

# A macroecological theory of microbial biodiversity

William R. Shoemaker<sup>†</sup>, Kenneth J. Locey<sup>†\*</sup> and Jay T. Lennon

**Microorganisms are the most abundant, diverse and functionally important organisms on Earth. Over the past decade, microbial ecologists have produced the largest ever community datasets. However, these data are rarely used to uncover law-like patterns of commonness and rarity, test theories of biodiversity, or explore unifying explanations for the structure of microbial communities. Using a global scale compilation of >20,000 samples from environmental, engineered and host-related ecosystems, we test the power of competing theories to predict distributions of microbial abundance and diversity–abundance scaling laws. We show that these patterns are best explained by the synergistic interaction of stochastic processes that are captured by lognormal dynamics. We demonstrate that lognormal dynamics have predictive power across scales of abundance, a criterion that is essential to biodiversity theory. By understanding the multiplicative and stochastic nature of ecological processes, scientists can better understand the structure and dynamics of Earth's largest and most diverse ecological systems.**

A central goal of ecology is to explain and predict patterns of biodiversity across evolutionarily distant taxa and scales of abundance<sup>1–4</sup>. Over the past century, this endeavour has focused almost exclusively on macroscopic plants and animals, giving little attention to the most abundant and taxonomically, functionally, and metabolically diverse organisms on Earth: microorganisms<sup>1–4</sup>. However, global scale efforts to catalogue microbial diversity across environmental, engineered and host-related ecosystems have created an opportunity to understand biodiversity using a scale of data that far surpasses the largest macrobial datasets<sup>5</sup>. While commonness and rarity in microbial systems have become increasingly studied over the past decade, such patterns are rarely investigated in the context of unified relationships that are predictable under general principles of biodiversity.

One of the most frequently documented patterns of microbial diversity in recent years is the 'rare biosphere', which describes how the majority of taxa in an environmental sample are represented by few gene sequences<sup>6,7</sup>. While the rare biosphere has become a primary pattern of microbial ecology<sup>6–8</sup>, it also reflects the universally uneven nature of one of ecology's fundamental patterns, that is, the species abundance distribution (SAD)<sup>9</sup>. The SAD is among the most intensively studied patterns of commonness and rarity. Furthermore, the SAD is central to biodiversity theory and macroecology, which aims to understand patterns in abundance, distribution and diversity across scales of space and time<sup>9</sup>. However, microbiologists have largely overlooked the connection of the SAD to theories of biodiversity and macroecology, and the ability of some of those theories to predict other intensively studied patterns such as the species–area curve or distance–decay relationship<sup>10</sup>.

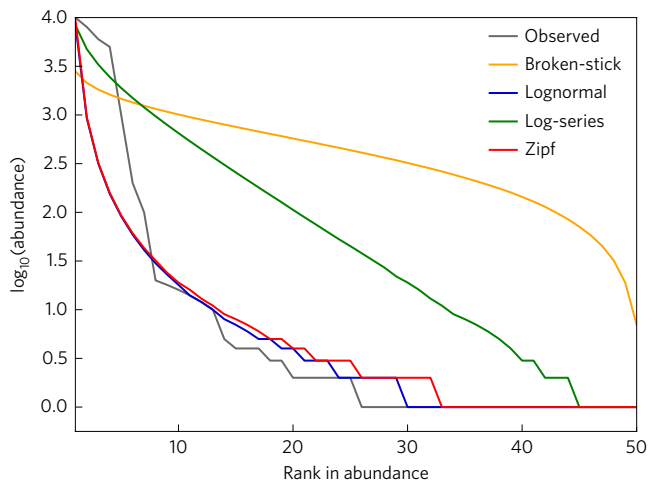
Since the 1930s, ecologists have developed more than 20 models that predict the SAD<sup>3</sup>. While some of these models are purely statistical and predict only the shape of the SAD, others encode the principles and mechanisms of competing theories<sup>2–4,9</sup>. Of all existing SAD models, none have been more successful than the distributions known as the lognormal and log-series, which often serve as standards against which other models are tested<sup>2</sup>. The lognormal is characterized by a right-skewed frequency distribution that becomes approximately normal under log-transformation; hence the name 'lognormal'. Historically, the lognormal is said to emerge from the multiplicative interactions of stochastic processes<sup>11</sup>. Examples of these lognormal dynamics include the multiplicative

nature of growth, the stochastic nature of population dynamics, and the energetic cost of individual dispersal across geographic distance. While most ecological processes are likely to have multiplicative interactions<sup>11</sup>, many theories of biodiversity (neutral theory, stochastic geometry, stochastic resource limitation theory) include a stochastic component<sup>2,12,13</sup>. Lognormal dynamics should become increasingly important for large communities, a result of the central limit theorem and law of large numbers<sup>11</sup>. Yet despite being one of the most successful models of the SAD among communities of macroorganisms, the lognormal does not seem to be predicted by any general theory of biodiversity and is used only rarely in microbial studies<sup>14–18</sup>.

Like the lognormal, the log-series has also been successful in predicting the SAD<sup>19</sup>. Although commonly used since the 1940s, the log-series is the form of the SAD that is predicted by one of the most recent, successful and unified theories of biodiversity, that is, the maximum entropy theory of ecology (METE)<sup>4</sup>. In ecological terms, METE states that the expected form of an ecological pattern is that which can occur in the greatest number of ways for a given set of constraints, that is, the principle of maximum entropy (MaxEnt)<sup>4,20</sup>. METE uses only the number of species ( $S$ ) and total number of individuals ( $N$ ) as its empirical inputs to predict the SAD. Using the most comprehensive global scale data compilations of macroscopic plants and animals, METE outperformed the lognormal and often explained >90% of variation in abundance within and among communities<sup>21,22</sup>. The success of METE has made the log-series the most highly supported model of the SAD<sup>4</sup>. But despite its success, METE has not been tested with microbial data and it is unknown whether it can predict microbial SADs, a crucial requirement for a macroecological theory of biodiversity<sup>23</sup>.

The lognormal, log-series and other models of biodiversity have competed to predict the SAD for several decades. However, few studies have gone beyond the SAD to test multiple models using several patterns of commonness and rarity. For example, recently discovered relationships show how aspects of commonness and rarity scale across as many as 30 orders of magnitude, from the smallest sampling scales of molecular surveys to the scale of all organisms on Earth<sup>5</sup>. Such scaling laws are among the most powerful relationships in biology, revealing how one variable (for example,  $S$ ) changes in a proportional way across orders of magnitude in another variable such as  $N$ . However, the mechanisms that give rise to these scaling

Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. <sup>†</sup>These authors contributed equally to this work. \*e-mail: [kjlocey@indiana.edu](mailto:kjlocey@indiana.edu)



**Figure 1 | Forms of predicted SADs in rank abundance form, that is, ordered from the most abundant species ( $N_{\max}$ ) to the least abundant on the x axis.** The grey line represents one SAD that was randomly chosen from our data. Each model was fit to the observed SAD; see Methods. The broken-stick is known to produce an overly even SAD. The log-series often explains SADs for plant and animal communities but has gone untested among microorganisms<sup>22</sup>. The Zipf distribution is a power-law model that produces one of the most uneven forms of the SAD, often predicting more singletons and greater dominance ( $N_{\max}$ ) than other models. Finally, the Poisson lognormal, a lognormal model with Poisson-based sampling error, tends to be similar to the unevenness of the Zipf distribution, but predicts more realistic  $N_{\max}$ . Importantly, each model used here predicts an SAD with the same richness of the observed SAD, which is often not the case in other studies that fail to use maximum likelihood expectations<sup>24</sup>.

laws were not reported and it remains to be seen whether any biodiversity theory can predict and unify them. It also remains to be seen whether the model that best predicts the SAD would also best explain how aspects of commonness and rarity scale with  $N$ .

In this study we ask whether the lognormal and log-series can reasonably predict microbial SADs, and whether either model can reproduce recently discovered diversity–abundance scaling relationships<sup>5</sup>. We used a compilation of 16S ribosomal RNA (rRNA) community-level surveys from more than 20,000 unique locations, ranging from glaciers to hydrothermal vents to hospital rooms. We contextualize the results of the lognormal and the log-series against two other well-known SAD models: one that predicts a highly uneven form (the Zipf distribution), and one that predicts a highly even form (the simultaneous broken-stick). As general theories of biodiversity should make accurate predictions regardless of the size of a sample, community, or microbiome, we tested whether the performance of these four long-standing models is influenced by sample abundance ( $N$ ), which is a primary constraint on the form of the SAD. We discuss our findings in the context of greater unification across domains of life, paradigms of biodiversity theory, and in the context of how lognormal dynamics may underpin microbial ecological processes.

## Results

**Predicting distributions of microbial abundance.** The lognormal explained nearly 94% of the variation within and among microbial SADs, compared with 91% for the Zipf distribution and 64% for log-series predicted by METE (Fig. 1, Table 1). In addition, the lognormal consistently had the highest corrected Akaike information criteria (AIC<sub>c</sub>) weight in a bootstrap analysis and was on average the best fitting model 57% of the time (Supplementary Fig. 6, Supplementary Table 6). The performance of the simultaneous

**Table 1 | Comparison of the performance of SAD models for microbial datasets.**

Model	Mean $r^2_m$	Standard error
Lognormal	0.94	0.0044
Zipf	0.91	0.0031
Log-series	0.64	0.014
Broken-stick	−0.32	0.034

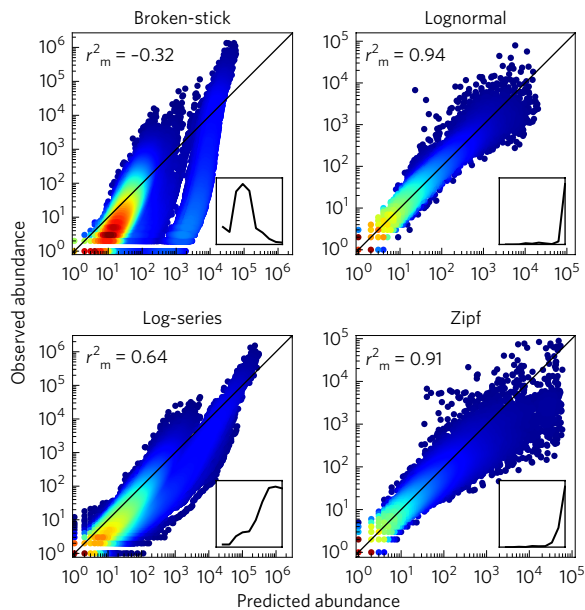
The mean site-specific modified  $r$ -square ( $r^2_m$ ) and standard error for each model from 10,000 bootstrapped samples of 200 SADs: broken-stick, the log-series predicted by METE, the lognormal and the Zipf power-law distribution. The lognormal and the Zipf provide the best predictions for how abundance varies among taxa. The lognormal and the Zipf are also characterized by lower standard errors than the broken-stick and the log-series.

broken-stick (hereafter referred to as the broken-stick) was too poor to be evaluated using the modified coefficient of determination ( $r^2_m$ ). Although close to the predictive power of the lognormal, the Zipf distribution greatly over-predicted the abundance of the most abundant taxa ( $N_{\max}$ ). In some cases, the predicted  $N_{\max}$  was greater than the empirical value for sample abundance ( $N$ ). The Zipf distribution was also sensitive to the exclusion of singleton operational taxonomic units (OTUs) and percent cutoff in sequence similarity (Supplementary Table 3, Supplementary Fig. 3). In this way, the Zipf reasonably predicts the abundance of intermediately abundant taxa, but often fails for the most dominant and rare taxa (Supplementary Tables 1 and 2)<sup>22,24</sup>. In contrast to the other models, the lognormal produced unbiased predictions for the abundances of dominant and rare taxa, regardless of cutoffs in percent similarity and the exclusion of singleton OTUs (Supplementary Figs 1 and 2, Supplementary Tables 1 and 2).

**Predictive power across scales of sample abundance.** The performance of SAD models across scales of  $N$  is rarely, if ever, examined. While the log-series has been successful among communities of macroscopic plants and animals<sup>21,22</sup>, for the vast majority of these samples  $N$  was less than a few thousand organisms<sup>21,22</sup>. In contrast, the log-series predicted by METE has yet to be tested using microbial data, that is, where  $N$  often represents millions of sampled 16S rRNA gene reads.

We found that the lognormal performed well across all orders of magnitude in  $N$ , with no indication of weakening at higher orders of magnitude. The performance of METE's log-series, however, was much more variable and often provided fits to microbial SADs that were too poor to interpret. As a result, the form of the SAD predicted by the most successful theory of biodiversity for macroorganisms (that is, METE) failed across orders of magnitude in microbial  $N$ . This was the case for SADs from different systems and within SADs that were resampled to smaller  $N$  (Fig. 2, Supplementary Fig. 3). While the Zipf distribution also provided reasonable fits that improved with increasing  $N$ , the broken-stick increasingly failed for greater  $N$ . This latter result supports previously documented patterns of decreasing species evenness with increasing  $N$ <sup>5,25</sup>; a trend that the lognormal captures without apparent bias.

**Diversity–abundance scaling laws.** Recently, aspects of taxonomic diversity have been shown to scale with  $N$  at rates similar to molecular surveys of microorganisms and individual counts of macroorganisms<sup>5</sup>. These aspects of diversity include dominance (the abundance of the most abundant OTU;  $N_{\max}$ ), evenness (similarity in abundance among OTUs) and rarity (concentration of taxa at low abundances). We found that the lognormal best reproduced these diversity–abundance scaling relationships<sup>5</sup> (Table 2, Fig. 3). While the Zipf approximated the rate at which  $N_{\max}$  scaled with  $N$ , it greatly over-predicted the  $y$ -intercept and, hence, the actual value



**Figure 2 | Relationships between predicted abundance and observed abundance.** All species of all examined SADs are plotted; hotter colours (such as red) reveal a greater density of species abundances. The black diagonal line is the 1:1 line, around which a perfect prediction would fall. Insets show histograms of the per-SAD modified  $r_m^2$  values from a range of zero to one, with left-skewed histograms suggesting a better fit of the model to the data. The value at the top-left of each sub-plot is the mean  $r_m^2$  value for 10,000 bootstrapped samples (see Methods). Each dot represents the observed abundance versus the predicted abundance for each species in the data.

of  $N_{\max}$  (Fig. 3). Additionally, neither the log-series predicted by METE nor the broken-stick came close to reproducing the observed diversity–abundance scaling relationships (Fig. 3, Table 2).

## Discussion

In this study, we asked whether widely known and successful models of biodiversity could predict microbial SADs and also unify SADs with recently discovered diversity–abundance scaling laws. We found that the lognormal provided the most accurate predictions for nearly all patterns in our study. This is in sharp contrast to studies of macroorganisms where the log-series distribution predicted by METE was overwhelmingly supported<sup>21,22</sup>. Such discrepancies in model performance suggest that there are fundamental differences between macroorganisms and microorganisms that point to the importance of lognormal dynamics. Specifically, that multiplicative processes (such as growth) and stochastic outcomes (such as population fluctuations) produce a central limiting pattern within large and heterogeneous communities where species partition multiple resources<sup>11</sup>. Instead of identifying a particular process (for example, dispersal limitation, resource competition) or a set of specific environmental factors (such as, pH, temperature), we propose that lognormal dynamics underpin the fundamental nature of microbial communities<sup>11,12</sup>.

There are fundamental differences in how ecologists study communities of microscopic and macroscopic organisms. As ecologists tend to sample microbial communities on spatial scales that greatly exceed the scales of their interactions, samples of microbial communities are likely to lump together many ecologically distinct taxa that do not partition the same resources or occupy the same microhabitats<sup>26</sup>. If microbial studies commonly lump together species that belong to different ecological communities, then this may lead to the emergence of a power-law SAD (for example, the Zipf)<sup>27</sup>.

**Table 2 | The lognormal comes closest to reproducing the scaling exponents of diversity–abundance scaling relationships<sup>5</sup>.**

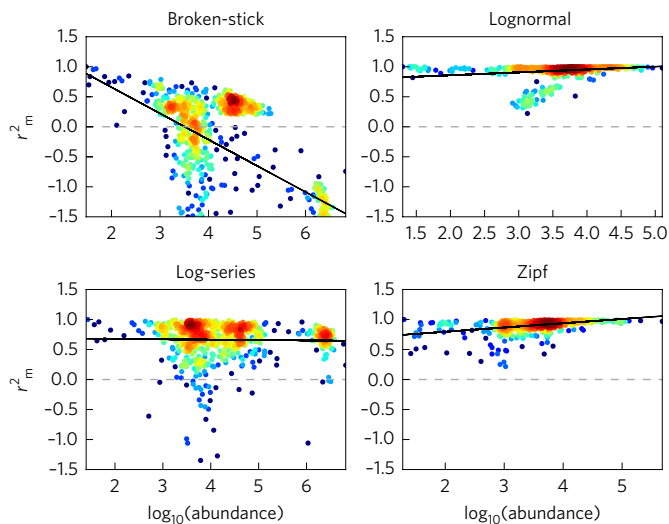
Model	Diversity metric	Slope	Difference (%)
Lognormal	$N_{\max}$	1.0	1.5
	Evenness	−0.48	42.0
	Skewness	0.10	23.0
Zipf	$N_{\max}$	1.0	0.28
	Evenness	−0.53	53.0
	Skewness	0.086	41.0
Log-series	$N_{\max}$	0.86	16.0
	Evenness	−0.16	66.0
	Skewness	0.048	92.0
Broken-stick	$N_{\max}$	0.73	32.0
	Evenness	−0.022	170.0
	Skewness	0.014	160.0

These scaling relationships pertain to absolute dominance ( $N_{\max}$ ), Simpson's metric of species evenness and skewness of the SAD. The percent difference is given between the scaling exponents predicted from each SAD model and the mean of the scaling exponents for the EMP, HMP, and MG-RAST reported in Table 1 of ref. <sup>5</sup>, that is, where the mean for  $N_{\max}$  was 1.0, the mean for evenness was −0.48, and the mean for skewness was 0.10.  $P < 0.0001$  for all scaling exponents.

We expect that the increasing performance of the Zipf with greater  $N$  is evidence of a power-law SAD arising from the mixture of lognormal microbial communities. Although the connection between the lognormal and the Zipf needs further study, a macroecological theory of microbial biodiversity should allow for this dynamic.

In our macroecological study, we used data that microbial ecologists have collected and made available (16S rRNA sequences). As is well known, these amplicon-based data may contain artefacts that potentially affect the shape of microbial SADs. We accounted for some of these artefacts by testing for the effects of the sequence similarity percent cut-off used to cluster OTUs, as well as the influence of singletons and sample size. There are also caveats that we could not address. First, we used the post-processed sequence abundances and OTU classifications provided in publicly available datasets<sup>5,25</sup>. While different methods for processing amplicon reads may influence the shape of the SAD, we did not have the resources and computational capacity to re-process all of the raw sequence data from our various datasets. Second, the number of sampled 16S rRNA sequences is not equivalent to the number of cells in a sample. Instead of assuming this equivalency, we assumed that SADs based on 16S rRNA sequences are similar to those based on organismal abundances, as is common in microbial community studies. Future studies may reveal whether or not this assumption is justified.

Finally, in rejecting the log-series as a model for microbial SADs, we are not rejecting METE altogether. We are instead rejecting the log-series as METE's primary form of the SAD<sup>4</sup>. In fact, METE seems capable of predicting both the lognormal and the Zipf<sup>28</sup>. This is because in using METE, one tries to infer the most likely form of an ecological pattern for a particular set of variables such as  $N$  and  $S$  and constraints such as  $N/S$ . Consequently, the forms of ecological patterns predicted by METE could change depending on the constraints and state variables used<sup>28</sup>. For example, METE predicts that the SAD is a power law if it constrains the SAD to  $N/S$  while including a resource variable<sup>28</sup>. However, METE has not been developed to predict forms of the SAD other than the log-series and it remains to be seen whether METE can predict the form of the lognormal (that is, Poisson lognormal) used in our study. If so, and if it can reconcile why a log-series SAD works best for macroorganisms and a lognormal works best for microorganisms, then METE may indeed be a unified theory of biodiversity. Until then, microbial communities



**Figure 3 | The relationship of model performance to the total number of 16S rRNA reads ( $N$ ).** The modified coefficient of determination ( $r^2_m$ ) is the variation in the observed SAD that is explained by the predicted SAD (as in Fig. 2). The performance of the broken-stick model and of the log-series distribution predicted by the maximum entropy theory of ecology (METE) decreases for greater  $N$ . With the exception of a small group of points, the lognormal provides  $r^2_m$  values of 0.95 or greater across scales of  $N$ . The Zipf provides a better explanation of microbial SADs with increasing  $N$ . The grey dashed horizontal line is placed where the  $r^2_m$  equals zero. The  $r^2_m$  can take negative values because it does not represent a fitted relationship, that is, the  $y$ -intercept is constrained to 0 and the slope is constrained to 1. Results from the simple linear regression can be found in Supplementary Table 1.

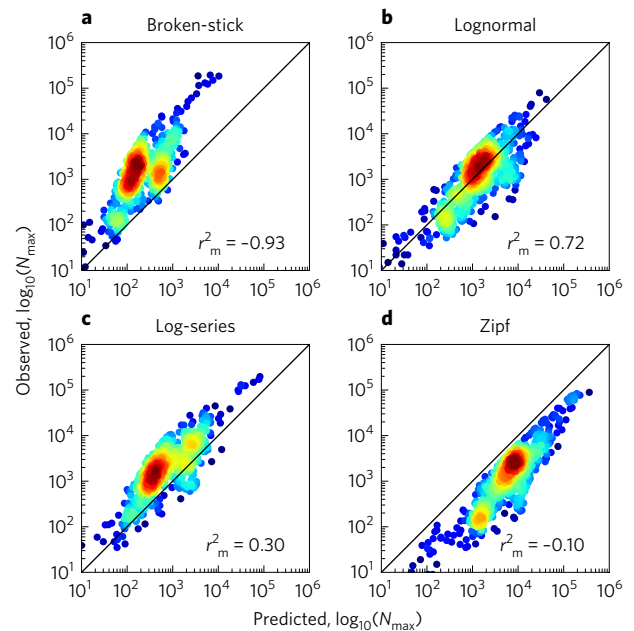
and microbiomes seem to be shaped by the multiplicative interactions of stochastic processes that, although highly complex, inevitably lead to predictable patterns of biodiversity.

## Methods

**Data.** We used one of the largest compilations of microbial community and microbiome data to date, consisting of bacterial and archaeal community sequence data from over 20,000 unique geographic sites. These data were compiled in a previous study<sup>5</sup> and include 14,962 sites from the Earth Microbiome Project (EMP)<sup>29</sup>, 4,303 sites from the Data Analysis and Coordination Center (DACC) for the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP)<sup>30</sup> and 1,319 non-experimental sequencing projects consisting of processed 16S rRNA amplicon reads from the Argonne National Laboratory metagenomics server MG-RAST<sup>31</sup>. All sequence data were previously processed using established pipelines to remove low-quality sequence reads and chimeras<sup>29–31</sup>. Additional information pertaining to the datasets can be found in the Supplementary Information and in previous studies<sup>5</sup>.

To assess the effect of sequence similarity on the fit of SAD models we analysed the same collection of MG-RAST data with different percent cutoffs. This collection was analysed at minimum percent sequence similarities of 95, 97 and 99% to the closest reference sequence in MG-RAST's M5 rRNA database, with a maximum  $e$  value (probability of observing an equal or better match in a database of a given size) of  $10^{-5}$  and a minimum alignment length of 50 base pairs<sup>32–37</sup>. As we did not have the computational capacity or resources to reclassify all sequences from all samples of each project, we used the sampled sequence abundances and OTU classifications provided in each study, as was done in similar scale studies of large microbial data compilations<sup>5,25</sup>. In addition, we assume that the criterion for an OTU holds across microbial taxonomic groups, though we acknowledge this as a potential source of error.

Similar to the majority of SAD studies, we cannot confirm that our data are representative random samples of their respective environments. While systemic methodological artefacts across a few studies of particular environments could produce systemic biases, our study rests on an implicit macroecological assumption, that many thousands of independently gathered data points from a diversity of studies and methods are unlikely to produce the



**Figure 4 | Predictions of absolute dominance (the greatest species abundance within an SAD,  $N_{max}$ ) using the dominance scaling relationships of each model (Table 1).** a–d, Owing to the negative  $r^2_m$  values for the broken-stick (a) and the Zipf (d), only the lognormal (b) and the log-series (c) are capable of providing meaningful predictions of  $N_{max}$ .

same artefact. Though 16S rRNA amplicon sequencing has several limitations, makes assumptions that are likely to be unrealistic (for example, several ecologically distinct taxa may be clustered as a single OTU) and is inherently limited by the fact that the number of 16S sequences is not equal to the number of individuals in a community, it is still one of the most widely used methods in microbial ecology and is regularly used to examine the structure and composition of natural and man-made microbiomes. Additional information pertaining to the datasets can be found elsewhere<sup>5</sup>.

**Description of SAD models.** In this study we ask whether the lognormal, log-series and two other classic SAD models that have some success in microbial ecology (the simultaneous broken-stick<sup>12</sup> and the Zipf distribution<sup>38,39</sup>), can reasonably predict microbial SADs (Fig. 4). We evaluated the performance of each model with and without singletons and across different percent cutoffs for sequence similarity used to cluster 16S rRNA reads into OTUs.

**Lognormal.** To avoid fractional abundances and to account for sampling error, we used a Poisson-based sampling model of the lognormal, which is known as the Poisson lognormal<sup>40</sup>. We used the maximum likelihood estimate of the Poisson lognormal as our species abundance model of lognormal dynamics. The likelihood estimate of the single composite parameter  $\lambda$  (composed of two parameters, the mean ( $\mu$ ) and standard deviation ( $\sigma$ )) of the Poisson lognormal is derived via numerical maximization of the likelihood surface<sup>40</sup>. Once  $\lambda$  is found, the probability that a randomly chosen species is represented by  $n$  individuals ( $p(n)$ ) under the Poisson lognormal (hereafter lognormal) is derived using:

$$p(n) = \int_0^{\infty} \frac{\lambda^n e^{-\lambda}}{n} p_{LN(\lambda, \sigma)} d\lambda$$

where  $p_{LN}$  is the lognormal probability.

Although the validity of the lognormal has previously been criticized as an appropriate null model for SADs, we chose to use the lognormal in our study for four reasons<sup>41</sup>. First, the primary source of criticism was based on results from a handpicked dataset of only three macrobial SADs. Second, the authors of the primary criticism did not use the Poisson lognormal described here, which has long been the recommended form. Third, the authors hold against the lognormal the historical shortcomings of fitting SADs by eye, an issue that is not relevant here, as we have designed our study to be quantitative and replicable. Finally, rather than use the lognormal as a null model, we frame the lognormal as capturing the multiplicative and stochastic nature of microbial community dynamics.

**METE.** METE uses only two empirical inputs to predict the SAD: species richness ( $S$ ) and total abundance ( $N$ ) of individuals (or sequence reads) in a sample.

To predict the SAD, METE assumes that the expected shape of the SAD is that which can occur in the highest number of ways, an assumption based on the principle of MaxEnt<sup>20</sup>. Using METE, the shape of the SAD was predicted by calculating the probability ( $\Phi$ ) that the abundance of a species is  $n$  given  $S$  and  $N$ :

$$\Phi(n|S, N) = \frac{1}{\log(\beta^{-1})} \frac{e^{-\beta n}}{n}$$

where  $\beta$  is a fitted energetic parameter representing the community-scaled metabolic rate.  $\beta$  is derived using numerical optimization and the following equation:

$$\frac{N}{S} = \frac{\sum_{n=1}^N e^{-\beta n}}{\sum_{n=1}^N e^{-\beta n} / n}$$

where  $N/S$  is the average abundance among species. This approach to predicting the MaxEnt form of the SAD yields the log-series distribution<sup>4,19</sup>.

It is worth noting that METE yields a Zipf distribution with an exponential cutoff if a resource in addition to energy (that is ideally energy-independent) is allocated across individuals (for example, space)<sup>28,42</sup>. Given that microbial cell density can be extremely high in certain systems (for example, a gram of soil can contain  $10^{10}$  cells), it is possible that the inclusion of space as a resource, measured in appropriate units, would increase the performance of METE for microbial systems. Unfortunately, there is little available microbial spatial data. In addition, METE can potentially predict the lognormal<sup>42</sup>. However, neither of these alternate forms of METE have been sufficiently developed or tested.

**Broken-stick.** The broken-stick model predicts a high similarity in abundance among species and, hence, predicts one of the most even SADs of any model. The broken-stick model predicts the SAD as the simultaneous breaking of a stick of length  $N$  at  $S - 1$  randomly chosen points<sup>12</sup>. The broken-stick also has a purely statistical equivalent, that is, the geometric distribution<sup>43,44</sup>, where  $f(k)$  is the probability mass function for  $k$  trials with probability  $p$  of success:

$$f(k) = (1 - p)^{k-1} p$$

The broken-stick has no free parameters and predicts only one form of the SAD for a given combination of  $N$  and  $S$ . Although rarely recognized, METE predicts a geometric distribution form of the SAD if energy (that is,  $\beta$  from the log-series) is not included as a state variable, with the constraint arising from an empirical value for the ratio  $N/S$ . The geometric distribution is a maximum entropy solution when  $N$  and  $S$  are the only state variables, with the constraint arising from the empirical value of the ratio  $N/S$ .

**Zipf distribution.** The Zipf distribution (also known as the discrete Pareto distribution) is a power-law model that predicts one of the most uneven forms of the SAD. This distribution is based on a power-law of the frequency of ranked data and is characterized by one parameter ( $\gamma$ ), where the frequency of the  $k$ th rank is inversely proportional to  $k$ :  $p(k) \approx k^{-\gamma}$ , with  $\gamma$  often ranging between  $-1$  and  $-2$ <sup>38,45-47</sup>. The Zipf distribution predicts the frequency of elements of rank  $k$  out of  $N$  elements with parameter  $\gamma$  as:

$$f(k; \gamma, N) = \frac{1/k^\gamma}{\sum_{n=1}^N (1/n^\gamma)}$$

We calculated the maximum likelihood estimate of  $\gamma$  using numerical maximization, which was then used to generate the predicted form of the SAD.

**Testing SAD predictions.** Our SAD predictions were based on the rank abundance form of the SAD. This form is a vector of species abundances ranked from most to least abundant (Fig. 4). As the predicted form of each model preserves the number of observed species ( $S$ ), we were able to directly compare (rank-for-rank) the observed and predicted SADs using regression analysis to find the percent of variation in abundance among species that is explained by each model. We generated the predicted forms of the SAD using previously developed code<sup>21</sup> (<https://github.com/weecology/white-et-al-2012-ecology>) and the public repository macroecotools (<https://github.com/weecology/macroecotools>).

To prevent bias in our results due to the overrepresentation of a particular dataset, we performed 10,000 bootstrap iterations using a sample size of 200 SADs drawn randomly from each dataset. The sample size was determined based on the number of SADs that the numerical estimator used to generate the Zipf distribution was able to solve for the smallest dataset (239 SADs from MG-RAST). This was necessary because numerical optimization can fail to arrive at a maximum likelihood solution (or take an exhaustively long time) for the parameter(s) of a given model. We then calculated the modified coefficient of

determination around the 1:1 line (as per previous tests of METE<sup>21,25,47</sup>) with the following equation:

$$r_m^2 = 1 - \frac{\sum (\log(\text{obs}_i) - \log(\text{pred}_i))^2}{\sum (\log(\text{obs}_i) - \log(\text{obs}_i))^2}$$

where  $\text{obs}_i$  and  $\text{pred}_i$  represent the observed and predicted abundance of the  $i$ th species, respectively. It is possible to obtain negative  $r_m^2$  values because the relationship is not fitted, but instead is performed by estimating the variation around the 1:1 line with a constrained slope of 1.0 and a constrained intercept of 0.0 (refs. <sup>21,25,47</sup>). Furthermore, we performed an extensive analysis using previously established methods<sup>22</sup> where we compared the fits of all four models using AICc weights, correcting both for the number of species observed in each site and the number of fitted parameters in each model. We then selected the model with the largest AICc weight as the best fitting model for that particular site. To prevent bias due to the overrepresentation of a particular dataset, we performed the same bootstrap analysis that was done for the  $r_m^2$ . We have provided the mean, standard deviation and kernel density estimates of the log-likelihood and parameter values for all models that contain a free parameter (Supplementary Table 5, Supplementary Fig. 5).

**Diversity–abundance scaling relationships.** To determine whether the SAD models tested here can explain previously reported diversity–abundance scaling relationships<sup>5</sup>, we first calculated the values of  $N_{\text{max}}$ , Simpson's measure of species evenness and the log-modulo of skewness as a measure of rarity derived from predicted SADs of each model, as in ref. <sup>5</sup>. We examined these diversity metrics against the values of  $N$  in the observed SADs. We used simple linear regression on log-transformed axes to quantify the slopes of the scaling relationships, which become scaling exponents when axes are arithmetically scaled, that is,  $\log(y) = z \log(x)$  is equivalent to  $y = x^z$ , where  $z$  is the slope and scaling exponent. These scaling exponents were compared to the reported exponents<sup>5</sup>. We calculated the percent difference between the diversity metrics reported by each SAD model and the mean of the exponents reported for the EMP, HMP and MG-RAST datasets.

We could not assess the ability of the SAD models to predict the scaling relationship of  $S$  to  $N$ , as in ref. <sup>5</sup>. This was because all of the SAD models used in our study return SADs with the same value of  $S$  as the empirical form.

**Influence of total abundance on model performance.** We used ordinary least-squares regression to assess the relationship between the performance of each SAD model and the number of sequences in a given sample ( $N$ ). While the aim of our study was to capture the influence of  $N$  on SAD model performance, we also rarefied within SADs. We performed bootstrapped resampling on rarefied sets of SADs to determine the influence of subsampled  $N$  on model performance. This bootstrap sampling procedure consisted of sampling SADs at given fractions of sample  $N$  and then calculating the mean  $r_m^2$ , repeating 100 times for each model. SADs were sampled at 50, 25, 12.5, 6.25, 3.125 and 1.5625% of sample  $N$ . This subsampling analysis was computationally exhaustive and required SADs with  $N$  large enough to be halved six times and still large enough to be analysed with SAD models. Likewise, we used only SADs for which predictions from each SAD model could be obtained at each scale of subsampled  $N$ . Altogether, we were able to use ten SADs that met these criteria.

**Computing code.** We used open source computing code to obtain the maximum-likelihood estimates and predicted forms of the SAD for the broken-stick, the lognormal, the prediction of METE (log-series distribution), and the Zipf distribution (<https://github.com/weecology/macroecotools>, <https://github.com/weecology/METE>). This is the same code used in studies that showed support for METE among communities of macroscopic plants and animals<sup>22-24</sup>. All analyses can be reproduced or modified for further exploration by using the code, data and directions provided here: <https://github.com/LennonLab/MicrobialBiodiversityTheory>.

**Data availability.** All data used in this study can be found in the public GitHub repository MicrobialBiodiversityTheory (<https://github.com/LennonLab/MicrobialBiodiversityTheory>).

Received 15 July 2016; accepted 30 January 2017; published 3 April 2017

## References

- Brown, J. H., Mehlman, D. W. & Stevens, G. C. Spatial variation in abundance. *Ecology* **76**, 1371–1382 (1995).
- Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, 2001).
- McGill, B. J. Towards a unification of unified theories of biodiversity. *Ecol. Lett.* **13**, 627–642 (2010).
- Harte, J. *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics* (Oxford Univ. Press, 2011).
- Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl Acad. Sci. USA* **113**, 5970–5975 (2016).

6. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored “rare biosphere.” *Proc. Natl Acad. Sci. USA* **103**, 12115–12120 (2006).
7. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat. Rev. Microbiol.* **13**, 217–229 (2015).
8. Reid, A. & Buckley, M. *The Rare Biosphere: A Report from the American Academy of Microbiology* (American Academy of Microbiology, 2011).
9. McGill, B. J. *et al.* Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol. Lett.* **10**, 995–1015 (2007).
10. Horner-Devine, M. C., Lage, M., Hughes, J. B. & Bohannan, B. J. M. A taxa-area relationship for bacteria. *Nature* **432**, 750–753 (2004).
11. Putnam, R. *Community Ecology* (Chapman & Hall, 1993).
12. MacArthur, R. On the relative abundance of species. *Am. Nat.* **94**, 25–36 (1960).
13. Sih, A., Englund, G. & Wooster, D. Emergent impacts of multiple predators on prey. *Trends Ecol. Evol.* **13**, 350–355 (1998).
14. Dunbar, J., Barns, S. M., Ticknor, L. O. & Kuske, C. R. Empirical and theoretical bacterial diversity in four Arizona soils. *Appl. Environ. Microbiol.* **68**, 3035–3045 (2002).
15. Curtis, T. P., Sloan, W. T. & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. USA* **99**, 10494–10499 (2002).
16. Bohannan, B. J. M. & Hughes, J. New approaches to analyzing microbial biodiversity data. *Curr. Opin. Microbiol.* **6**, 282–287 (2003).
17. Schloss, P. D. & Handelsman, J. Status of the microbial census. *Microbiol. Mol. Biol. Rev.* **68**, 686–691 (2004).
18. Pedrós-Alió, C. & Manrubia, S. The vast unknown microbial biosphere. *Proc. Natl Acad. Sci. USA* **113**, 6585–6587 (2016).
19. 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. *J. Anim. Ecol.* **12**, 42–58 (1943).
20. Jaynes, E. T. *Probability Theory: The Logic of Science* (Cambridge Univ. Press, 2003).
21. White, E. P., Thibault, K. M. & Xiao, X. Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology* **93**, 1772–1778 (2012).
22. Baldrige, E., Harris, D. J., Xiao, X. & White, E. P. An extensive comparison of species-abundance distribution models. *PeerJ* **4**, e2823 (2016).
23. McGill, B. Strong and weak tests of macroecological theory. *Oikos* **102**, 679–685 (2003).
24. Ulrich, W., Ollik, M. & Uglund, K. I. A meta-analysis of species-abundance distributions. *Oikos* **119**, 1149–1155 (2010).
25. Locey, K. J. & White, E. P. How species richness and total abundance constrain the distribution of abundance. *Ecol. Lett.* **16**, 1177–1185 (2013).
26. Fierer, N. & Lennon, J. T. The generation and maintenance of diversity in microbial communities. *Am. J. Bot.* **98**, 439–448 (2011).
27. Allen, A. P., Li, B. & Charnov, E. L. Population fluctuations, power laws and mixtures of lognormal distributions. *Ecol. Lett.* **4**, 1–3 (2001).
28. Harte, J. & Newman, E. Maximum information entropy: a foundation for ecological theory. *Trends Ecol. Evol.* **29**, 384–389 (2014).
29. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth microbiome project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).
30. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804–810 (2007).
31. Meyer, F. *et al.* The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
32. Flores, G. E. *et al.* Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environ. Microbiol.* **13**, 2158–2171 (2011).
33. Wang, J. *et al.* Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J.* **7**, 1310–1321 (2013).
34. Chu, H. *et al.* Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ. Microbiol.* **12**, 2998–3006 (2010).
35. Fierer, N. *et al.* Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J.* **6**, 1007–1017 (2012).
36. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
37. Amend, A. S., Seifert, K. A., Samson, R. & Bruns, T. D. Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *Proc. Natl Acad. Sci. USA* **107**, 13748–13753 (2010).
38. Gans, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**, 1387–1390 (2005).
39. Dumbrell, A. J., Nelson, M., Helgason, T., Dytham, C. & Fitter, A. H. Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME J.* **4**, 337–345 (2010).
40. Magurran, A. E. & McGill, B. J. *Biological Diversity Frontiers in Measurement and Assessment* (Oxford Univ. Press, 2011).
41. Williamson, M. & Gaston, K. J. The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *J. Anim. Ecol.* **74**, 409–422 (2005).
42. Harte, J., Zillio, T., Conlisk, E. & Smith, A. B. Maximum entropy and the state-variable approach to macroecology. *Ecology* **89**, 2700–2711 (2008).
43. Cohen, J. E. Alternate derivations of a species-abundance relation. *Am. Nat.* **102**, 165–172 (1968).
44. Heip, C. H. R., Herman, P. M. J. & Soetaert, K. Indices of diversity and evenness. *Oceanis* **24**, 61–87 (1998).
45. Zipf, G. K. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).
46. Newman, M. E. J. Power laws, Pareto distributions and Zipf’s law. *Contemp. Phys.* **46**, 323–351 (2005).
47. Xiao, X., McGlinn, D. J. & White, E. P. A strong test of the maximum entropy theory of ecology. *Am. Nat.* **185**, 70–80 (2015).

## Acknowledgements

We thank J. Gilbert and S. Gibbons for providing EMP data and guidance on using it. We also thank the researchers who collected, sequenced and provided metagenomic data on MG-RAST, as well as the individuals who maintain and provide the MG-RAST service. We also acknowledge the researchers who provided the open-source code for conducting some of our analyses. Finally, we thank the HMP Consortium for providing their data on the NIH’s publicly accessible DACC server. This work was supported by a National Science Foundation Dimensions of Biodiversity Grant (no. 1442246 to J.T.L. and K.J.L.) and the US Army Research Office (W911NF-14-1-0411 to J.T.L.).

## Author contributions

W.R.S. and K.J.L. conceived, designed and performed the experiments, analysed the data and contributed materials/analysis tools. W.R.S., K.J.L. and J.T.L. wrote the paper.

## Additional information

**Supplementary information** is available for this paper.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Correspondence and requests for materials** should be addressed to K.J.L.

**How to cite this article:** Shoemaker, W. R., Locey, K. J. & Lennon, J. T. A macroecological theory of microbial biodiversity. *Nat. Ecol. Evol.* **1**, 0107 (2017).

## Competing interests

The authors declare no competing financial interests.