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

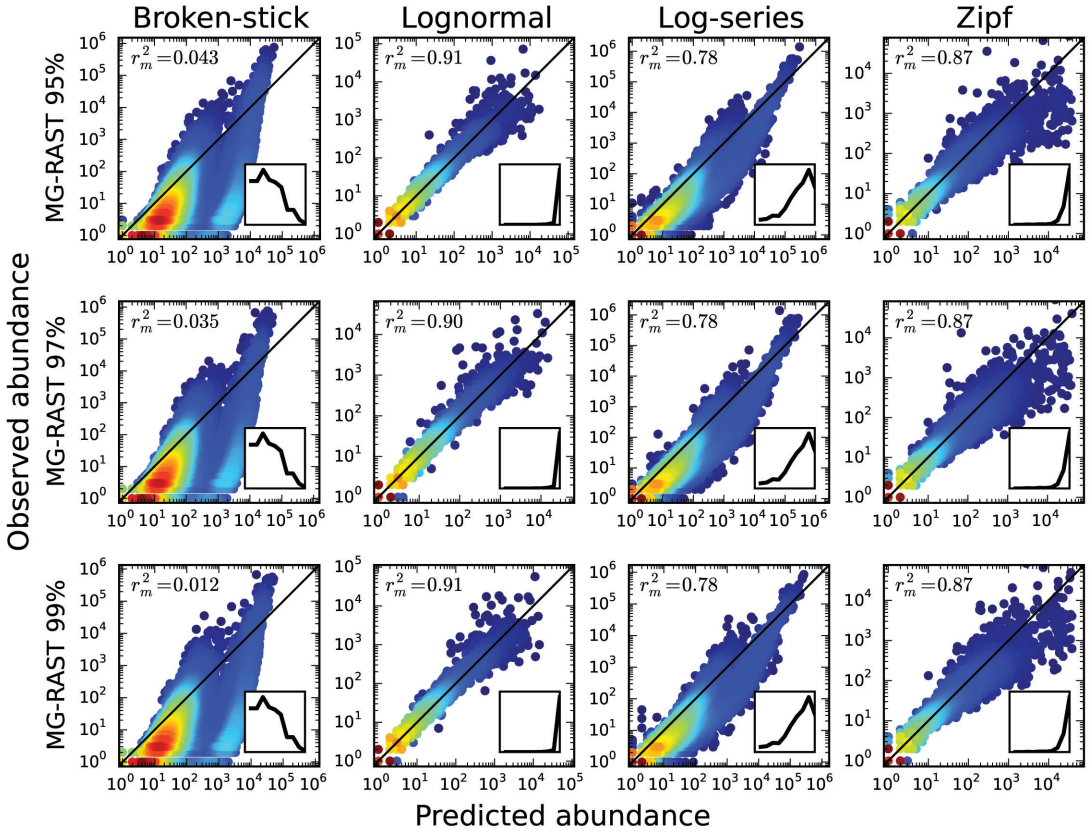
Supplementary Figures

1. The relationship between the predicted abundance of each rank and the observed abundance for different sequence similarity cutoffs.
2. The relationship between the predicted abundance of each rank and the observed abundance with singletons removed.
3. The relationship between model performance and the percent value of *N* used to sample SADs.
4. Kernel density estimates for the lognormal, log-series, and Zipf distribution from a single bootstrapped sample.
5. Kernel density estimates for the fitted parameters from the lognormal, log-series, and Zipf distribution.
6. A box-and-whisker plot of the percent of the time that each SAD model had the highest AICc weight.

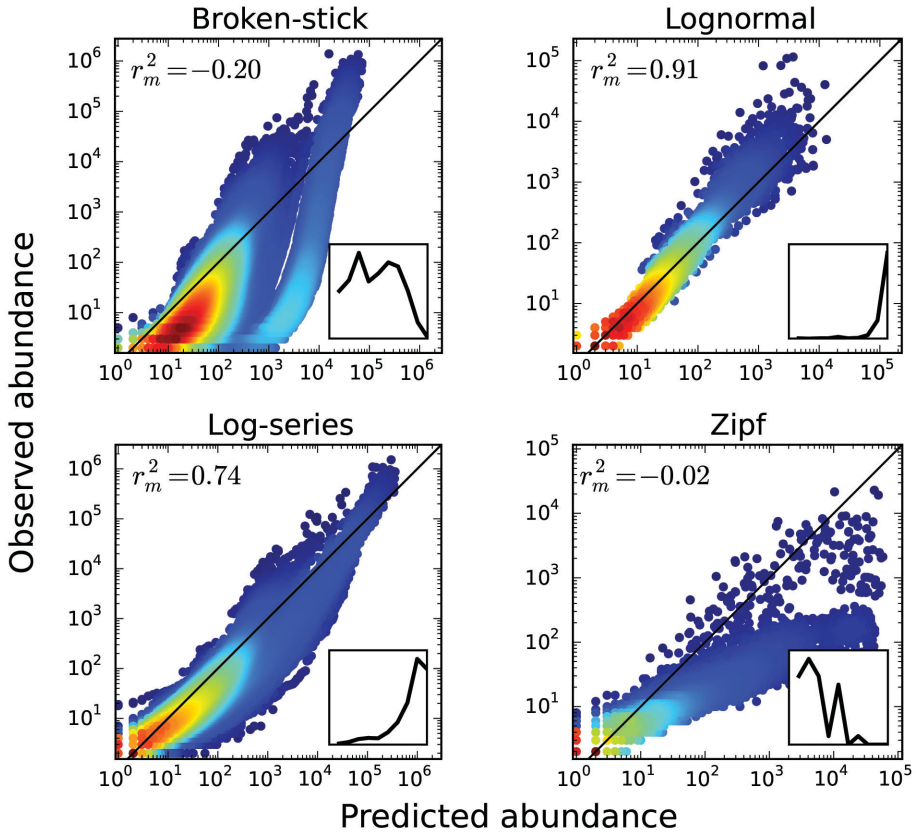
Supplementary Tables

1. The results of the simple linear regression of each SAD model from Figure 3 for a single bootstrap iteration.
2. Comparison of the performance of each species abundance distribution (SAD) model for different sequence similarity cutoffs.
3. Comparison of the performance of species abundance distribution (SAD) models for microbial datasets with singletons removed.
4. A comparison of the mean and standard deviation of the log-likelihood values for the lognormal, log-series, and Zipf distribution from 10,000 bootstrapped samples.
5. The mean and standard deviation of the fitted parameters for the lognormal, log-series, and Zipf distribution.
6. The mean and standard deviation of the fitted parameters for the lognormal, log-series, and Zipf distribution.

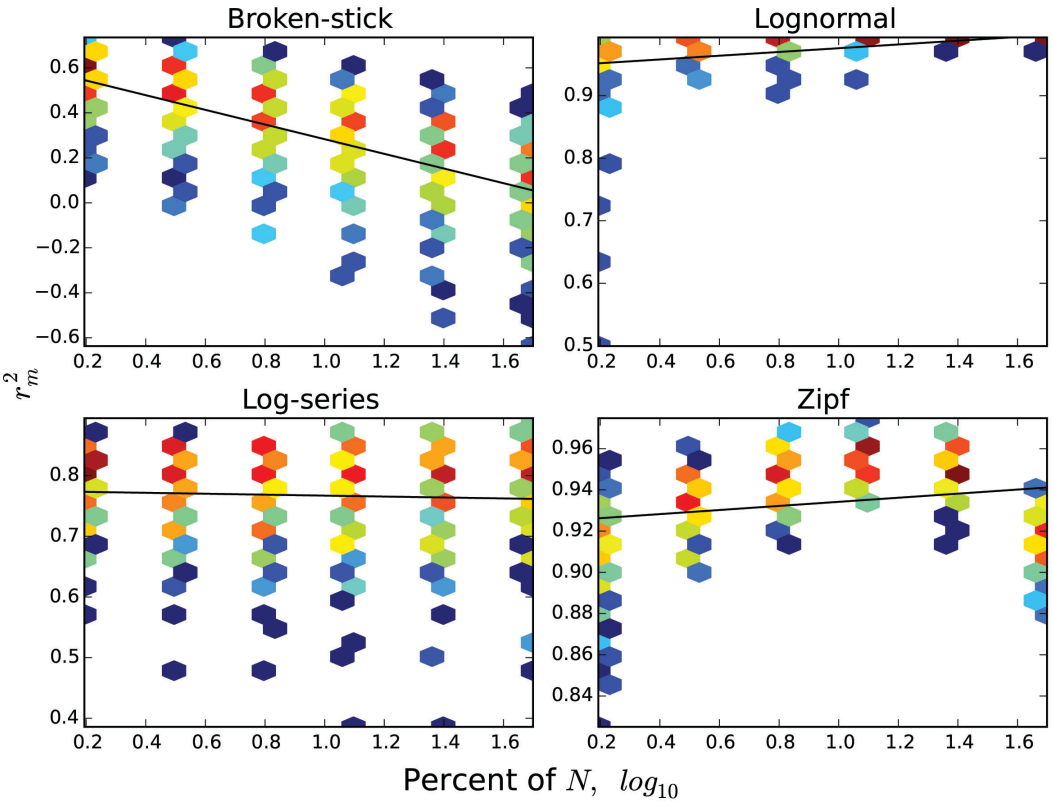
Supplementary Figure 1. The relationship between the predicted abundance of each rank and the observed abundance for each SAD model for sequence similarity cutoffs of 95, 97, and 99%. Sequence similarity had no measurable effect on model performance. Points are color-coded by the density of adjacent points. Hotter colors indicate a higher density of points.



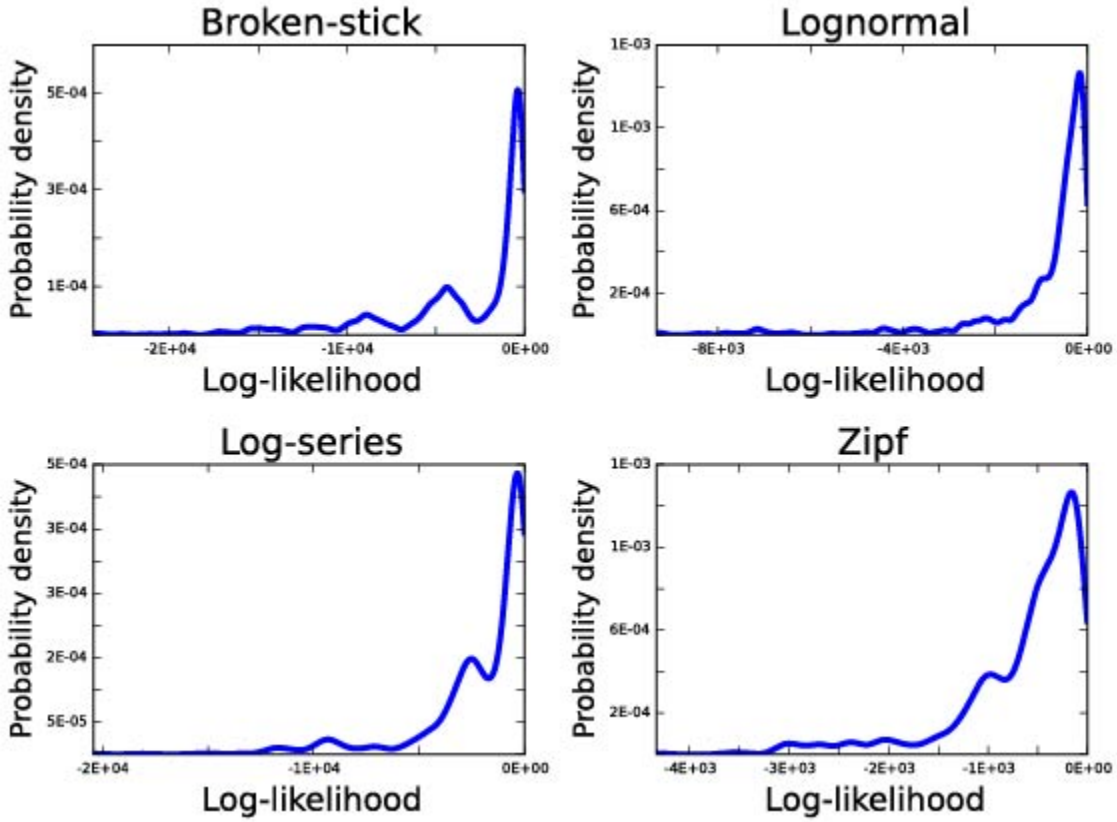
Supplementary Figure 2. The relationship between the predicted abundance of each rank and the observed abundance for each SAD model with singletons (i.e. species that have an abundance of one) removed. Removing singletons had little effect on the performance of the lognormal distribution. The performance of the log-series and Broken-stick increased while the performance of the Zipf distribution decreased, likely as a result of the models tendency to under and over predict, respectively, the number of singletons in a site. The value at the top-left of each sub-plot is the mean r^2_m value for 10,000 bootstrapped samples (see Methods). Points are color-coded by the density of adjacent points. Hotter colors indicate a higher density of points.



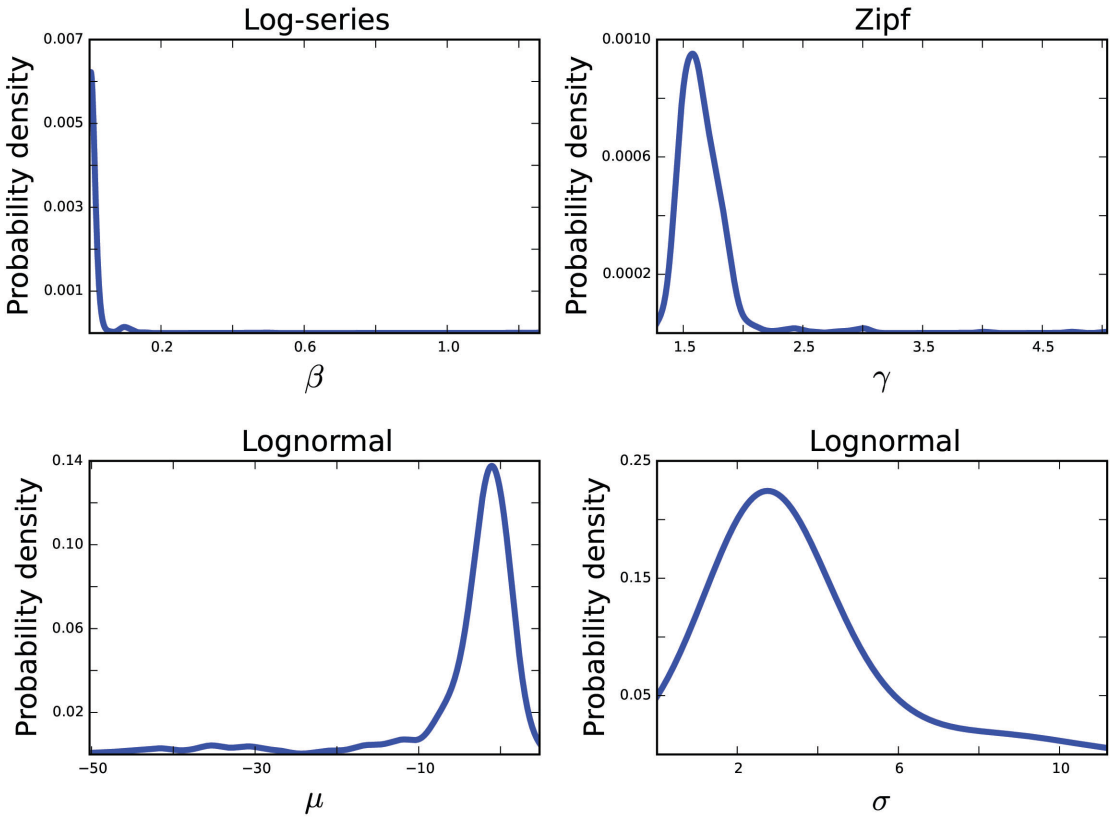
Supplementary Figure 3. The relationship between model performance (i.e. r^2_{m}) and the percent value of *N* used to sample SADs. The performance of the Broken-stick and the log-series decreases as *N* increases while the performance of the log-series increases as *N* increases. While the performance of the Zipf displays a humped-shaped relationship, performance increases up until a cut-off of 25%. Each sub-plot contains the results of 100 bootstrap iterations (see Methods).



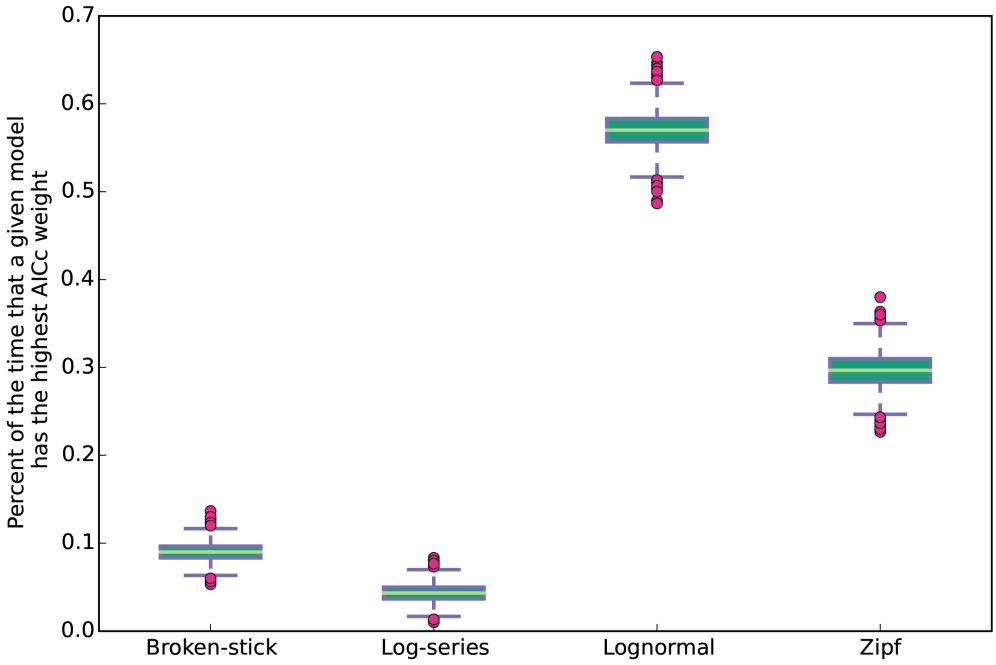
Supplementary Figure 4. Kernel density estimates for the lognormal, log-series, and Zipf distribution from a single bootstrapped sample. The lognormal and Zipf distribution had similar levels performance, while the log-likelihood of the log-series was larger by several orders of magnitude (Supplementary Table 4). In addition, the distribution of parameter values for all biodiversity models with a fitted parameter were highly peaked and unimodal. Kernel density estimates were selected based on cross-validation.



Supplementary Figure 5. Kernel density estimates for the fitted parameters from the lognormal, log-series, and Zipf distribution. The mean and standard deviation for each parameter can be found in Supplementary Table 5.



Supplementary Figure 6. A box-and-whisker plot of the percent of the time that each SAD model had the highest AICc weight. The mean and standard deviation of this data can be found in Supplementary Table 6.



Supplementary Table 1. The results of the simple linear regression of each SAD model from Figure 3 for a single iteration. Positive slopes indicate that model performance (r^2_m) increased as abundance (N) increased. The lognormal and Zipf provide better explanations of microbial SADs as N increases than the Broken-stick or the log-series.

Model	**Slope**	r^2	$p - value$
Lognormal	0.022	0.058	< 0.0001
Zipf	0.064	0.18	< 0.0001
Log-series	-0.010	0.00057	0.52
Broken-stick	-0.51	0.20	< 0.0001

Supplementary Table 2. Comparison of the performance of each species abundance distribution (SAD) model for sequence similarity cutoffs of 95, 97, and 99%. Results are reported as the reported as the mean and standard deviation of $r^2_{m.}$ Sequence similarity has little measurable effect on model performance. Additional information pertaining to the table can be found in the description for Table 1.

Model	Sequence similarity	$\overline{r^2_m}$	$\sigma_{r^2_m}$
Lognormal	95 %	0.91	0.11
	97 %	0.90	0.14
	99 %	0.91	0.12
Zipf	95 %	0.87	0.080
	97 %	0.87	0.11
	99 %	0.87	0.076
Log-series	95 %	0.78	0.18
	97 %	0.78	0.19
	99 %	0.78	0.18
Broken-stick	95 %	0.043	0.66
	97 %	0.034	0.67
	99 %	0.012	0.67

Supplementary Table 3. Comparison of the performance of species abundance distribution (SAD) models for microbial datasets with singletons removed. A sample of 100 SADs was randomly sampled from each dataset 10,000 times and reported as the mean and standard deviation of $r^2_{m.}$ Removing singletons had no measurable effect on model performance. Additional information pertaining to the results summarized in this table can be found in the description of Table 1.

Model	$\overline{r^2_m}$	$\sigma_{\bar{r^2_m}}$
Lognormal	0.91	0.087
Zipf	−0.021	0.73
Log-series	0.74	0.28
Broken-stick	−0.20	0.82

Supplementary Table 4. A comparison of the mean and standard deviation of the log-likelihood values for the lognormal, log-series, and Zipf distribution from 10,000 bootstrapped samples. Similar to the results using the r^2_m, the Zipf and lognormal perform similarly well while the log-series performs poorly. These samples are drawn from the set of sites that a given model was able to arrive at a prediction. The log-likelihood values reported here are for all sites where numerical estimation arrived a solution for a given model.

Model	$\mu_{Log-likelihood}$	$\sigma_{Log-likelihood}$
Lognormal	-970	47
Zipf	-630	19
Log-series	-2000	89
Broken-stick	-3400	130

Supplementary Table 5. The mean and standard deviation of the fitted parameters for the lognormal, log-series, and Zipf distribution. The Broken-stick has no fitted parameters and is not included in this table.

Model	Parameter	Mean	Standard deviation
Lognormal	μ	1.8	0.044
	σ	1.9	0.029
Zipf	γ	1.4	0.0038
Log-series	β	0.0055	0.00069

Supplementary Table 6. The mean and standard deviation for the percent of the time that a given SAD model has the highest corrected Akaike Information Criterion (AICc) weight from 10,000 bootstrap samples. Because the AICc needs to be calculated on a site-by-site basis, the values in this table were calculated using the set of sites where all models arrived at a successful prediction. AICc values were calculated using the set of sites where numerical estimation arrived at a successful prediction for all SAD models.

Model	$\mu_{\%\, winning}$	$\sigma_{\%\, winning}$
Lognormal	57.0	2.1
Zipf	30.0	2.0
Log-series	4.3	1.0
Broken-stick	9.0	1.1