A unifying ecological theory of microbial biodiversity

William R. Shoemaker[1*], Kenneth J. Locey[1*], Jay T. Lennon[1]

[1]Department of Biology, Indiana University, Bloomington, IN 47405 USA

[*]Authors contributed equally to the study

Correspondence: K Locey, Department of Biology, Indiana University, 261 Jordan Hall,

1001 East 3rd Street, Bloomington, IN 47405 USA. E-mail: kjlocey@indiana.edu

Conflict of interest

The authors declare no conflict of interest.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1450v3 | CC BY 4.0 Open Access | rec: 8 Jun 2016, publ: 8 Jun 2016

Abstract

An ecological theory of microbial biodiversity has yet to be developed. This shortcoming leaves patterns of abundance, distribution, and diversity for the most abundant and diverse organisms on Earth without a predictive framework. However, because of their high abundance and complex dynamics, microbial communities may be underpinned by lognormal dynamics, i.e., synergistic interactions among complex stochastic variables. Using a global-scale compilation of 20,456 sites from a diverse set of natural and host-related environments, we test whether a lognormal model predicts microbial distributions of abundance and diversity-abundance scaling laws better than other well-known models, including the most successful macroecological theory of biodiversity, i.e., maximum entropy theory of ecology. We found that the lognormal explains the greatest percent variation in abundance, that the success of the lognormal increased with abundance while other models decreased, and that the lognormal was the only model to reproduce recently documented diversity-abundance scaling laws. Our unifying ecological theory of microbial biodiversity explains and predicts macroecological patterns based on dynamics that capture the complex large number dynamics of microbial life.

Introduction

A central goal of ecology is to discover, predict, and unify patterns of biodiversity across evolutionarily distant taxa and scales of space, time, and abundance (Brown, 1995; Hubbell, 2001; McGill, 2010; Harte, 2011). The aim of this goal is to provide a coherent understanding of the dynamics that shape biodiversity and the means to predict the assembly, structure, and response of ecological communities. Over the past century, this endeavor has disproportionately focused on macroscopic plants and animals, giving little attention to the most abundant and taxonomically, functionally, and metabolically diverse organisms on Earth, i.e., microorganisms. However, this trend is beginning to change as global-scale sampling efforts and repositories of molecular survey data allow studies of microbial biodiversity to rival or surpass the scale of the largest macroecological datasets (e.g., Locey and Lennon, 2016). Yet patterns of abundance, distribution, and diversity in microbial systems are rarely studied as relationships that are predictable and unified under principles of biodiversity theory.

Studies of microbial biodiversity have documented patterns of commonness and rarity across space, time, and taxa for over a decade (e.g., Horner-Devine *et al.*, 2004; Sogin *et al.*, 2006; Locey and Lennon, 2016). Among these, the tendency for most taxa to account for a minority of relative abundance, i.e., the microbial "rare biosphere" is the most ubiquitous (Sogin *et al.,* 2006; Lynch and Neufeld, 2015). While the rare biosphere is primarily studied with respect to the biology of microorganisms (Reid and Buckley, 2011), the underlying pattern reflects the universally uneven nature of the species abundance distribution (SAD) (McGill *et al*., 2007). Nearly 20 theories of biodiversity have been developed in attempt to predict the SAD. Of these, the more general theories

predict additional patterns that are well documented in microbial ecology. For example, the decreasing similarity in compositional diversity with increasing geographic distance (Wang *et al.,* 2013; Zinger *et al*., 2014) and the rate at which taxa are discovered with increasing area (Horner-Devine *et al*., 2004; Green and Bohannan, 2006). These patterns often differ between microbes and macrobes, yet microbial ecologists rarely ask whether these differences support different ecological theories, whether any standing theories explain multiple patterns of microbial biodiversity, and what new predictions could be made by knowing how patterns of microbial biodiversity are related.

Microbial ecologists have used classic models of biodiversity for over a decade to predict the SADs of microorganisms (e.g., Dunbar *et al*., 2002; Gans *et al.,* 2005). These models include the Broken-stick, lognormal, Zipf, and log-series (Fig 1). Among these, the lognormal has been the most widely used and discussed (Curtis *et al*., 2002; Dunbar *et al*., 2002; Bohannan and Hughes, 2003; Schloss and Handelsman, 2006; Pedrós-Alió and Manrubia, 2016). The lognormal is underpinned by multiplicative interactions and stochastic processes, both of which characterize population, community, and trophic dynamics (MacArthur, 1960; Sih *et al*., 1998; Hubbell, 2001). In terms of the SAD, multiplicative interactions of random variables produce right-skewed histograms of species abundances that are approximately normal under log-transformation, hence the name "lognormal" (May, 1975). This outcome becomes increasingly likely for large communities where species partition multiple resources, a result of the central limit theorem and law of large numbers (Putnam, 1993). In this way, larger more heterogeneous communities should conform to a lognormal distribution. Additionally, the lognormal is the only biodiversity model that has been modified to estimate the number

of microbial species from local to global scales (Curtis *et al.,* 2002). Altogether, the lognormal appears to be an appropriate model on which a macroecological theory of microbial biodiversity could be developed.

In contrast to the frequent use and success of the lognormal model in microbial ecology, an expansive and unifying ecological theory has recently proven to be overwhelmingly successful in predicting several patterns of biodiversity among macroscopic plants and animals. The maximum entropy theory of ecology (METE) holds that the expected forms of ecological patterns are those that can occur in the greatest number of ways for a given set of constraints (Harte, 2011). Using the number of species (S) and total number of individuals (N) as the only empirical inputs, METE often explains $\geq$ 90 % of variation in abundance within and among communities of plants and animals (White *et al.,* 2012; Baldridge *et al.,* 2015). This success has made METE the most highly supported model of the species abundance distribution (SAD), the central pattern from which other patterns are predicted, e.g., species-area relationship, distance-decay relationship, spatial-abundance distribution, (Harte, 2011; Xiao *et al.,* 2015). Though METE is a relatively young theory, the form of the SAD that METE predicts is actually the log-series distribution, one of the oldest and most successful SAD models in ecology (Fisher *et al.,* 1943). Despite its successes, METE has not been tested on microbial community or microbiome datasets.

Here, we test whether lognormal dynamics explain microbial SADs and diversity-abundance scaling relationships better than METE and two other classic SAD models that have seen some success in microbial ecology, i.e., the Broken-stick model and the Zipf distribution (Gans *et al.,* 2005; Dumbrell *et al.,* 2010) (Fig 1). We also test the hypothesis

that the explanatory power of the lognormal increases with *N*, reflecting the tendency for large and heterogeneous communities to resemble the lognormal distribution (Putnam, 1993). In addition, we examine whether the lognormal is able to predict additional patterns of biodiversity by testing its ability, as well as that of other models, to reproduce the empirical diversity-abundance scaling relationships that have been recently described (Locey and Lennon, 2016). We conduct these tests using the largest compilation of microbial community data to date, including the Earth Microbiome Project, the Human Microbiome Project, and molecular surveys downloaded from MG-RAST. We discuss our findings in the context of a lognormal theory of microbial biodiversity and the tendency for microbial communities and microbiomes to exemplify lognormal dynamics.

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 15,535 unique geographic sites. These data were compiled in a previous study (i.e., Locey and Lennon, 2016) and include 14,962 sites from the Earth Microbiome Project (EMP) (Gilbert *et al.*, 2014), 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) (Turnbaugh *et al*., 2007), as well as 1,319 non-experimental sequencing projects consisting of processed 16S rRNA amplicon reads from the Argonne National Laboratory metagenomics server MG-RAST (Meyer *et al.,* 2008). Additional information pertaining to the datasets can be found elsewhere (Locey and Lennon, 2016).

A common convention in lieu of traditional species classification for microbial community sequence data is to cluster 16S rRNA amplicon reads into Operational Taxonomic Units (OTUs) based on a sequence similarity cutoff. It has been previously demonstrated that the cutoff for percent sequence similarity (95%, 97%, 99%) in determining taxonomic units does not change the general shape of the SAD (Locey and White, 2013). However, it is less common for investigators to evaluate how the percent cutoff affects the fit of SAD models (Woodcock *et al.*, 2007; Dumbrell *et al.*, 2010). To assess the effect of sequence similarity on the fit of SAD models we analyzed the same collection of MG-RAST data with different percent cutoffs. This collection was analyzed 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, and (Flores *et al.,* 2011; Wang, *et al.*, 2014; Chu *et al.*, 2010; Fierer *et al.,* 2012; Yatsunenko *et al.,* 2012; Amend *et al.*, 2010).

Description of SAD models

In this section, we provide a general overview of the different SAD models and how they were used in our analyses.

Lognormal — The lognormal distribution arises as a consequence of the central limit theory and the multiplicative interaction of random variables. It is one of the most popular models of species abundance. The general shape of the lognormal can be envisioned as right skewed frequency distribution with an internal mode. However,

because the lognormal is a continuous distribution, it has the undesirable characteristic of allowing fractional abundances, e.g., a species with 1.5 individuals. Instead of using the canonical lognormal of Preston (1948), modern macroecology studies use a Poisson-based sampling model of the lognormal, i.e., the Poisson lognormal (Magurran and McGill, 2007). The Poisson lognormal assumes that the true species abundance distribution for the community is lognormal but that sampling errors prevent samples from being truncated versions of the entire community. This sampling error is captured by assuming a Poisson random sampling of individuals from a lognormal community, which produces integer value abundances while accounting for sampling error. By accounting for variance due to sampling, the Poisson lognormal solves the issue of applying Preston's lognormal to empirical SADs (see Pedrós-Alió and Manrubia, 2016)

We used the maximum likelihood estimate of the Poisson lognormal as our species abundance model of lognormal dynamics. The likelihood estimate of the parameter (λ) of the Poisson lognormal must be derived using numerical maximization of the likelihood surface and can be computationally intensive (Magurran and McGill, 2007). Once λ is found, the probability mass function for the Poisson lognormal (hereafter lognormal) is derived using:

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

where p_{LN} is the lognormal probability.

METE — The SAD prediction from the maximum entropy theory of ecology (METE) (Harte 2011) is based on two empirical inputs: species richness (*S*) and total abundance (*N*) of individuals (or sequence reads) in a sample. Four constraints are produced from

these empirical inputs and an inferred rate of community-level metabolism (E): the average number of individuals per species (N/S), the average per species metabolic flux ($\varepsilon = E/S$), and the constraints that no species has more than N individuals or a greater total metabolic rate than E. The energetic constraint ε is integrated out, leaving the predicted SAD independent of ε, meaning that METE predicts only one form of the SAD for a given combination of N and S. The predicted SAD is based on a joint conditional probability distribution that describes the distribution of individuals (n) over species and of metabolism (ε) over individuals within a species (Harte *et al.,* 2008; Harte, 2011). Entropy of the SAD is then maximized according to the method of Lagrange multipliers. The SAD is then derived by integrating out energy and dropping terms that are vanishingly small. METE predicts the shape of the SAD by calculating the probability that the abundance of a species is n given S and N:

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

where β is defined by the equation

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

This approach to predicting the MaxEnt form of the SAD yields the log-series distribution of Fisher *et al.* (1943). The log-series is one of the two most successful SAD models, the other being the lognormal.

Broken-stick — The simultaneous Broken-stick (hereafter Broken-stick) model of MacArthur (1960) predicts the SAD as the simultaneous breaking of a stick of length N at S -1 randomly chosen points. The lengths of segments represent the predicted abundance

of species. The Broken-stick predicts one of the most even forms of the SAD. Like the prediction of METE, the Broken-stick has a general statistical equivalent, i.e., an exponential distribution that, for discrete cases, is the geometric distribution (Cohen, 1968; Heip *et al.,* 1998):

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

Zipf distribution — The Zipf-distribution (Zipf, 1949) is based on a power-law for frequencies of ranked data and is characterized by one free parameter (α), where the frequency of the k^{th} rank abundance is inversely proportional to k, i.e., $p(k) \approx k^{\alpha}$, with α often ranging between -1 and -2 (Gans, 2005; Newman, 2006). In contrast to the Broken-stick, the Zipf distribution predicts one of the most uneven forms of the SAD and can be shown to predict both more singletons than METE as well as greater dominance (i.e., the abundance of the most abundant species). The Zipf distribution predicts the frequency of elements of rank k out of N elements with parameter α as:

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

Testing SAD predictions

Our SAD predictions are based on the rank-abundance form of the SAD, i.e., a vector of species abundances ranked from most to least abundant (Fig. 1). The predictions of each model (i.e., METE, lognormal, Broken-stick, Zipf) yields the same value of S (i.e., number of species) that is given in the empirical input. This means that the observed and predicted SADs can be directly compared (rank-for-rank) using regression analyses to reveal the percent variation explained by each model.

We generated the predicted forms of the SAD using the code of White *et al.* (2012) (https://github.com/weecology/white-etal-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 1,000 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 (i.e. 239 SADs from MG-RAST). We then calculated the modified coefficient of determination (r^2_m) around the 1:1 line (as per White *et al.,* 2012; Locey and White, 2013, Xiao *et al.,* 2015) with the following equation.

$$r_m^2 = 1 - \frac{\sum(log_{10}(obs_i) - log_{10}(pred_i))^2}{\sum(log_{10}(obs_i) - \overline{log_{10}(obs_i)})^2}$$

Negative values are possible because the relationship is not a fitted one, i.e., estimating variation around a line with a constrained slope of 1.0 and a constrained intercept of zero (White *et al.,* 2012; Locey and White, 2013; Xiao *et al.,* 2015).

Diversity-abundance scaling relationships

Scaling relationships take the form of $y = x^z$ and reveal how one variable changes in a linear and proportional way across orders of magnitude in another variable. The primary feature of scaling relationships is the scaling exponent z, which becomes the slope of a linear relationship when axes are arithmetically scaled, i.e., $\log(y) = z\log(x)$ is equivalent to $y = x^z$. Scaling relationships are mathematically simplistic and are among the most powerful statistical relationships in ecology.

Recently, aspects of taxonomic diversity have been shown to scale with N (Locey and Lennon, 2016). These aspects include richness (i.e., the number of OTUs; S), dominance (i.e., the abundance of the most abundant OTU; N_{max}), evenness (i.e., similarity in abundance among OTUs captured by the variance of the SAD), and rarity (i.e., concentration of abundance among low abundant taxa captured by the skewness of the SAD). Except for the scaling of S with N, these relationships suggested universal scaling behavior for microorganisms and macroscopic plants and animals. Additionally, the scaling of N_{max} with N spans an unprecedented 30 orders of magnitude. However, the mechanisms that explain these scaling relationships have yet been reported and it is unknown whether any single biodiversity theory can explain, and hence unify, them. We used the values of N_{max}, evenness, and rarity derived from the predicted SADs of each model. We could not assess the ability of the SAD models to predict richness, as all the models rely on S as an empirical input, returning an SAD with the same number of species as the empirical SAD. We then examined these values 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, i.e., $\log(y) = z\log(x)$ is equivalent to $y = x^z$, where z is the slope and scaling exponent. These scaling exponents were compared to the exponents reported in Locey and Lennon (2016). 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 using the following equation:

$$\%\, Difference = \frac{|E_1 - E_2|}{\frac{1}{2}(E_1 + E_2)} * 100$$

Where E_1 and E_2 represent, respectively, quantities produced by an SAD model and predicted by the empirical scaling relationship. We then calculated the percent error using the following equation:

$$\%\ Error = |\frac{E_1 - E_2}{E_2}| * 100$$

Influence of total abundance on model performance

The form of the SAD is mathematically constrained by the number of individuals sampled (i.e. *N*) (Locey and White, 2013). For example, as *N* increases, rarity will tend to increase while species evenness will tend to decrease (Locey and White 2013, Locey and Lennon, 2016). Because models of the SAD can predict characteristically different forms, the success of a model may depend on the scale of *N*. In this way, models that predict relatively even SADs, e.g., Broken-stick, should increasingly fail with larger *N*. In contrast, the lognormal should perform better with larger *N* because it arises as a consequence of the law of large numbers and the central limit theorem (May, 1975; Putnam, 1993) However, to the best of our knowledge, the performance of SAD models across scales of *N* has rarely if ever been examined. We used ordinary least-squares regression to assess the relationship between the performance of each SAD model (measured by r^2_{m}) and *N*.

Computing code

We used open source computing code for obtaining the maximum-likelihood estimates for the Broken-stick, the lognormal, the prediction of METE (i.e. the log-series distribution), and the Zipf distribution (github.com/weecology/macroecotools, github.com/weecology/METE). This is the same code used in studies that showed

support for METE among communities of macroscopic plants and animals (White *et al.,* 2012; Baldridge *et al*., 2015; Xiao *et al*., 2015). If microbial SADs do not meaningfully differ from the SADs of these other taxa, then METE will perform better than the lognormal, the Zipf, and the Broken-stick. All analyses can be reproduced or modified for further exploration by using code, data, and following directions provided here: https://github.com/LennonLab/MicrobialBiodiversityTheory.

RESULTS

The sampling form of the lognormal, i.e., the Poisson lognormal, explained nearly 97% of variation in abundance among microbial taxa, compared to 91% for the Zipf distribution and 60% for the log-series predicted by the maximum entropy theory of ecology (METE) (Fig 2, Table 1). The overall performance of the Broken-stick was too poor to be evaluated (r^2_m = -0.60). Though close in predictive power, the results of the lognormal and the Zipf distribution differed in that the Zipf often greatly over-predicted the abundance of the most abundant taxa (N_{max}). In some cases, the predicted N_{max} of the Zipf distribution was greater than observed total abundance (N). In contrast, the lognormal was not biased towards over or under predicting both dominant and rare taxa. Neither the percent cutoff for sequence similarity used to cluster 16S rRNA reads into OTUs, nor the inclusion or exclusion of singleton OTUs had a substantial effect on the explanatory power of the various models (Fig S1, S2; Table S1, S2).

The lognormal best reproduced empirical diversity-abundance scaling relationships (Locey and Lennon, 2016) (Table 2). In each case, the lognormal closely approximated the values of the exponents for how rarity, evenness, and dominance have been found to scale with N. While the Zipf rivaled the lognormal in predicting microbial

SADs, it was only able to more closely approximate the scaling of absolute dominance (N_{max}) with N. The lognormal and Zipf were also the only models to explain the variation in dominance among sites (Fig 3). Neither the percent difference in the N_{max} relationship of the log-series predicted by METE nor the Broken-stick model came close to reproducing those scaling relationships or explaining the variation in dominance. Finally, models that closely reproduced the empirical dominance scaling law (i.e., lognormal, Zipf) were also able to make reasonable predictions of N_{max} (Fig 4).

The total number of reads (N) influenced the success of SAD models, where increasing N led to decreasing performance of the Broken-stick and log-series, and increasing performance of the lognormal and Zipf (Fig 3). In contrast, the performance of the lognormal and the Zipf increased with N, with the relationship being strongest for the lognormal (Fig 3). These results reflect how N numerically constrains the form of the SAD and how the SAD is likely to assume a lognormal form as N increases, a consequence of the central limit theorem and law of large numbers, i.e., lognormal dynamics (Putnam, 1993).

DISCUSSION

Overview

In this study, we sought a unifying explanation for common patterns of microbial biodiversity. These include recently documented diversity-abundance scaling relationships (i.e. Locey and Lennon, 2016) and the highly uneven forms of abundance distributions that characterize the microbial "rare biosphere". We focused on the two most historically accurate models of community structure (i.e., the lognormal and the log-

series), both of which have a history of use in microbial ecology (Curtis *et al.,* 2002; Bohannan and Hughes, 2003; Schloss and Handelsman, 2006; Dumbrell *et al.,* 2010). We also included two other well-known models (i.e., Broken-stick and Zipf) that predict qualitatively disparate forms the species abundance distribution (SAD), but have only been tested on relatively small microbial datasets from few sites (e.g., Gans *et al.,* 2005; Dumbrell *et al.,* 2010). In contrast to recent overwhelming support among communities of macroscopic plants and animals for the log-series distribution predicted by the maximum entropy theory of ecology (METE) (White *et al.,* 2012; Baldridge *et al.,* 2015), the lognormal provided the most accurate predictions for nearly all patterns in this study. Likewise, results from the lognormal provide the first explanation for the diversity-abundance scaling relationships, suggesting that the lognormal is able to capture general ecological features of microbial communities (Locey and Lennon, 2016).

A unifying theory of microbial biodiversity

Altogether, our findings point to a unifying lognormal theory of microbial biodiversity. This conclusion is supported by the ability of the lognormal model to make local-to-global-scale predictions of microbial richness using total abundance (N) and the abundance of the most abundant taxon (N_{max}) (e.g., Curtis *et al.,* 2002; Locey and Lennon, 2016). These predictions of a lognormal theory of microbial biodiversity span many orders of magnitude in total abundance and are based on how microbial communities exemplify the dynamics that underpin the lognormal distribution. In community ecology, the classic explanation of the lognormal states that it arises from the multiplicative interactions of many random variables (May, 1975; Putnam, 1993). In

short, the synergistic outcome of stochastic events among large numbers of individuals results in a predictable distribution of abundance. As ecological dynamics are often stochastic and multiplicative (MacArthur, 1960; Sih *et al.,* 1998; Hubbell, 2001) and as microbial communities inherently represent large number ecological systems, a lognormal theory of microbial biodiversity is an appropriate and promising starting point for unifying patterns of microbial commonness and rarity across scales of abundance.

The success of the lognormal increased with greater *N* (Fig. 2). In contrast, the success of the log-series predicted by METE decreased with greater *N*. As a result, the most successful ecological theory of biodiversity for macroscopic plants and animals fails at relatively small magnitudes of microbial abundance. This finding is particularly important because a successful theory of biodiversity for microorganisms must have predictive power across communities and microbiomes of vastly different size. In fact, the scales of abundance across which a successful theory must hold would include the ~25 orders of magnitude in *N* that could not be accounted for in this study, but that characterize abundances of microbial life from gut microbiomes to all of Earth (Whitman *et al.,* 1998; Kallmeyer *et al.,* 2012). To our knowledge, models based on lognormal dynamics are the only biodiversity models that have shown this degree of promise (Curtis *et al.,* 2002; Locey and Lennon, 2016), though both the lognormal and the Zipf successfully reproduced the dominance-abundance scaling law.

In predicting the form of the abundance distribution, the success of the Zipf often rivaled that of the lognormal and, like the lognormal, the performance of the Zipf increased with *N*. These similarities have a sound and ecologically relevant explanation. Specifically, the type of power-law behavior that characterizes the Zipf may emerge from

the mixing of lognormal distributions (Allen *et al.,* 2001). That is, communities that are well explained by the Zipf distribution could actually be composed of smaller communities that are each characterized by a lognormal distribution. This dynamic readily applies to microbial communities and microbiomes, where environmental samples lump together taxa that may not interact. Consequently, a theory of microbial biodiversity based on lognormal dynamics allows for the emergence of power-law behavior such as diversity-abundance scaling laws and the success of the Zipf distribution.

Lognormal distributions emerge as a central limiting pattern that results from the multiplicative interactions of many random variables and the law of large numbers (MacArthur, 1960; Putnam, 1993). In community ecology, the lognormal has been envisioned to characterize large or heterogeneous communities of species that respond to many complex variables and independent processes. However, a lognormal theory of microbial biodiversity invokes an even more ecologically meaningful interpretation, i.e., one that captures the macroecology of microorganisms. In a microbiome of immense abundance, stochastic dynamics within and among taxa that partition many different resources is a prime example of a large-number ecological system operating under lognormal dynamics (Putnam, 1993). The natural complexity of microbiomes, the capacity for microorganisms to grow on a large numbers of resources, their high abundances and diversity, and the contribution of stochasticity (e.g., Epstein, 2009; Stegen *et al.,* 2012) further emphasizes that lognormal dynamics underpin the assembly and structure of microbial communities and microbiomes.

Acknowledgements

We thank Jack Gilbert and Sean 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. Finally, we would like to thank The Human Microbiome Project Consortium for providing their data on the National Institutes of Health's publically accessible Data Analysis and Coordination Center server. This work was supported by a National Science Foundation Dimensions of Biodiversity Grant (#1442246 to JTL and KJL) and the U.S. Army Research Office (W911NF-14-1-0411 to JTL).

References

Allen AP, Li B-L, Charnov EL. (2001). Population fluctuations, power laws and mixtures of lognormal distributions. *Ecol Lett* **4**:1–3.

Amend AS, Seifert KA, Samson R, Bruns, TD. (2010). Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics. *Proc Natl Acad Sci USA* **107**:13748–13753.

Baldridge E, Xiao X, White EP. (2015). An extensive comparison of species-abundance distribution models. bioRxiv. doi: http://dx.doi.org/10.1101/024802

Brown JH, Mehlman DW, Stevens GC. (1995). Spatial variation in abundance. *Ecology* **76**: 1371–1382.

Chu H, Fierer N, Lauber CL, Caporaso JG, Knight R, Grogan P. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ Microbiol* **12**: 2998–3006.

Cohen JE. (1968). Alternate derivations of a species-abundance relation. *Amer Nat* **102**: 165–172.

Curtis TP, Sloan WT, Scannell JW. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99**: 10494–104999.

Dunbar J, Barns SM, Ticknor LO, Kuske CR. (2002). Empirical and theoretical Bacterial diversity in four Arizona soils. *Appl Environ Microbiol* **68**: 3035–3045.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH. (2010). Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME J* **4**: 337–345.

Epstein S. (2009). Microbial awakenings. *Nature* **457**: 1083.

Fierer N, Lauber CL, Ramirez KS, Zaneveld J, Bradford MA, Knight R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J* **6**: 1007–1017.

Fisher RA, Corbet AS, Williams CB. 1943. 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.

Flores GE, Campbell JH, Kirshtein JD, Meneghin J, Podar M, Steinberg JI, *et al.,* (2011). Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environ Microbiol* **13**: 2158–2171.

Gans J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**: 1387–1390.

Gilbert JA, Jansson JK, Knight R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biol* **12**: 69.

Green J, Bohannan BJM. (2006). Spatial scaling of microbial biodiversity. *Trends Ecol Evolut* **21**: 501–507.

Harte AJ, Zillio T, Conlisk E, Smith AB, Harte J. (2008). Maximum entropy and the state-variable approach to macroecology. *Ecology* **89**: 2700–2711.

Harte J. (2011). Maximum entropy and ecology. Oxford University Press: New York.

Heip CHR, Herman PMJ, Soetaert K. (1998). Indices of diversity and evenness. *Oceanis* **24**: 61–87.

Horner-Devine MC, Lage M, Hughes JB, Bohannan BJM. (2004). A taxa-area relationship for bacteria. *Nature* **432**: 750–753.

Hubbell SP. (2001). The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press: Princeton.

Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci USA* **109**: 16213–16216.

Locey KJ, Lennon JT. (2016). Scaling laws predict global microbial diversity. *Proc Natl Acad Sci USA* **113**: 5970–5975.

Locey KJ, White EP. (2013). How species richness and total abundance constrain the distribution of abundance. *Ecol Lett* **16**: 1177–1185.

Lynch MDJ, Neufeld JD. (2015). Ecology and exploration of the rare biosphere. *Nature Rev. Microbiol.* **13**: 217–229.

MacArthur R. (1960). On the relative abundance of species. *Amer Nat* **94**: 25–36.

Magurran AE, McGill BJ. (2011). Biological diversity frontiers in measurement and assessment. Oxford University Press: New York.

May RM. (1975). Patterns of species abundance and diversity. In: Cody ML, Diamond JM (ed). Ecology and evolution of communities. Harvard University Press: Cambridge, pp 81-120.

McGill BJ. (2010). Towards a unification of unified theories of biodiversity. *Ecol Lett* **13**: 627–642.

McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, *et al.,* (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett* **10**: 995–1015.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, *et al.,* (2008). The metagenomics RAST server—a public resource for the automatic phylo- genetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Newman MEJ. (2005). Power laws, Pareto distributions and Zipf's law. *Contemp Phys* **46**: 323-351.

Pedrós-Alió C, Manrubia S. The vast unknown microbial biosphere. *Proc Natl Acad Sci USA* 2016; e-pub ahead of print 3 June 2016, doi:10.1073/pnas.1606105113

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

Reid A, Buckley M. (2011) The Rare Biosphere: A report from the American Academy of Microbiology. Washington, DC: American Academy of Microbiology

Schloss PD, Handelsman J. (2004). Status of the microbial census. *Microbiol Mol Biol R* **68**: 686–691.

Sih A, Englund G, Wooster D. (1998). Emergent impacts of multiple predators on prey. *Trends Ecol Evolut* **13**: 350–355.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, *et al.,* (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc Natl Acad Sci USA* **103**: 12115–12120.

Stegen JC, Lin X, Konopka AE, Fredrickson JK. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J* **6**: 1653–64.

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. (2007). The human microbiome project. *Nature* **449**: 804–810.

Wang J, Shen J, Wu Y, Tu C, Soininen J, Stegen JC, *et al.,* (2013). Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J* **7**: 1310–1321.

White EP, Thibault KM, Xiao X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology* **93**: 1772–1778.

Whitman WB, Coleman DC, Wiebe WJ. (1998). Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA* **95**: 6578–6583

Woodcock S, Van Der Gast CJ, Bell T, Lunn M, Curtis TP, Head IM, Sloan WT. (2007). Neutral assembly of bacterial communities. *FEMS Microbiol Ecol* **62**: 171–180.

Xiao X, McGlinn DJ, White EP. (2015). A strong test of the Maximum Entropy Theory of Ecology. *Amer Nat* **185**: 70–80.

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, *et al.,* (2012). Human gut microbiome viewed across age and geography. *Nature* **486**: 222–7.

Zinger L, Boetius A, Ramette A. (2014). Bacterial taxa-area and distance-decay relationships in marine environments. *Mol Ecol* **23**: 954–64.

Zipf GK. (1949). Human behavior and the principle of least effort. Addison-Wesley: Cambridge.

Table 1. Comparison of the performance of species abundance distribution (SAD) models for microbial datasets. The mean site-specific *r*-square (r^2_{m}) and standard error ($\sigma_{r^2_m}$) for each model from 1,000 bootstrapped samples of 200 SADs: Broken-stick, the log-series predicted by the Maximum Entropy Theory of Ecology (METE), the lognormal, and the Zipf power law distribution. The lognormal and the Zipf provide the best predictions for how abundance varies among taxa, and are also characterized by lower standard errors than the Broken-stick and the log-series.

Model	$\overline{r^2_m}$	$\sigma_{\bar{r^2_m}}$
Lognormal	0.97	0.0092
Zipf	0.91	0.0085
Log-series	0.60	0.0023
Broken-stick	−0.60	0.0024

Table 2. A comparison of how closely each model reproduces the scaling exponents from abundance-diversity scaling relationships in Locey and Lennon (2016). These scaling relationships pertain to absolute dominance (N_{max}), Simpson's metric of species evenness, and skewness of the SAD. The percent difference and percent error 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 Locey and Lennon (2016), i.e., where the mean for N_{max} was 1.0, the mean for evenness was -0.31, and the mean for skewness was 0.13. In general, the lognormal comes closest to simultaneously reproducing the empirical scaling exponents. The p-values were < 0.0001 for all scaling exponents.

Model	Diversity metric	Slope	% Difference
Lognormal	N_{max}	1.0	1.4
	Evenness	−0.48	42.0
	Skewness	0.10	23.0
Zipf	N_{max}	1.0	0.42
	Evenness	−0.53	53.0
	Skewness	0.086	41.0
Log-series	N_{max}	0.85	15.0
	Evenness	−0.16	66.0
	Skewness	0.049	91.0
Broken-stick	N_{max}	0.73	32.0
	Evenness	−0.022	170.0
	Skewness	0.014	16.0

Figure 1. Demonstrations of typical shapes of species abundance distributions (SAD) in rank-abundance form, as predicted by our SAD models. The grey line represents a single empirical SAD randomly chosen from our data (see methods). To provide an example of all the models used, each model was fit to this observed SAD as described in the Methods. The Broken-stick is well known to produce an overly even SAD, while the log-series is generally considered to be uneven enough to produce realistic SADs for plant and animal communities (White *et al.,* 2012). In contrast, the Zipf distribution is among the most uneven SAD models, often predicting more singletons than other models. Finally, the lognormal, based on the Poisson sampling of a lognormal distribution to capture actual sampling effects, tends to be less even than the canonical lognormal and more similar to the unevenness of the Zipf distribution.

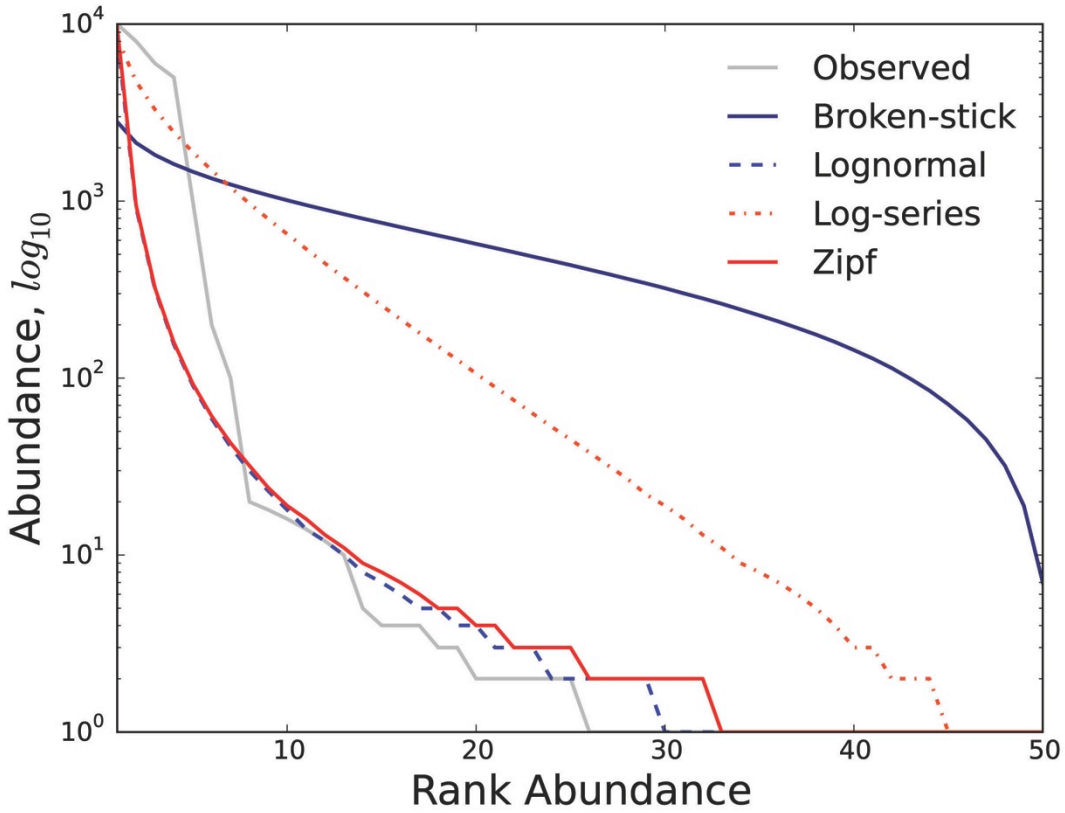

Figure 2. The relationship between the predicted abundance and the observed abundance of each rank for a single sample of each SAD model. The black diagonal line represents the 1:1 line. The box within each subplot is a histogram of the per-site modified r-squared (r^2_m) values from a range of zero 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^2_m value for 1,000 bootstrapped samples (see methods). Points are color-coded by the density of adjacent points. Hot colors (i.e., red) indicate a high density of adjacent points and cool colors (i.e., blue) indicate a low density of adjacent points. Each dot represents the observed abundance versus the predicted abundance for each species in the data.

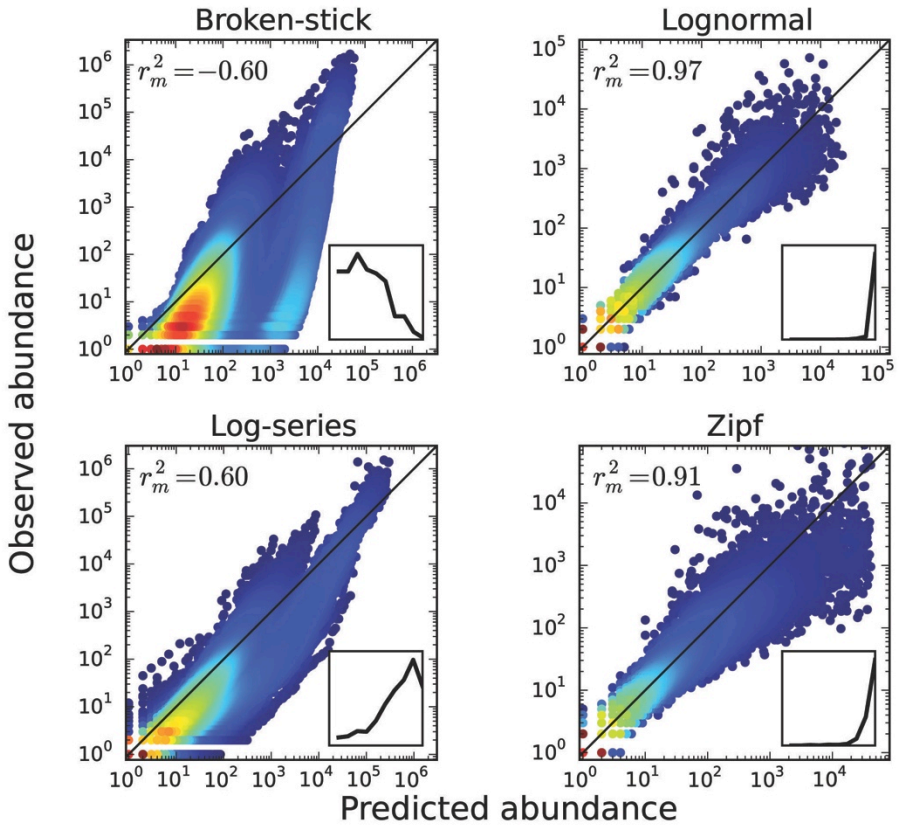

Figure 3. The relationship of model performance (via modified r^2_m) to the total number of 16S rRNA reads (N) for each SAD analyzed from a single sample. Here, r^2_m is the variation in the observed SAD that is explained by the predicted SAD (as in Fig 2). The performance of the Broken-stick model and the log-series distribution predicted by the maximum entropy theory of ecology (METE) decreases for greater N. In contrast, the lognormal and Zipf provide better explanations 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, i.e., the y-intercept is constrained to 0 and the slope is constrained to 1. Results from the simple linear regression can be found in Table S1.

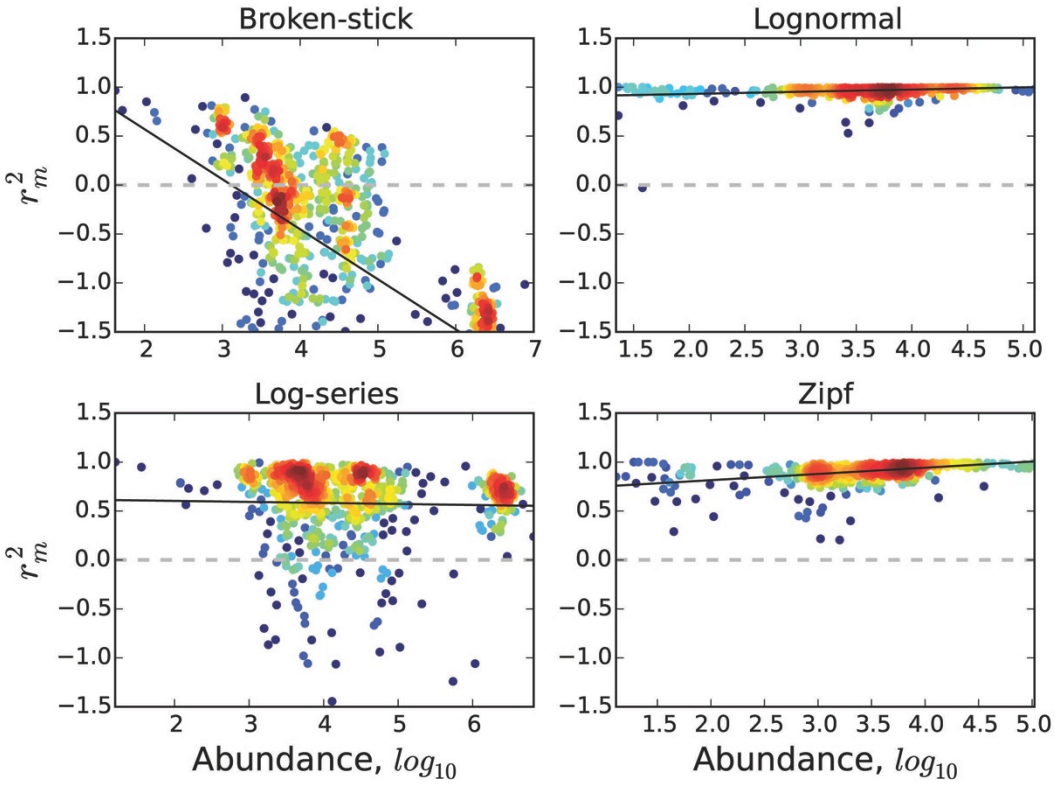

Figure 4. Predictions of absolute dominance (i.e., the abundance of the most abundant species, N_{max}) using the dominance scaling relationships of each model (Table 1) and the r^2_m of the relationship. Because of the negative r^2_m values for the Broken-stick and the log-series, only the lognormal and the Zipf are capable of providing meaningful predictions of N_{max}. This figure demonstrates the variability in N_{max} produced by models that closely approximate the empirical scaling law. Points are color-coded by the density of adjacent points. Hotter colors indicate a higher density of points.