000 species of the main primary producers on land, the vascular plants, are (Fig. 1*B*). When we turn to the main primary producers in the oceans, however, we are at a loss to say how many species are there and how they are distributed. We are left with just the chlorophyll concentrations in Fig. 1*A*, because the marine primary producers are microbes. The contrast between our knowledge of the diversity of macrobes and microbes is thus apparent; the paper by Locey and Lennon (1) tries to put together data on both to estimate the total number of microbial species on Earth.

Microbes and macrobes have seldom been considered together in the same study and with the same tools. Basically, considerations of the total number of species have ignored bacteria and archaea and estimates about protists have not been detailed enough. On the other hand, several extrapolation methods have been feeding a lively debate about the total number of species of insects and, particularly, of beetles. Erwin (2) created quite a stir with his estimate of 30×10^6 species of insects in tropical forests. This number was based on canopy fogging of 19 trees of *Luehea seemannii*. The tree canopy was packed in a bag and an insecticide fog spread inside. Then the dead insects were collected at the bottom. This was a major effort in itself, but it was just the beginning: 1,200 species of beetles were sorted out. Eventually, 162 species were considered to be specific to the particular tree host species. Next, a series of assumptions were needed to estimate the total number of insects. The number of tree species (50,000) was multiplied by this number of tree-specific beetles to give 8×10^6 species of canopy beetles. The proportion of beetles among insects was further used to estimate 20×10^6 arthropods in the canopy. Finally, the ground species were added to give a grand total of 30×10^6. This number was scary. Only about 1×10^6 species of insects had been described thus far; if the total were 30×10^6 the task of describing them seemed impossible. Since then, several other exercises have been carried out and the more optimistic scientists estimate that there may be "only" between $5 \pm 3 \times 10^6$ (3) and $8.7 \pm 1.3 \times 10^6$ species (4). These lower estimates have been used to argue that describing all of the species is possible in the next century in the face of increasing extinction rates (3).

Of course the microbes were completely absent from such estimates. However, microbes rule the world. They are the most numerous, the most diverse, and the most important species in terms of carbon and nutrient cycling globally. How can a biologist speculate on the number of species on Earth and avoid considering microbes?

The difficulty is that describing a microbial species is a tricky issue. In the case of bacteria and archaea, one needs to isolate it in pure culture, describe it biochemically, morphologically, and genetically, the 16S rRNA has to be sequenced, and the description has to be published in one of the few official bacterial taxonomy journals. This process is not that different from the effort necessary to describe a beetle. However, the main problem is that many bacterial and archaeal species are not easily isolated in pure culture. Some estimate than only 1% of the species present in an ecosystem can be retrieved in culture with current techniques.

One shortcut came from sequencing of the 16S rRNA from natural communities. All of a sudden, this approach provided a wealth of novel microbial diversity that had remained beyond the reach of

[a]Systems Biology Program, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC), 28049 Madrid, Spain; and [b]Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain

Author contributions: C.P.-A. and S.M. wrote the paper.

The authors declare no conflict of interest.

See companion article on page 5970 in issue 21 of volume 113.

[1]To whom correspondence should be addressed. Email: cpedros@cnb.csic.es.

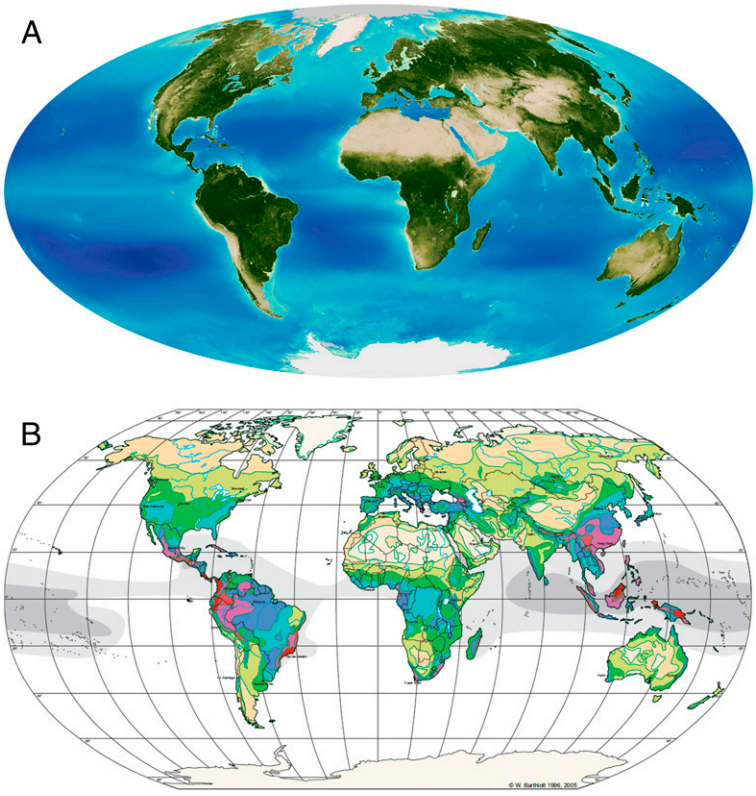

Fig. 1. (*A*) **Abundance of primary producers on Earth. Tan to green colors on land show density of green vegetation as the normalized difference vegetation index. Deep blue to yellow color in the oceans show chlorophyll concentration (mg/m^{-3}). Image courtesy of NASA Earth Observatory.** (*B*) **Number of vascular plant species on land, from <100 species per 10,000 km^2 (white) to >5,000 (red). Image courtesy of ref. 2 and Wilhelm Barthlott (University of Bonn, Bonn), copyright Wilhelm Barthlott.**

taxonomists. Some experts claim that 16S rRNA sequences overestimate the number of species. But all of the evidence points in the opposite direction (5). Every time an experimental test has been carried out, ribosomal RNA has been shown to underestimate the biological species concept. Therefore, using these sequences as proxies, one is on the "safe" prudent side. In the last decade the implementation of high-throughput sequencing techniques has resulted in a massive amount of sequence data from most ecosystems, and these are now ready to be used in further extrapolations.

Locey and Lennon (1) have made extensive use of the microbial sequence data from many different environments. However, the authors have combined them with data from animals and plants, trying to find a universal law that would explain diversity across the whole tree of life. Interestingly, these datasets come both from high-technology massive high-throughput sequencing efforts and from eager birdwatchers participating in Christmas counts with just binoculars and a finely tuned ear. By using these datasets and some techniques that we will examine next, they arrived at an estimate of 10^{12} microbial species on Earth. Obviously the 30×10^6 estimates of Erwin (2) become insignificant in comparison with this number. But, how good is this estimate?

There are several quantities that have been directly measured for the communities in this study (1), among them richness (total number of taxa, S), size of the sample (total number of individuals, N), and abundance of the most frequent taxon (N_{max}). The dataset spans eight orders-of-magnitude in N. However, extrapolation to the whole planet implies 10^{30} individuals: that is, 22 additional orders-of-magnitude, which requires a sufficiently good functional fit to the data, with good statistics, to allow the extrapolation within reasonable confidence intervals.

Locey and Lennon (1) tried different functions and found out that N and N_{max} closely fulfilled an allometric relationship of the form $N_{max} = 0.38\ N^{0.93}$ as best fit. The confidence intervals shown were small, therefore supporting a priori the feasibility of an extrapolation to larger communities. To provide additional independent estimates of how good this extrapolation was, the authors used actual numbers for N and N_{max} from three experimentally available pairs of data: the human gut, the cow rumen, and the pelagic oceans. The extrapolation fit these data extremely well, thus nicely encompassing both macro- and microorganisms across 30 orders-of-magnitude of N.

Direct measurements of the taxa diversity S are reported in the original dataset for communities with up to $N = 10^8$ individuals, although the fit is not as good as the one relating N and N_{max}. However, S can also be inferred by means of ecological theory. Here, Locey and Lennon (1) resorted to a framework that goes back to studies by Preston (6) on the species abundance distribution (SAD). Preston was first to describe how the relationship between the number of species represented by N individuals [$S(N)$] and N itself often followed log-normal distributions (Fig. 2). With N and N_{max} known, two additional assumptions are needed to estimate S. First, Preston's canonical hypothesis, stating that the most-abundant class of individuals coincides with that of the most-abundant species, has to hold. Second, the less-abundant species should be represented by a single individual ($N_{min} = 1$). If this is the case, then the number of species can be derived in the form shown by Curtis et al. (8). As an independent test, Locey and

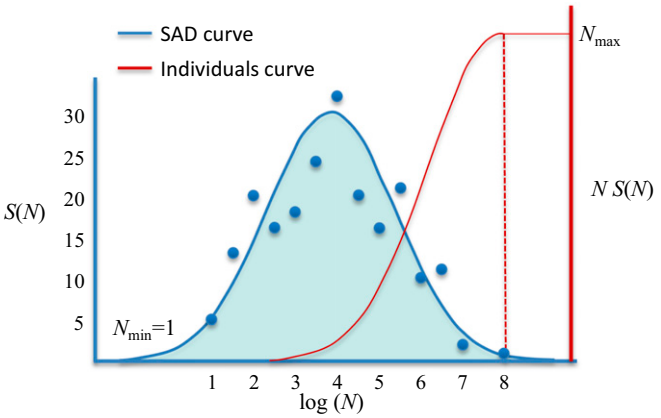

Fig. 2. Summary of the ecological theory used to estimate *S*. The species abundance distribution is a canonical log-normal (blue curve), with the most-abundant species yielding the maximum of the individuals curve. The less-abundant species is represented by a single individual. The area under the SAD (in pale blue) yields the diversity of taxa *S*. Blue circles correspond to data of bird species abundance (from ref. 7) and are intended as an illustration of a good fit between field data and the log-normal SAD.

Lennon (1) used their N_{max}-to-N relationship to estimate N_{max} for four large communities: the human gut ($N = 10^{14}$), the cow rumen ($N = 10^{15}$), the oceans ($N = 10^{29}$), and the whole Earth ($N = 10^{30}$), and then used the predicted values of N_{max} to validate their S-to-N relationship. Again, the predictions fit quite well the statistical expectations, giving additional strength to the estimates of diversity: 10^6, 10^7, 10^9, and 10^{12} taxa for the four previous communities, respectively.

There is much more in a particular function fitting the data than the possibility to predict unknown quantities, such as N_{max} or S. Mathematical relationships between variables speak for the mechanisms that underlie the observations and can discriminate between theories. It would be of interest to explore whether there are alternative fits to the power-law N vs. N_{max} or S vs. N functions. Even in cases where the functional relationship seemed incontestable in the past, such as the 3/4 power-law of metabolism, alternative fits suggested alternative theories to explain the origin of the pattern, not always suited for all organisms (9). An early explanation for log-normal distributions of abundance was that they arose from random demographic processes. However, distributions obtained in that way do not fulfill in general Preston's canonical hypothesis (10). Therefore, those processes are not sufficient to explain log-normal SADs. The shape and origin of SAD curves remains a controversial issue, with at least three different functional forms supported by data and dozens of models coexisting in the literature (11). A further step prompted by the present study (1) is to explore the effects of species abundance distributions different from the log-normal in the prediction of ecological diversity.

Actually, the same distribution does not necessarily have to be applicable to all living beings, and there is a great intellectual beauty in the difference. For example, although terrestrial or freshwater animal communities tend to follow log-normal SADs more often than log-series or power-law distributions, plant and marine communities tend to deviate from this terrestrial pattern by more often following log-series (marine communities) or power laws (plants) (12). Therefore, the differences observed by Locey and Lennon (1) between macro- and microorganisms in rarity and evenness may be quite revealing about basic differences in the ecology of these organisms.

The paper by Locey and Lennon (1) spurs efforts to understand the evolutionary origins of microbial diversity and brings back to the stage unsolved questions on the characterization and understanding of ecological communities. Wanting to know the global number of species and to uncover the mechanisms that explain ecosystem organization are not only intellectually attractive questions. These questions are relevant for taking appropriate conservation policies, are essential to map the potential living resources, to allocate funding and efforts to taxonomy, and to understand and compare ecosystems. Knowing the quantity and quality of the organisms that we share the Earth with also has symbolic and ethical implications (13). If the total number of microbial species on Earth is 10^{12}, the challenge ahead is formidable in its multiple facets.

1 Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. *Proc Natl Acad Sci USA* 113(21):5970–5975.
2 Erwin TL (1982) Tropical forests: Their richness in Coleoptera and other arthropod species. *Coleopt Bull* 36(1):74–75.
3 Costello MJ, May RM, Stork NE (2013) Can we name Earth's species before they go extinct? *Science* 339(6118):413–416.
4 Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the ocean? *PLoS Biol* 9(8):e1001127.
5 Pedrós-Alió C (2012) The rare bacterial biosphere. *Annu Rev Mar Sci* 4:449–466.
6 Preston FW (1948) The commonness, and rarity, of species. *Ecology* 29(3):254–283.
7 Trnka A (1997) *Current List of Birds of Slovakia* (Trnava Univ, Trnava, Slovakia).
8 Curtis TP, Sloan WT, Scannell JW (2002) Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* 99(16):10494–10499.
9 Glazier DS (2014) Metabolic scaling in complex living systems. *Systems* 2(4):451–540.
10 Sugihara G (1980) Minimal community structure: An explanation of species abundance patterns. *Am Nat* 116(6):770–787.
11 McGill BJ, et al. (2007) Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett* 10(10): 995–1015.
12 Ulrich W, Ollik M, Ugland KI (2010) A meta-analysis of species–abundance distributions. *Oikos* 119(7):1149–1155.
13 Caley MJ, Fisher R, Mengersen K (2014) Global species richness estimates have not converged. *Trends Ecol Evol* 29(4):187–188.