# Do modern theories of biodiversity fail to predict commonness and rarity among microbes?

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

Department of Biology, Indiana University, Bloomington, IN 47405 USA

[†]Authors contributed equally to the study

*Corresponding author, email: kjlocey@indiana.edu

## Abstract

Ecological theories of biodiversity seek to predict and unify patterns of commonness and rarity across taxa. The maximum entropy theory of ecology (METE) is among the most unifying theories of biodiversity, explaining >90% of variation in abundance among species of plant animal using the total number of individuals ($N_0$) and number of species as empirical inputs. However, METE has not been tested among the most abundant and diverse organisms on Earth, i.e., microorganisms. Using ~20,000 sites of microbial communities, we show that METE often explains <10% of variation in abundance and increasingly fails for larger $N_0$. In contrast, a more uneven distribution with a maximum entropy solution, the Zipf, often explains >90% of variation among microbes and performs better as $N_0$ increases. Our findings suggest that theories of biodiversity could produce accurate predictions across the tree of life and scales of abundance if they capture how disparities in abundance increase with $N_0$.



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

## Introduction

A primary goal of biodiversity theory is to predict patterns of biodiversity across evolutionarily distant taxa and scales of space, time, and abundance (Brown 1995, Hubbell 2001, McGill 2010, Harte 2011). Among the most universal of these patterns is the observation that few species in most ecological communities are highly abundant, while most are relatively rare, i.e., the canonical hollow-curve species abundance distribution (SAD) (McGill et al. 2007). The ubiquity of this pattern is a unifying assertion of biodiversity theory (McGill 2010) and explaining it has been a focus of community ecology, macroecology, and biogeography theory for decades (Whittaker 1972, Hubbell 2001, McGill et al. 2007). While SADs are often predicted as the result of resource partitioning, dispersal limitation, demographic stochasticity, competition and coexistence, the most successful models often have purely statistical explanations (e.g., Fisher et al. 1943, Preston 1948, Harte 2011).

One of the newest paradigms in biodiversity theory predicts patterns of commonness and rarity using the principle of maximum entropy (MaxEnt) from information theory (Pueyo et al. 2006, Harte et al. 2008, 2009). In short, the principle holds that the most likely form of a distribution is that having the most ways of occurring according to a set of state variables and any prior information that constrains the form of the distribution (Harte 2011). Recognizing that the form of the SAD is constrained by the state-variables of total abundance ($N_0$) and the number of species ($S_0$), MaxEnt models predict the most likely form of the SAD based on $N_0$ and $S_0$. However, because MaxEnt models often use additional constraints derived from $N_0$ and $S_0$ (e.g., average abundance, maximum abundance) and require assumptions such as whether species and individuals are distinguishable, different predictions are possible under different MaxEnt models (Haegeman and Etienne 2010).

Among the various MaxEnt frameworks, the maximum entropy theory of ecology (METE) of Harte (2011) has been the most successful in predicting the SAD and in coupling it to



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

other primary ecological patterns such as the species-area relationship, the distance-decay relationship, the spatial-abundance distribution, body-size distributions, among others (Harte 2011, Xiao et al. 2014). METE has been widely successful in predicting the SAD among communities of birds, mammals, trees, and invertebrates, often explaining >90% of observed variation in abundance among species (White et al. 2012, Baldridge et al. 2015). Despite its success in predicting SADs and other patterns of commonness and rarity, METE has not yet been tested among the most abundant and taxonomically, metabolically, and functionally diverse organisms on Earth, i.e., bacteria and archaea.

Within natural and host-associated ecosystems, most microbial taxa account for the minority of abundance. This seemingly universal pattern of microbial commonness and rarity is known as the microbial "rare biosphere" (Sogin et al. 2006, Reid and Buckley 2011). While the causes of the rare biosphere are typically studied with respect to the biology and ecology of microorganisms (Reid and Buckley 2011), the pattern reflects the universally uneven nature of SADs that characterize communities of macroscopic plants and animals. Yet it remains to be seen whether the theory that most often accurately predicts SADs among macrobes (i.e. METE) also succeeds in predicting SADs among microbes. If so, then patterns of commonness and rarity among microbes and macrobes may be unified by the assertions of METE and the principle of MaxEnt. However, the failure of METE to predict SADs among microbes would suggest a difference in patterns of commonness and rarity between microbes and macrobes that has yet been realized and accounted for in modern biodiversity theory.

Here, we test the ability of METE to predict microbial SADs using the largest compilation of microbial community yet assembled from publicly available sources. These data include 20,216 sites of bacterial and archaeal communities from the Earth Microbiome Project, the Human Microbiome Project, and datasets from Argonne National Laboratory's metagenomic server MG-RAST. For comparison to METE, we use the predictions of the Broken-stick model



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

(MacArthur 1960) and the Zipf distribution (Zipf 1949). The Broken-stick model is also a MaxEnt prediction (Haegeman and Etienne 2010) based on $N_0$ and $S_0$ but produces one of the most even forms of the SAD. In contrast, the Zipf distribution predicts a highly uneven SAD as the result of a power-law. While a previous study predating modern sampling methods and large molecular surveys has suggested that the Zipf provides a good characterization of microbial SADs, the authors were not able to test the prediction of METE and we cannot account for their method of fitting the Zipf and whether it conformed to best practices (see White et al. 2008). Because METE has been recently shown to out-perform both the relatively even Broken-stick and the relatively uneven Zipf at predicting SADs among mammals, trees, birds, and invertebrates (Harte 2011, White et al. 2012, Baldridge et al. 2015), we expect to see the same results if METE accurately characterizes microbial SADs.

## METHODS

### Data

We used bacterial and archaeal community sequence data from 20,456 sites. 14,962 of these sites were from the Earth Microbiome Project (EMP) (Gilbert et al., 2014) obtained on 22 August, 2014. Sample processing and sequencing of the V4 region of the 16S ribosomal RNA gene are standardized by the EMP and all are publicly available at www.microbio.me/emp. The EMP data consist of open and closed reference datasets, which are defined in the QIIME tutorial (http://qiime.org/tutorials/otu_picking.html) as follows. QIIME defines closed-reference as a classification scheme where any reads that do not hit a sequence in a reference collection are excluded from analysis. In contrast, open-reference refers to a scheme where reads that do not hit a reference collection are subsequently clustered de novo and represent unique but unclassified taxonomic units. Our main results are based on closed-reference data, due to the greater accuracy of the approach and because unclassified sequences were excluded from other microbial datasets



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

(below). However, we also examined the open-reference dataset, the results of which are consistent with our main findings (see Supplemental file).

We also used 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). These data consisted of samples taken from 15 to 18 locations (including the skin, gut, vagina, and oral cavity) on each of 300 healthy individuals. In each sample the V3-V5 region of the 16S rRNA gene was sequenced and analyzed using the mothur pipeline (Turnbaugh, et al., 2007). We excluded sites from pilot phases of the HMP as well as time-series data; see http://hmpdacc.org/micro_analysis/microbiome_analyses.php for details on HMP sequencing and sampling protocols. We also included 1,191 non-experimental sequencing projects consisting of processed 16S rRNA amplicon reads from the Argonne National Laboratory metagenomics server MG-RAST (Meyer, et al., 2008). Represented in this compilation were samples from arctic aquatic systems (CATLIN: 130 sites; MG-RAST id: mgp138), hydrothermal vents (HYDRO: 123 sites; MG-RAST id: mgp327) (Flores et al., 2011), freshwater lakes in China (187 sites; MG-RAST id: mgp2758) (Wang, et al., 2014), arctic soils (CHU: 44 sites; MG-RAST id: mgp69) (Chu et al., 2010), temperate soils (LAUB: 84 sites; MG-RAST id: mgp68) (Fierer et al., 2012), bovine fecal samples (BOVINE: 16 sites; MG-RAST id: mgp14132), human gut microbiome samples not part of the HMP project (529 sites; MG-RAST id: mgp401) (Yatsunenko, et al., 2012), and freshwater, marine, and intertidal river sediments (34 sites; MG-RAST id: mgp1829).

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 shown that the cutoff for percent sequence similarity in determining species-level units (95%, 97%, 99%) does not change the general shape of the SAD (Locey & While, 2013). However, how the percent cutoff



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

affects the fit of SAD models to empirical data is rarely tested (Woodcock et al., 2007; Dumbrell et al., 2010). The use of MG-RAST data allowed us to choose common parameter values for percent sequence similarity (i.e. the % for species-level) and taxa assignment including a maximum e-value (probability of observing an equal or better match in a database of a given size) of $10^{-5}$, a minimum alignment length of 50 base pairs, and minimum percent sequence similarities of 95, 97, and 99% to the closest reference sequence in MG-RAST's M5 rRNA database (Chu et al., 2010; Flores et al., 2011; Wang, et al., 2014; Fierer et al., 2012; Yatsunenko, et al., 2012). These latter analyses were conducted on MG-RAST datasets for which we obtained 95, 97, and 99% sequence similarity data: CHU, LAUB, HYDRO, CATLIN, BOVINE. All analyses can be reproduced or modified for further exploration by using code and data provided here: https://github.com/LennonLab/MicroMETE.

**METE**

The maximum entropy theory of ecology (METE) (Harte et al. 2008, 2009, Harte 2011) is based on two empirical inputs: species richness ($S_0$) and total abundance ($N_0$). These, along with an inferred rate of community-level metabolism ($E_0$), form the state variables of METE. Four constraints are produced from these state variables. These are the average number of individuals per species ($N_0/S_0$), the average per species metabolic flux ($\varepsilon = E_0/S_0$), and the constraints that no species has more than $N_0$ individuals or a greater total metabolic rate than $E_0$. The energetic constraint $\varepsilon$ is eventually integrated out, which leaves the predicted SAD independent of $\varepsilon$, meaning that METE predicts only a single form of the SAD for a given combination of $N_0$ and $S_0$.

The prediction of METE is based on a joint conditional probability distribution that describes the distribution of individuals ($n$) over species and of metabolism ($\varepsilon$) over individuals within a species (Harte et al. 2008, Harte 2011). Entropy of the distribution is then maximized



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

according to the method of Lagrange multipliers (Harte 2011). The SAD is then derived by integrating out energy and dropping terms that are vanishingly small. This approach, in fact, yields the log-series distributions of Fisher et al. (1943). As it happens, the log-series is among the oldest and most successful SAD models. In this case, METE predicts the shape of the SAD by calculating the probability that the abundance of a species is $n$ given $S_0$ and $N_0$:

$$\Phi\left(n \mid S_0, N_0\right) = \frac{1}{log(\beta^{-1})} \frac{e^{-\beta n}}{n}$$

where $\beta$ is defined by the equation

$$\frac{N_0}{S_0} = \frac{\sum_{n=1}^{N_0} e^{-\beta n}}{\sum_{n=1}^{N_0} e^{-\beta n}/n}$$

While METE uses $N_0$ as $S_0$ as state variables, neither are used as hard-constraints. That is, METE does not predict an SAD with $S_0$ species whose abundances are constrained to sum to $N_0$. This "soft" constrained nature (see Haegeman and Etienne 2010) is not exceptional, as other MaxEnt models produce similar or identical predictions (Pueyo et al. 2007, Dewar and Porté 2008).

**Broken-stick**

The simultaneous Broken-stick (SBS) model of MacArthur (1960) predicts the distribution of abundance as the simultaneous breaking of a stick of length $N_0$ at $S_0$ -1 randomly chosen points. The length of each segment represents the predicted abundance of each species. The SBS predicts one of the most even forms of the SAD, where the most dominant species are not abundant enough and where the rarest species are too abundant (McGill et al. 2007, Hubbell 2001). It also known that the form of the SBS is equivalent to the exponential distribution (Heip et al. 1998) which, for discrete cases, is the geometric distribution:

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



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

The geometric distribution is not to be confused with the geometric series of Motomura (1932); though it often is. Cohen (1968) shows that the geometric distribution is equivalent to discrete SBS, both of which, are known MaxEnt solutions when the predicted distribution is hard-constrained to have $N_0$ unlabeled individuals among $S_0$ labeled species (Harte et al. 2008, Haegeman and Etienne 2010).

**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 ($\alpha$), where the frequency of the $k^{th}$ rank abundance is inversely proportional to $k$, i.e., $p(k) \approx k^{\alpha}$, where $\alpha$ often ranges between -1 and -2 (Gans 2005, Newman 2006). In contrast to the simultaneous 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), as we show in this study. It is perhaps, interesting, that METE has been derived to not follow a power-law scaling behavior, which defines the Zipf distribution. It has also recently been shown that METE out-performs the Zipf distribution in predicting SADs for trees, mammals, birds, and invertebrates (Baldridge et al. 2015).

**Computing code**

We used open source computing code for obtaining the maximum-likelihood estimates for the geometric distribution, the prediction of METE (i.e. the log-series distribution), and the Zipf distribution (github.com/weecology/macroecotools, https://github.com/weecology/METE). This is the same code used in studies that showed support for METE or the general failure of the Zipf distribution among communities of macroscopic plants and animals (White et al. 2012,



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

Baldridge et al. 2015, Xiao 2015). If microbial SADs do not meaningfully differ from the SADs of these other taxa, then METE will perform better than both Zipf and the Broken-stick.

**Testing MaxEnt predictions**

Both METE (which predicts a log-series distribution) and the Broken-stick (i.e., the geometric distribution) can produce predictions for the rank-abundance form of the SAD. This form of the SAD is simply a vector of species abundances ranked from greatest to least. Both predictions yield the same value of *S* that is given as 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 (METE, Broken-stick).

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 (https://github.com/weecology/macroecotools), which contains functions for fitting maximum-likelihood forms of species abundance models. We calculated the modified coefficient of determination ($r^2_m$) around the 1-to-1 line (as per White et al. 2012, Locey and White 2013, Xiao et al. 2014) with the following equation.

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

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. 2014).

## RESULTS

SAD predictions from the maximum entropy theory of ecology (METE) generally explained 0-60% of variation in abundance among microbial species from microbiome projects,



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

i.e., EMP and HMP (Fig 1, Fig S1-S2, Table 1). This is a poor degree of explanatory power given that METE commonly explains 90-96% of variation among macroscopic plants and animals (White et al. 2012, Baldridge et al. 2015, Xiao et al. 2015). METE performed considerably better for MG-RAST datasets, often explaining 60-70% of variation, though the Zipf distribution consistently explained more (~87%) (Fig 1, Fig S1-S2, Table 1). Likewise, the Zipf distribution explained, on average, 85% of variation within the HMP data and 58% within the EMP open-reference dataset (where METE explained ~0.06%). The performance of the Broken-stick model was generally too poor to be interpreted, often resulting in negative values for the modified r-square, which again, are possible because the relationship is not fitted (White et al. 2012, Locey and White 2013).

The percent cutoff for sequence similarity used to cluster 16S rRNA reads into taxonomic units had no effect on the explanatory power of SAD models (Table 1). However, across datasets, the success of METE and the Broken-stick were influenced by $N_0$, where increasing $N_0$ led to decreasing performance of each model (Fig 2; Fig S3-S8). In contrast, the performance of the Zipf increased with $N_0$. We also found that the value of the Zipf exponent (for the rank distribution) was often close to -1.5 to -2, and that this result was also dependent on $N_0$, where increasing $N_0$ led to a value between -1.5 and 2 (Figs S9-S10).

## DISCUSSION

Within and among communities of macroscopic plants and animals, METE often explains 90 to 96% of observed variation in abundance among species. Here, we showed that while METE performs better than an alternative MaxEnt prediction (i.e., Broken-stick) it often fails to explain the majority of variation within and among communities of bacteria and archaea from a range of diverse natural and host-related ecosystems. These results are primarily due to the tendency of both the Broken-stick and METE (i.e., geometric distribution and log-series



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

distribution) to under-predict the dominance of the most abundant species and to over predict the abundance of the rarest. In effect, while it has been well-known that the Broken-stick predicts an overly even SAD, it appears that, for microbes, METE suffers from the same shortcoming. Ecologists familiar with research on the microbial "rare biosphere" may have anticipated this outcome, as SADs from samples of microbial communities and microbiomes appear to have exceptionally uneven forms (Reid and Buckley 2011).

In contrast, we found that the Zipf distribution generally out-performs METE in explanatory power and that the performance of the Zipf increased with greater $N_0$. However, it also appears that the Zipf distribution often over-predicts the abundance of the most abundant taxa. In a study predating ultra high throughput sequencing methods and large-scale microbiome surveys, the Zipf-distribution was shown to provide the best fit of any general model to microbial SADs of pristine and polluted soils, and typically had an exponent between -1.5 and -2 (Gans et al. 2005). This particular finding has received little attention, while the study itself was likely unable to use neither a MaxEnt form of the Zipf nor a maximum likelihood estimate which can be problematic (White et al. 2008, Baldridge et al. 2015). Online methods and supplementary files for that study appear to be inaccessible due to a failed link, so we cannot say how many communities the authors sampled or what methods they used for modeling. However, we found close agreement to this earlier study when using over 20,000 samples of microbial communities from a diverse array of natural and host-associated ecosystems.

METE's success is heavily influenced by one of its primary state variables ($N_0$). As a result, increasing $N_0$ causes METE as well as the Broken Stick to fail more severely. Importantly, these conditions also characterize numerical differences between microbial and macrobial SAD datasets. That is, $N_0$ for microbial datasets often represents tens of thousands to millions of processed rRNA reads. In contrast, $N_0$ for macrobial SADs typically ranges from a few hundred to a few thousand individual organisms. In short, METE might fail for microbes because it



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

simply fails with increasing $N_0$. The consequence of this finding is two-fold: First, METE either fails for microbes or when $N_0$ exceeds a few tens of thousands. Second, our findings suggest an increasing disparity in abundance for greater $N_0$ that one of the most accurate and unifying theories of biodiversity theory fails to track.

While it is surprising to see the Broken-stick and METE fail so greatly in predicting SADs among microorganisms, the failure of these models was not unforeseeable. It has been shown that as $N_0$ increases, the evenness of the SAD can be expected to decrease as a result of numerical constraints (Locey and White 2013). In the same way, as average abundance ($N_0/S_0$) increases, the evenness of the SAD can be expected to naturally decrease (Xiao et al. 2015). In both cases, constraints on the form of the SAD imposed by $N_0$ and $N_0/S_0$ lead to increasingly uneven SADs that outstrip the highly even form predicted by the Broken-stick (i.e. the geometric distribution) as well as the form predicted by METE (i.e. the log-series distribution). Still, it remains to be seen whether the inability of METE to predict microbial SADs is entirely driven by numerical constraints.

Our study suggests that highly uneven SADs are driven by mechanisms that lead to high $N_0$. However, uneven microbial SADs could also be driven by factors suggested to explain the microbial rare biosphere. For example, widespread dispersal and the ability of microbes to persist in suboptimal environments may allow many small populations of dormant or slow-growing organisms to have prolonged life spans that lead to accumulation of N (Reid and Buckley 2011; Lennon & Jones, 2011). Additionally, microorganisms may have unparalleled capacities to partition limited resources which, along with their microscopic size, may contribute to overall greater $N_0$. Consequently, the failure of the Broken-stick and METE may owe as much to the statistical influence of $N_0$ as to the ecological mechanisms that cause differences in abundances among specific species.



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

Our study suggests that ecology lacks a theory of biodiversity that captures the increasingly uneven nature of SADs with increasing $N_0$. Until now, ecology may have lacked an appropriate model to predict abundances when $N_0$ scales beyond a few tens of thousands, as is common in microbial community datasets. Yet, while the Zipf seems to perform better with increasing $N_0$ it is known to provide a relatively poor fit among communities of macroscopic organisms (Baldridge et al. 2015). Consequently, a greater synthesis is needed to establish a maximum entropy theory of ecology that works across scales and is not limited to predicting the log-series. Fortunately, it has been shown that the Zipf-distribution also has a MaxEnt solution (Baek et al. 2011, Visser 2013). If METE can be modified to predict an increasingly Zipf-like (i.e. power-law) SAD with increasing $N_0$, then perhaps the field of ecology will have arrived at a more unifying theory of biodiversity.

## Conclusion

The maximum entropy theory of ecology (METE) provides a first-principle framework for predicting biodiversity patterns based solely on small numbers of universal empirical inputs. Yet, it is clear from our study that METE will fail for communities of very large $N_0$, such as microbiomes where sampled $N_0$ is increasingly numbered in the millions. Consequently, while microbial SADs appear to be exceptional in their unevenness, we cannot conclude whether the cause is due to biological factors that drive rarity independent of their influence on $N_0$. It may be the biology which allows microbes to attain such high degrees of $N_0$, which then drives the SAD through statistical constrain-based mechanisms towards decreasing evenness.

## Acknowledgements



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

We thank Jack Gilbert and Sean Gibbons for providing EMP data and guidance on using it. This work was supported by a National Science Foundation Dimensions of Biodiversity Grant (#1442246) awarded to JTL and KJL.

## References

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

2. Brown, J. H. (1995). *Macroecology*. University of Chicago Press.

3. Chu, H. et al. (2010). Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ. Microbiol*. 12:2998–3006.

5. Dewar, R. C., Porté, A. (2008). Statistical mechanics unifies different ecological patterns. *Journal of Theoretical Biology*, 251:389–403.

6. Fierer, N. et al. (2012) Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *ISME J*. 6:1007–1017.

7. Flores, G. E. et al. (2011). Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environ. Microbiol*. 13:2158–2171.

8.

Gans, J. (2005). Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil. *Science*. 309:1387–1390.

9.

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC biology*. 12:69.

10.

Goris, J. et al. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Micr.* 57:81–91.

11.

Haegeman, B., Etienne, R. S., The, S., Naturalist, A., & April, N. (2010). Entropy Maximization and the Spatial Distribution of Species. *The American Naturalist*. 175:74–90.

12.

Harte, J., Zillio, T., Conlisk, E., Smith, A. B. (2008). Maximum Entropy and the State-Variable Approach to Macroecology. *Ecology*. 89:2700–2711.

13.

Harte, J. (2011). *Maximum Entropy and Ecology*. Oxford University Press. New York, New York, USA.

14.

Harte, J., Smith, A. B., Storch, D. (2009). Biodiversity scales from plots to biomes with a universal species-area curve. *Ecology Letters*. 12:789–797.

15.

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press. Princeton, New Jersey, USA.

16.



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

Lennon, J. T., & Jones, S. E. (2011). Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology*. 9:119–30.

17.

Locey, K. J., & White, E. P. (2013). How species richness and total abundance constrain the distribution of abundance. *Ecology Letters*. 16:1177–1185.

19.

McGill, B. J. (2010). Towards a unification of unified theories of biodiversity. *Ecology Letters*. 13:627–642.

20.

McGill, B. J., et al. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*. 10:995–1015.

21.

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

22.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*. 46:323-351.

23.

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

24.

Pueyo, S., He, F., & Zillio, T. (2007). The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecology Letters*. 10:1017–1028.

25.



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

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

394 27.

Sogin, M. et al. (2006). Microbial Diversity in the Deep Sea and the Underexplored "Rare

396 Biosphere." *Proc Natl Acad Sci USA*., 103(32), 12115–12120.

29.

398 Turnbaugh, P. J., et al. (2007). The human microbiome project. *Nature* 449:804–810.

31.

400 Visser, M. (2013). Zipf's law, power laws and maximum entropy. *New Journal of Physics*. 15.

32.

402 Wang, J., et al. (2014). Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J*. 7:1310-1321.

404 33.

White, E. P., Enquist, B. J., & Green, J. L. (2008). On estimating the exponent of power-law

406 frequency distributions. *Ecology*. 89, 905–912.

34.

408 White, E. P., Thibault, K. M., & Xiao, X. (2012). Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model Reports. *Ecology*,

410 93, 1772–1778.

35.

412 Whitman, W. B., Coleman, D. C., Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA*. 95:6578–6583.

414 37.

Woodcock, S., et al. (2007). Neutral assembly of bacterial communities. FEMS *Microbiol Ecol*.

416 62:171-80.

38.

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

39.

Xiao, X., Locey, K. J., & White, E. P. (2015). A Process-Independent Explanation for the General Form of Taylor's Law. *The American Naturalist*. 186:E51-E60.

40.

Yatsunenko, T. et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486:222-227.

41.

Zipf, G.K. (1949). Human Behavior and the Principle of Least Effort. Addison-Wesley. Cambridge, Massachusetts.



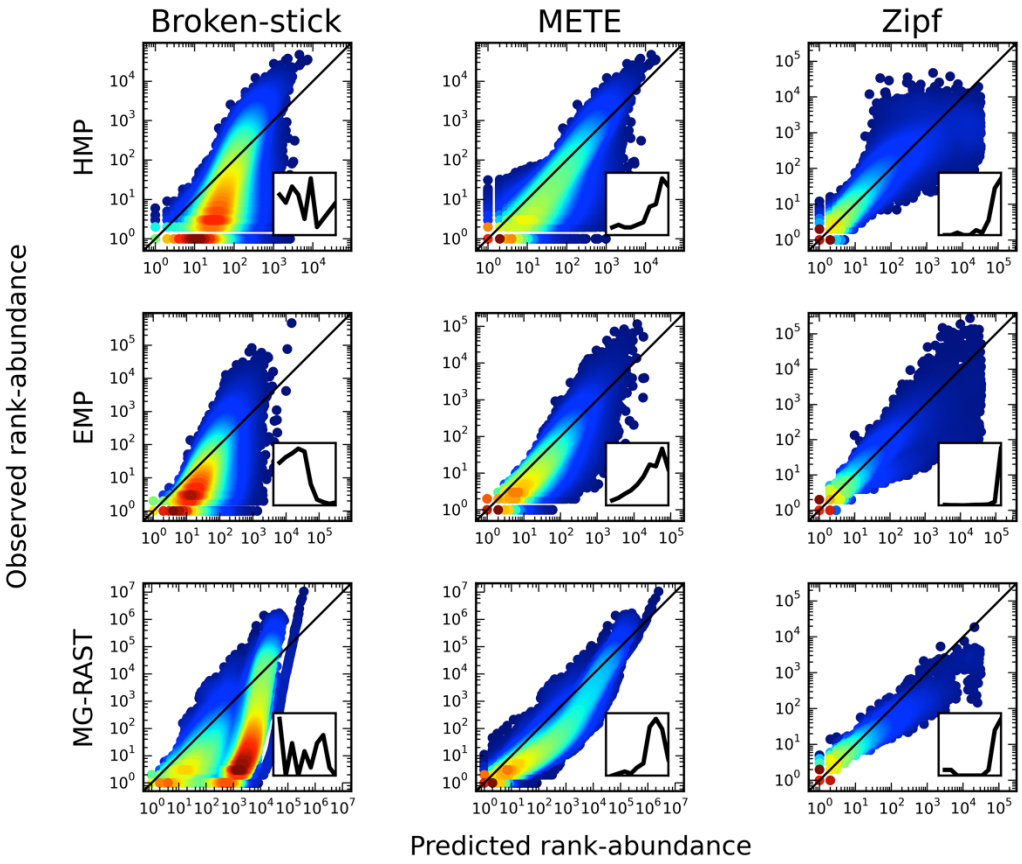
PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015

| Dataset | Model | $\overline{r^2_m}$ | $\sigma_{\overline{r^2_m}}$ | $\overline{N_0}$ | $\overline{S_0}$ |
|---|---|---|---|---|---|
| HMP | BS | −0.543 | 0.0170 | 5050 | 78 |
| | METE | 0.520 | 0.00846 | | |
| | Zipf | 0.854 | 0.0125 | | |
| EMP Closed | BS | −0.434 | 0.00851 | 44779 | 1189 |
| | METE | 0.562 | 0.00377 | | |
| | Zipf | 0.498 | 0.0250 | | |
| EMP Open | BS | −0.881 | 0.0101 | 88751 | 7247 |
| | METE | 0.0619 | 0.00526 | | |
| | Zipf | 0.577 | 0.0163 | | |
| MG − RAST | BS | −0.787 | 0.0294 | 1190013 | 366 |
| | METE | 0.693 | 0.00650 | | |
| | Zipf | 0.877 | 0.00842 | | |
| MG − RAST 95% | BS | 0.0493 | 0.0335 | 44346 | 306 |
| | METE | 0.785 | 0.00903 | | |
| | Zipf | 0.863 | 0.0102 | | |
| MG − RAST 97% | BS | 0.0397 | 0.0348 | 39788 | 270 |
| | METE | 0.779 | 0.00954 | | |
| | Zipf | 0.857 | 0.0122 | | |
| MG − RAST 99% | BS | 0.0214 | 0.0344 | 42674 | 299 |
| | METE | 0.778 | 0.00946 | | |
| | Zipf | 0.869 | 0.0116 | | |

Table 1. The mean and standard error of the modified r-square ($r^2_m$) for each dataset against either the Broken-stick (BS) or the Maximum Entropy Theory of Ecology (METE) and the mean per site total abundance ($\overline{N}$) and species abundance ($\overline{S}$) for each dataset. Note that percent sequence similarity for datasets obtained from MG-RAST did make a substantial difference in the performance of BS, METE, or the Zipf.



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015



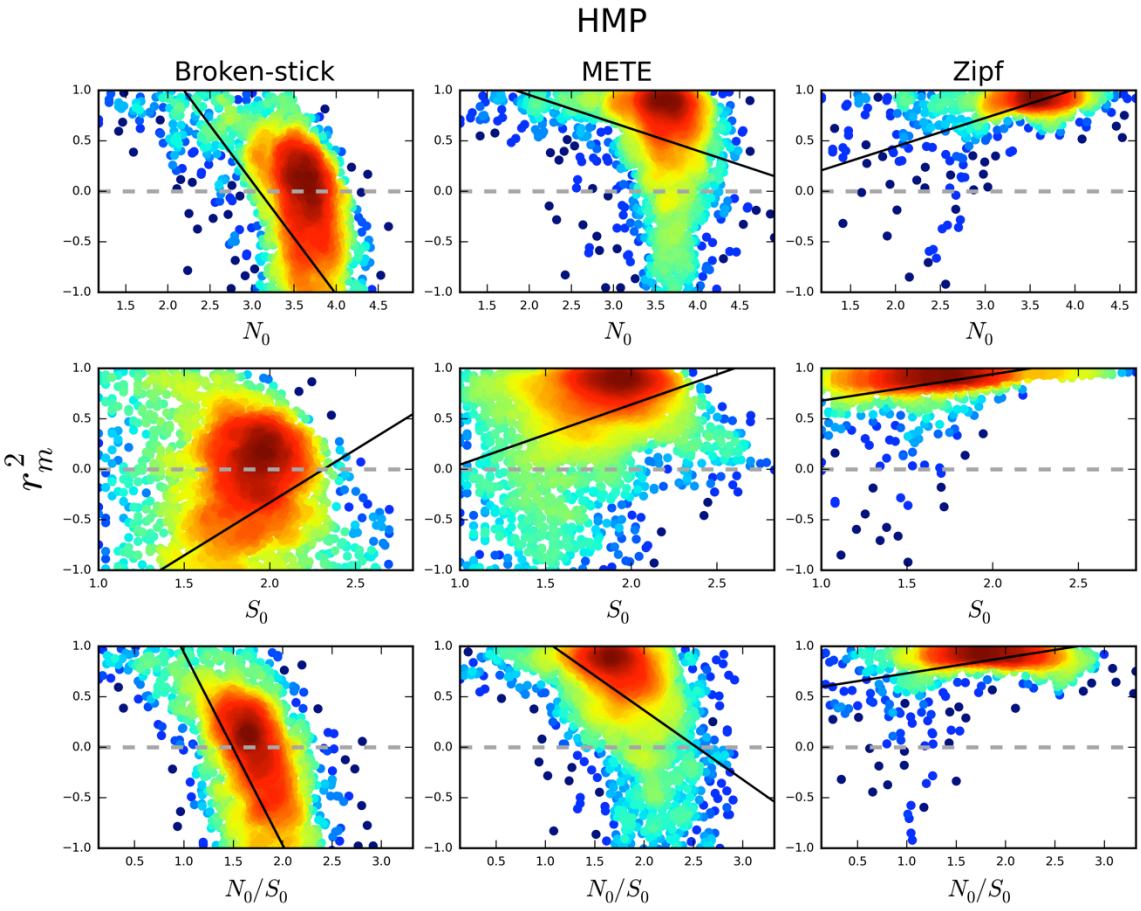

438

440 Figure 1: The relationship between the predicted rank-abundance and the observed rank-abundance across models and datasets. The Human Microbiome Project (HMP), the closed 442 reference Earth Microbiome Project (EMP) and environmental datasets obtained from the MG-RAST server clustered at 97% sequence similarity are arranged by row. SAD models are 444 arranged by column. The diagonal line represents the 1:1 line. The box within each subplot is a histogram of the modified r-squared ($r^2_m$) values from a range of zero to one.

446



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015




448

450 Figure 2: Ordinary least-squares regressions using either the total abundance ($N_0$), number of species ($S_0$), or the average species abundance ($N_0/S_0$) as the explanatory variable and the $r^2_m$ as 452 the response variable for HMP dataset. The black line is the slope of the relationship between the predictor and response variables. The grey dashed horizontal line is placed where the $r^2_m$ equals 454 zero as a point of reference. Similar trends were found for the EMP and MG-RAST datasets.

456



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1450v1 | CC-BY 4.0 Open Access | rec: 26 Oct 2015, publ: 26 Oct 2015