A macroecological 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1450v4 | CC BY 4.0 Open Access | rec: 11 Nov 2016, publ:

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 abundance [1-4]. Over the past century, this endeavor has focused almost exclusively on macroscopic plants and animals (i.e., macroorganisms), giving little attention to the most abundant and taxonomically, functionally, and metabolically diverse organisms on Earth, i.e., microorganisms [1-4]. However, global-scale efforts to catalog microbial diversity across environmental, engineered, and host-related ecosystems has created an opportunity to understand biodiversity using a scale of data that far surpasses the largest macrobial datasets [5]. While commonness and rarity in microbial systems has 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 [6, 7]. While the rare biosphere has become a primary pattern of microbial ecology [6-8], it also reflects the universally uneven nature of one of ecology’s fundamental patterns, i.e., the species abundance distribution (SAD) [9]. The SAD is among the most intensively studied patterns of commonness and rarity, and is central to biodiversity theory and the study of patterns in abundance, distribution, and diversity across scales of space and time (i.e., macroecology) [9]. However, microbiologists have largely overlooked the connection of the SAD to theories of biodiversity and macroecology and the ability for some of those theories to predict other intensively studied patterns such as the species-area curve or distance-decay relationship [10].

Since the 1930’s, ecologists have developed more than 20 models that predict the SAD [3]. While some of these models are purely statistical and only predict the shape of the SAD (e.g., Gamma, Inverse Gamma), others encode the principles and mechanisms of competing theories [2-4, 9]. 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 [2]. 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 [11]. Examples of these “lognormal dynamics” are the multiplicative nature of growth and the stochastic nature of population dynamics. Another example is the stochastic nature of individual dispersal and the energetic costs that are multiplied across geographic distance. While most ecological processes likely have multiplicative interactions [11], many theories of biodiversity (e.g., neutral theory, stochastic geometry, stochastic resource limitation theory) include a stochastic component [2, 12-13]. Lognormal dynamics should become increasingly important for large communities, a result of the central limit theorem and law of large numbers [11]. 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 only rarely used in microbial studies [14-18].

Like the lognormal, the log-series has also been successful in predicting the SAD [19]. Though commonly used since the 1940's, the log-series is the form of the SAD that is predicted by one of the most recent, successful, and unified theories of biodiversity, i.e., the maximum entropy theory of ecology (METE) [4]. 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, i.e., the principle of maximum entropy [4, 20]. 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 [21, 22]. The success of METE has made the log-series the most highly supported model of the SAD [4]. But despite its success, METE has not been tested with microbial data and it is unknown whether METE can predict microbial SADs, a crucial requirement for a macroecological theory of biodiversity [23].

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 [5]. Such scaling laws are among the most powerful relationships in biology, revealing how one variable (e.g., *S*) changes in a proportional way across orders of magnitude in another variable (e.g., *N*). However, the mechanisms that give rise to these scaling 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 [5]. We used a compilation of 16S ribosomal RNA (rRNA) community-level surveys from over 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, i.e., the Zipf distribution, and one that predicts a highly even form, i.e., the Broken-stick. Because 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 are influenced by a primary constraint on the form of the SAD, i.e., sample abundance (*N*). 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 to 91% for the Zipf distribution and 64% for log-series predicted by METE (Fig. 2 and Table 1). The performance of the Simultaneous Broken-stick (hereafter referred to as the Broken-stick) was too poor to be evaluated. While 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 OTUs and percent cutoff in sequence similarity (Table S3 Fig. S3). In this way, the Zipf reasonably predicts the abundance of intermediately abundant taxa, but often fails for the most dominant and rare taxa [22, 24] (Tables S1 and S2). 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 (Figs. S1, S2; Tables S1, S2).

Predictive power across scales of sample abundance (*N*)

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 [21, 22], *N* for the vast majority of these samples was less than a few thousand organisms [21, 22]. In contrast, the log-series predicted by METE has yet to be tested using microbial data, i.e., 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 (i.e., 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. 3, Fig. S3). 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* [5,25]; 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 that were similar for molecular surveys of microorganisms and individual counts of macroorganisms [5]. These aspects of diversity include dominance (i.e., the abundance of the most abundant OTU; N_{max}), evenness (i.e., similarity in abundance among OTUs), and rarity (i.e., concentration of taxa at low abundances). We found that the lognormal best reproduced these diversity-abundance scaling relationships [5] (Table 2 and Fig. 4). 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 of N_{max} (Fig. 4). Additionally, neither the log-series predicted by METE nor the Broken-stick came close to reproducing the observed diversity-abundance scaling relationships (Fig. 4, 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 the maximum entropy theory of ecology (METE) was overwhelmingly supported [21, 22]. Such discrepancies in model performance suggest there are fundamental differences between macroorganisms and microorganisms that point to the importance of lognormal dynamics. Specifically, that multiplicative processes (e.g., growth) and stochastic outcomes (i.e., population fluctuations) produce a central limiting pattern within large and heterogeneous communities where species partition multiple resources [11]. Instead of identifying a particular process (e.g., dispersal limitation, resource competition), we propose that lognormal dynamics underpin the fundamental nature of microbial communities [11,12].

There are fundamental differences in how ecologists study communities of microscopic and macroscopic organisms. In our study, we accounted for some of the artifacts that could potentially contribute to the highly uneven microbial SADs. For example, we tested for the effects of percent similarity cutoffs that are used for defining an OTU, along with the influence of singletons and sample size. However, there are other caveats that deserve attention. First, ecologists sample microbial communities at spatial scales that greatly exceed the scales of their interactions [26]. As a result, samples of microbial communities probably lump together many ecologically distinct taxa that do not partition the same resources or occupy the same microhabitats. If microbial studies commonly lump together species that belong to different ecological communities, then this may in fact lead to the emergence of a power-law SAD (e.g., the Zipf) [27]. 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. While the connection between the lognormal and the Zipf needs further study, a macroecological theory of microbial biodiversity should allow for this dynamic.

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 [4]. In fact, METE appears capable of predicting both the lognormal and the Zipf [40]. This is because in using METE, one tries to infer the most likely form of an ecological pattern for a particular set of variables (e.g., N, S) and constraints (e.g., N/S). Consequently, the forms of ecological patterns predicted by METE could change depending on the constraints and state variables used [40]. For example, METE predicts that the SAD is a power law if it constrains the SAD to N/S while including a resource variable [40]. However, METE has not been as fully 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 (i.e., Poisson lognormal) used in our study. If so, and if it can reconcile why a log-series SAD works best for macrobes and a lognormal works best for microbes, then METE may indeed be a unified theory of biodiversity. Until then, microbial communities and microbiomes appear to be shaped by the multiplicative interactions of stochastic processes that, while 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 over 20,000 unique geographic sites. These data were compiled in a previous study [5] and include 14,962 sites from the Earth Microbiome Project (EMP) [28], 4,303 sites from the Data Analysis and Coordination

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1450v4 | CC BY 4.0 Open Access | rec: 11 Nov 2016, publ:

Center (DACC) for the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP) [29], 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 [30]. All sequence data were previously processed using established pipelines to remove low quality sequence reads and chimeras [28-30]. Additional information pertaining to the datasets can be found in the supplement and in previous studies [5].

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, i.e., the Simultaneous broken-stick [12] and the Zipf distribution [31, 32] can reasonably predict microbial SADs (Fig. 1). 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 operational taxonomic units (OTUs).

Lognormal — To avoid fractional abundances and to account for sampling error, we used a Poisson-based sampling model of the lognormal, i.e., the Poisson lognormal [33]. 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 λ (composed of the mean (μ) and standard deviation (σ)) of the Poisson lognormal is derived via numerical maximization of the likelihood surface [33]. 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 maximum entropy theory of ecology (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 maximum entropy (MaxEnt) [20]. Using METE, the shape of the SAD was predicted 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 the single fitted parameter β is defined by the equation

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

Where *N*/*S* is the average abundance of species. This approach to predicting the MaxEnt form of the SAD yields the log-series distribution [4, 19].

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 [12]. The Broken-stick has also has a purely statistical equivalent, i.e., the geometric distribution [34, 35]:

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

The Broken-stick has no free parameters and predicts only one form of the SAD for a given combination of *N* and *S*. Though rarely recognized, the geometric distribution is a maximum entropy solution when using *N* and *S* as “hard” constraints, i.e., the predicted SAD must have *S* species and a sum of *N* individuals.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1450v4 | CC BY 4.0 Open Access | rec: 11 Nov 2016, publ:

Zipf distribution — The Zipf (i.e., the discrete Pareto distribution) 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 frequency of ranked data and is characterized by one parameter (γ), where the frequency of the k^{th} rank is inversely proportional to k, i.e., $p(k) \approx k^{\gamma}$, with γ often ranging between -1 and -2 [31, 36-38] . The Zipf distribution predicts the frequency of elements of rank k out of N elements with parameter γ as:

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

We calculated the maximum likelihood estimate of γ 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, i.e., a vector of species abundances ranked from most to least abundant (Fig. 1). Because the predicted form of each model preserves S (i.e., number of species), we were able to directly compare (rank-for-rank) the observed and predicted SADs using regression 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 [21] (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 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 (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 previous tests of METE [21, 25, 39]) with the following equation.

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

It is possible to obtain negative r^2_{m} 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 [21, 25, 39]. In addition, 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 (Tables S5, Figures S5).

Diversity-abundance scaling relationships

To determine whether the SAD models tested here can explain previously reported diversity-abundance scaling relationships [5], we first calculated the values of N_{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. 5. 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, 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 reported exponents [5]. 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. 5. 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 sample sequence abundance (*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, repeated 100 times for each model. SADs were sampled at 50%, 25%, 12.5%, 6.25%, 3.125%, and 1.5625% percent of sample *N*. This subsampling analysis was computationally exhaustive and required SADs with *N* large enough to be halved 6 times and still large enough to be analyzed with SAD models. Likewise, we only used SADs for which predictions from each SAD model could be obtained at each scale of subsampled *N*. Altogether; we were able to use 10 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 (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 [22-24]. All analyses can be reproduced or modified for further exploration by using code, data, and following directions provided here: https://github.com/LennonLab/MicrobialBiodiversityTheory.

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. We also acknowledge the researchers who provided the open-source code for conducting some of our analyses. 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).

Author contributions

WRS and KJL conceived, designed, and performed the experiments, analyzed the data, and contributed materials/analysis tools. WRS, KJL, and JTL wrote the paper.

Competing financial interests

The authors declare no conflict of interest.

References

1. Brown, J.H., Mehlman, D.W. & Stevens, G.C. Spatial variation in abundance. *Ecology* **76**, 1371–1382 (1995).

2. Hubbell, S.P. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press: Princeton (2001).

3. McGill, B.J. Towards a unification of unified theories of biodiversity. *Ecol. Lett.* **13**, 627–642 (2010).

4. Harte, J. Maximum entropy and ecology: a theory of abundance, distribution, and energetics. (Oxford University Press, New York, USA, 2011).

5. 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. *Nature Rev. Microbiol.* **13**, 217–229 (2015).

8. Reid, A. & Buckley, M. The Rare Biosphere: A report from the American Academy of Microbiology. Washington, DC: 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, London, United Kingdom (1993).

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

13. Sih, A., Englund, G. & Wooster, D. Emergent impacts of multiple predators on prey. *Trends Ecol. Evolut.* **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. R.* **68**, 686–691 (2004).

18. Pedrós-Alió, C. & Manrubia, S. The vast unknown microbial biosphere. *Proc. Natl. Acad. Sci. USA*; e-pub ahead of print 3 June 2016, doi:10.1073/pnas.1606105113 (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 University Press: New York (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. Baldridge, E., Xiao, X. & White, E.P. An extensive comparison of species-abundance distribution models. bioRxiv. doi: http://dx.doi.org/10.1101/024802 (2015).

23. McGill, B. Strong and weak tests of macroecological theory. *Oikos* **102**, 679–685 (2003).

24. Ulrich, W., Ollik, M., & Ugland, 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. Gilbert, J.A., Jansson, J.K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).

29. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R. & Gordon, J.I. The human microbiome project. *Nature* **449**, 804–810 (2007).

30. Meyer, F. *et al.* The metagenomics RAST server—a public resource for the automatic phylo-genetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).

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

32. 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).

33. Magurran, A.E. & McGill, B.J. Biological diversity frontiers in measurement and assessment. Oxford University Press: New York (2011).

34. Cohen, J.E. Alternate derivations of a species-abundance relation. *Amer. Nat.* **102**, 165–172 (1968).

35. Heip, C.H.R., Herman, P.M.J. & Soetaert, K. Indices of diversity and evenness. *Oceanis* **24**, 61–87 (1998).

36. Zipf, G.K. Human behavior and the principle of least effort. Addison-Wesley, Cambridge, USA (1949).

37. Newman, M.E.J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323-351 (2005).

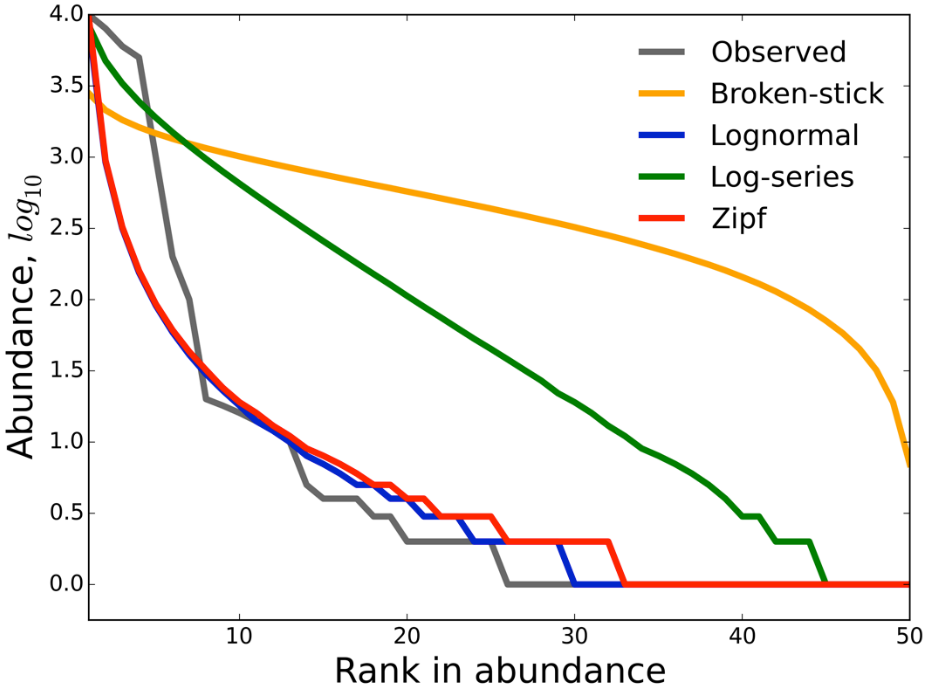
38. Newman, R.E.J. Power laws, Pareto distributions and Zipf's law. bioRxiv. doi: 10.1016/j.cities.2012.03.001 (2006).

39. Xiao, X., McGlinn, D.J. & White, E.P. A strong test of the Maximum Entropy Theory of Ecology. *Amer. Nat.* **185**, 70–80 (2015).

40. Harte, J., & Newman, E. Maximum information entropy: a foundation for ecological theory. *Trends Ecol. Evolut*. **29**, 384–389 (2014).

41. Ulrich, W., Ollik, M., & Ugland, K. I. A meta-analysis of species-abundance distributions. *Oikos* **119**, 1149–1155 (2010).

Figure 1. Forms of predicted species abundance distributions (SAD) in rank-abundance form, i.e., ordered from the most abundant species (N_{max}) to least the 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 Simultaneous 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 microbes [22]. 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 (i.e., 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 [41].



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1450v4 | CC BY 4.0 Open Access | rec: 11 Nov 2016, publ:

Figure 2. Relationships between predicted abundance and observed abundance for each SAD model. All species of all examined SADs are plotted, with hotter colors (e.g. red) reveal a greater density of species abundances. The black diagonal line is the 1:1 line, around which a perfect prediction would fall. The box within each subplot is a histogram of the per-SAD 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 10,000 bootstrapped samples (see methods). Each dot represents the observed abundance versus the predicted abundance for each species in the data.

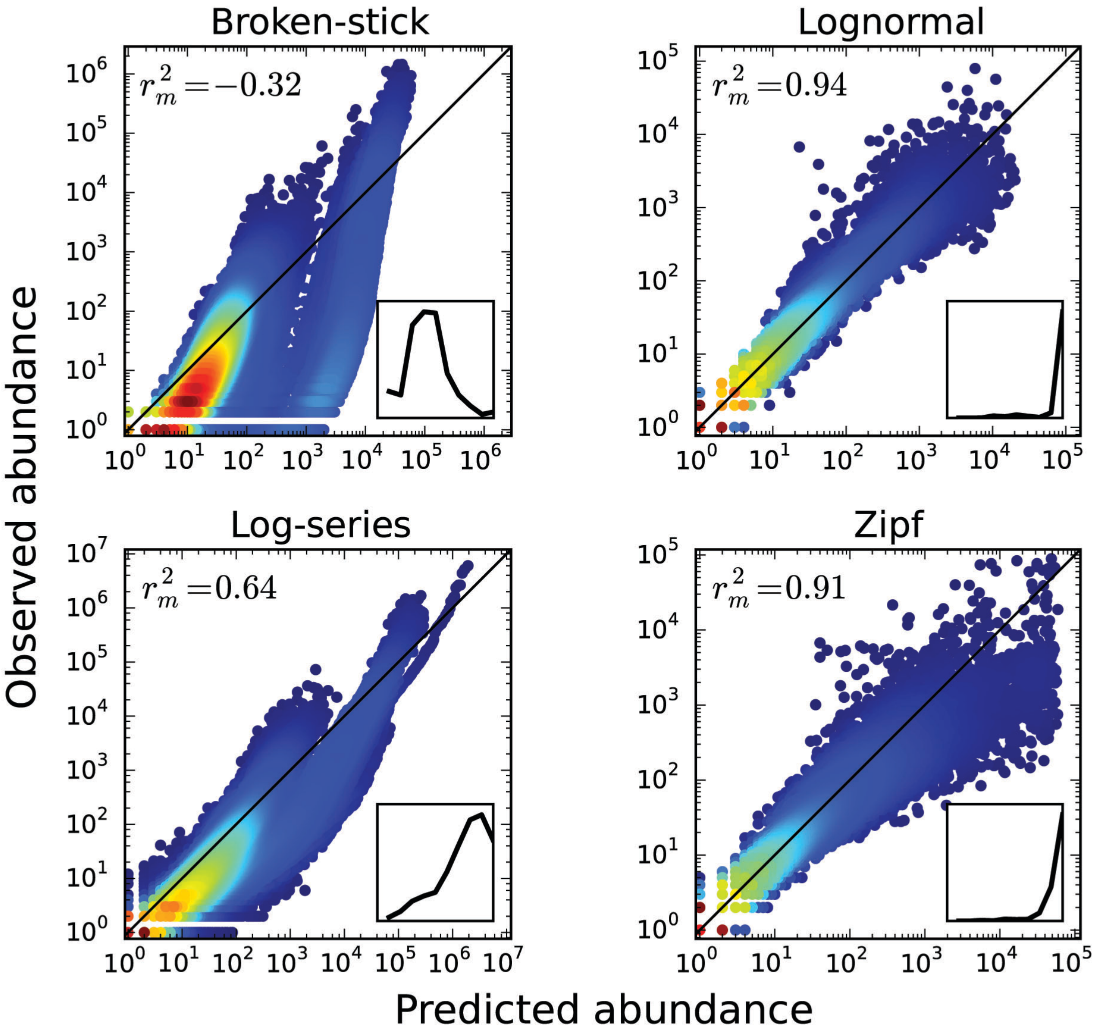


Figure 3. The relationship of model performance (via modified r^2_m) to the total number of 16S

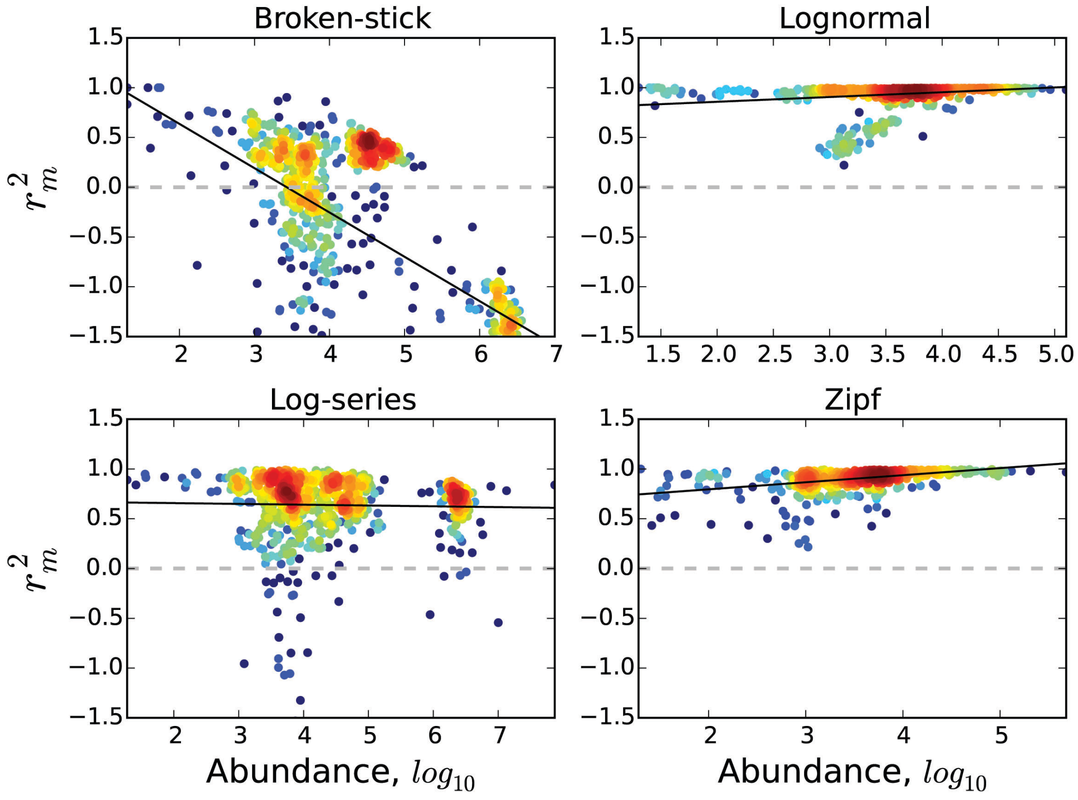
464 rRNA reads (N) for each SAD. The modified r-square value r^2_m is the variation in the observed SAD that is explained by the predicted SAD (as in Fig. 2). The performance of the Broken-stick

466 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 point, the lognormal

468 provides r^2_m values of 0.95 or greater across scales of N. The Zipf provide better explanations of microbial SADs with increasing N. The grey dashed horizontal line is placed where the r^2_{m}

470 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

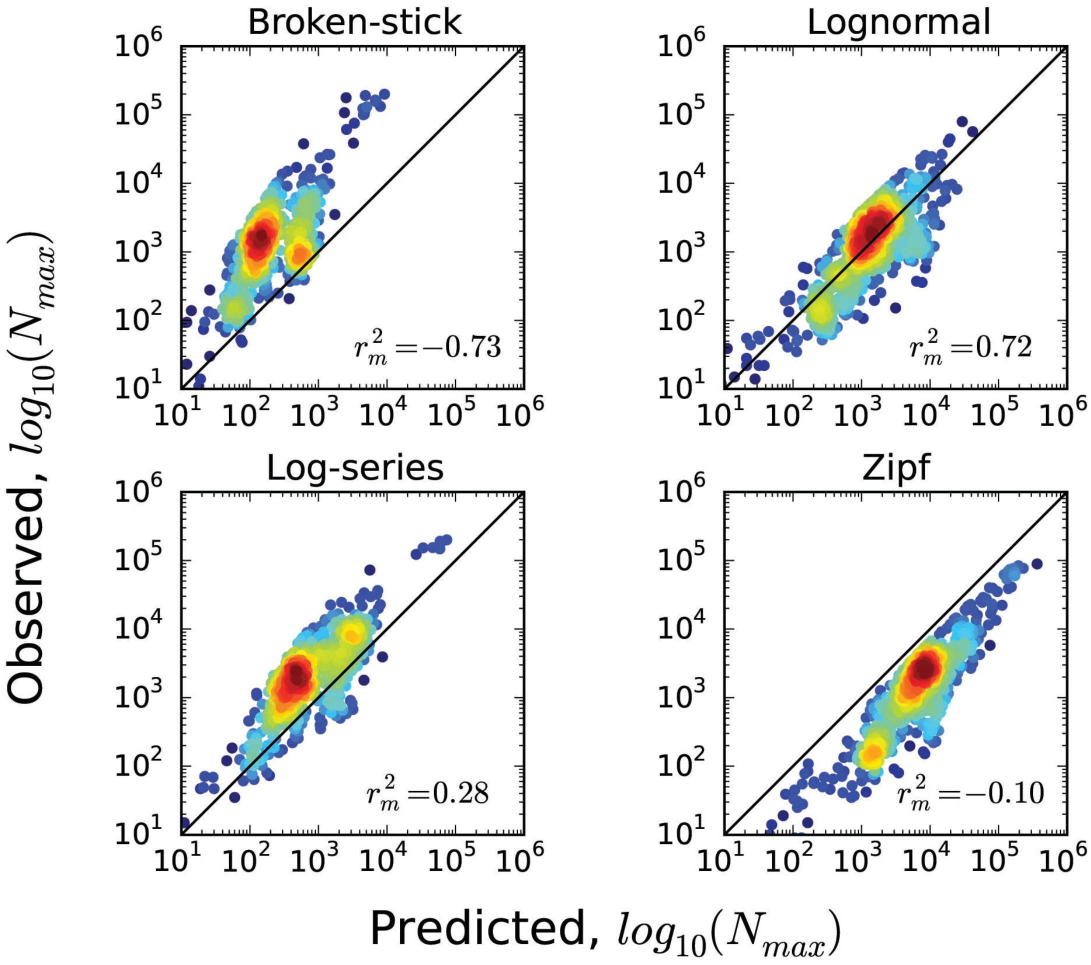
472 linear regression can be found in Table S1



474

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1450v4 | CC BY 4.0 Open Access | rec: 11 Nov 2016, publ:

Figure 4. Predictions of absolute dominance (i.e., greatest species abundance within an SAD,
476 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
478 lognormal and the Zipf are capable of providing meaningful predictions of N_{max}. This figure demonstrates the differences in N_{max} produced by models (i.e., lognormal and Zipf) that perform
480 well at predicting the SAD and closely approximate the dominance scaling exponent (Table 2). Hotter colors indicate a higher density of data points, i.e., results from SADs.



482

Table 1. Comparison of the performance of species abundance distribution (SAD) models for microbial datasets. The mean site-specific *r*-square ($r^2{}_{\mathrm{m}}$) and standard error ($\sigma_{r^2_m}$) for each model from 10,000 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. The lognormal and the Zipf 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	0.94	0.0044
Zipf	0.91	0.0031
Log-series	0.64	0.014
Broken-stick	−0.32	0.034

Table 2. In general, the lognormal comes closest to reproducing the scaling exponents of diversity-abundance scaling relationships [5]. 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 [5], i.e., where the mean for N_{max} was 1.0, the mean for evenness was -0.48, and the mean for skewness was 0.10. The p-values were < 0.0001 for all scaling exponents.

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