Supplementary Information Appendix

Figure S1. Species rarity. In the main body of the manuscript we present results for how rarity (log-modulo of skewness; **left**) scales with sample abundance, i.e., the number of individual organisms or gene reads in a sample (*N*) (see Fig. 1a). The log-modulo transformation adds a value of one to each measure of skewness and converts negative values to positive values, making them all positive and able to be log-transformed. The analysis showed similar scaling but a greater intercept for microbes, revealing greater rarity. We also quantified rarity as a logarithmically transformed measure of skewness (1) (**right**), however, this relationship which also showed increasing rarity (as decreasing log-skew) was weaker than the relationship based on the log-modulo transformation of skewness. Consequently, we used the log-modulo measure in the main body and for the main result. The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

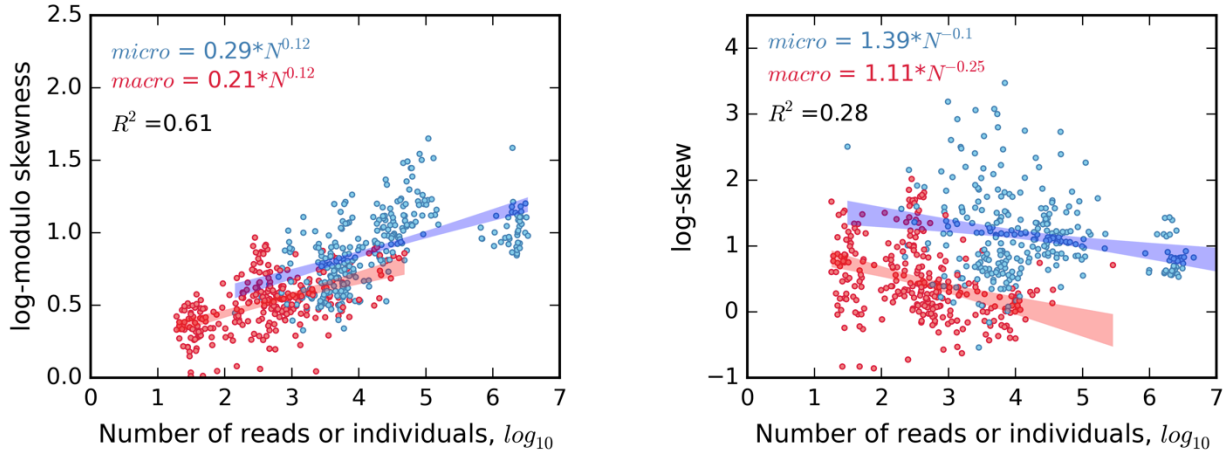


Figure S2. Dominance. In the main body of the manuscript we present results for how the number of individuals or gene reads belonging to the most abundant species (N_{max}) or species-level taxon scales with sample abundance, i.e., number of individual organisms or gene reads detected (N) (see Fig. 1b). For N_{max}, we observed strong and largely similar scaling slopes for microbes and macrobes. Because *Nmax* is an absolute measure of dominance and because the relationship is nearly isometric (i.e. nearly 1:1), we would expect no relationship for relative measures of dominance such as McNaughton's measure (% relative abundance of the two most abundant taxa), the Berger-Parker index (relative abundance of the single most abundant taxa), nor Simpson's Diversity (probability that the next sampled individual belongs to a different species) (1). The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

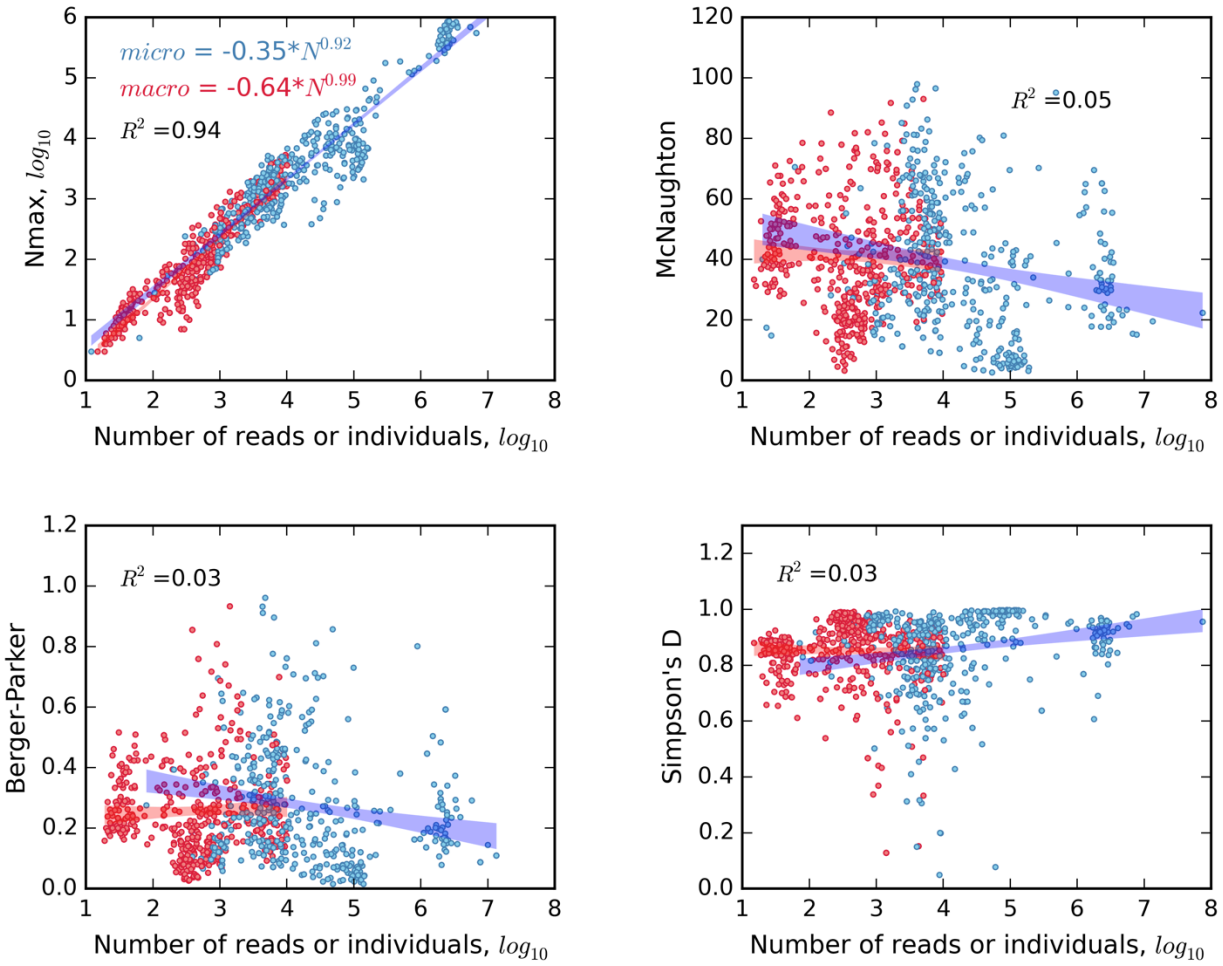


Figure S3. Species evenness. In the main body of the manuscript we presented results for how similarity in abundance (i.e. Simpson's evenness; ESimp) relates to the number of individual organisms or individual gene reads (*N*) (see Fig. 1c). We also observed similar slopes for microbes and macrobes using Heip's evenness index, Smith and Wilson's evenness index (Evar), and the O evenness index (See Methods). Slopes differ more greatly when using Evar, which gives less weight to highly abundant species than do other indices. The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

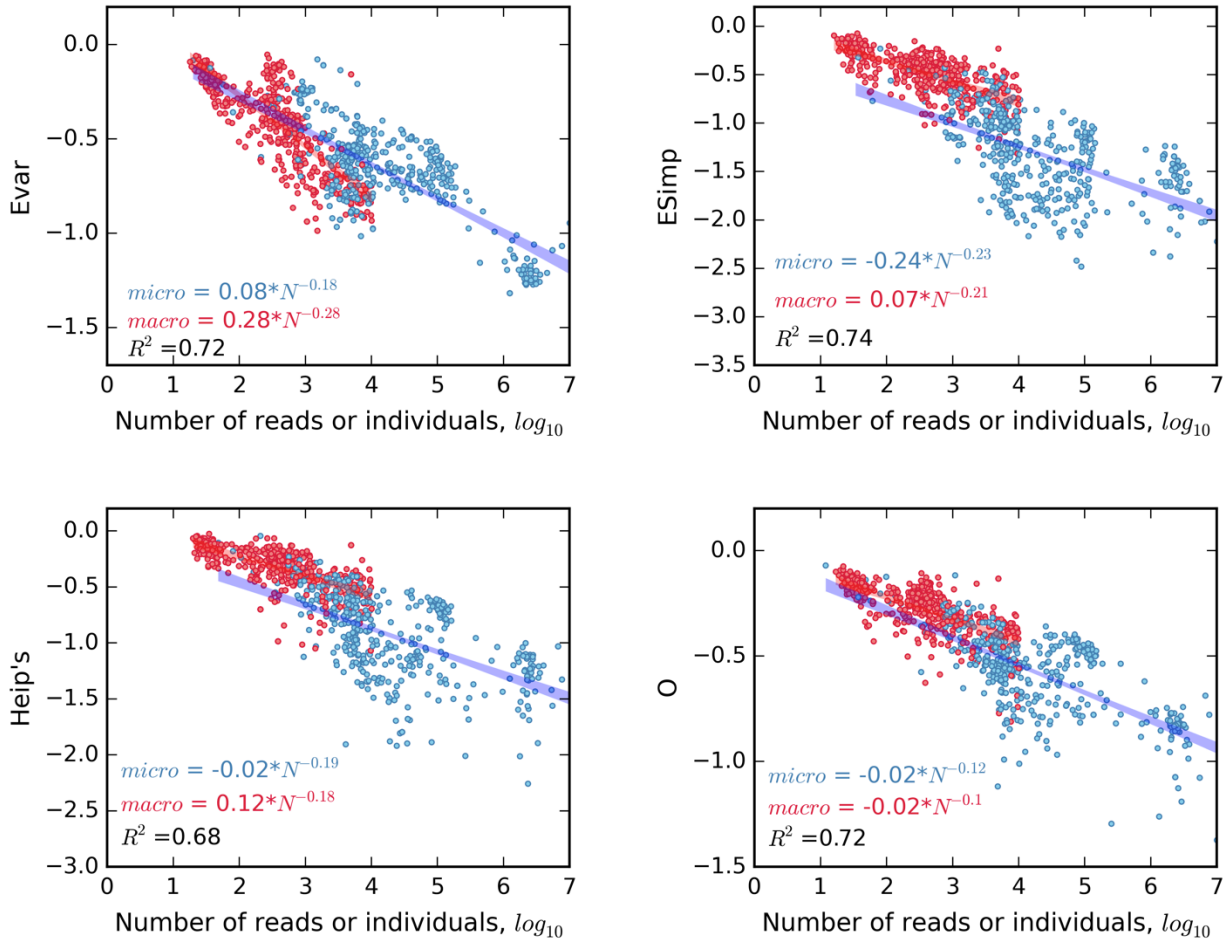


Figure S4. Species Richness. In the main body of the manuscript we present results for how observed numbers of species or species level taxonomic units (for microbes) relate to sample abundance, i.e., number of individual organisms or gene reads detected (*N*) (see Fig. 1d). We observed a steeper relationship and stronger scaling for microbes than macrobes. These results were qualitatively similar to estimates of richness: Chao1, ACE, Jackknife1, and Margalef's. These additional results reveal the same qualitative pattern and for all but Margalef's index, the same quantitative result. The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

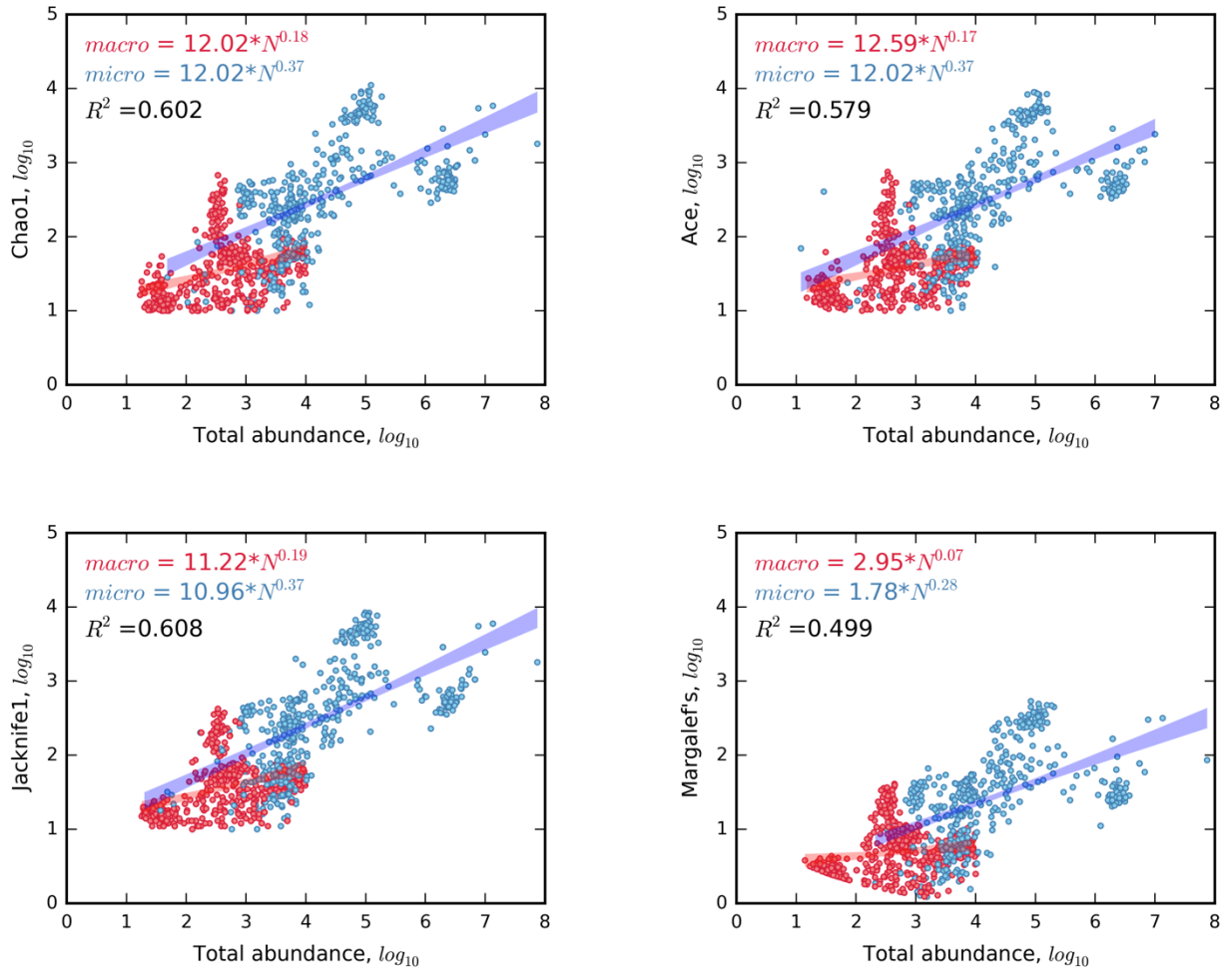


Figure S5. Robust responses to samples size. Our analyses relied on ordinary least squares regression, which includes several assumptions, not all of which are fatal when violated. We tested assumptions of linearity, normality, homoscedasticity (no change in error structure across the *x*-axis), and serial correlation, across a range of sample sizes because larger samples are more likely to uncover a real difference (greater statistical power) but are more likely to fail parametric tests of regression assumptions. While passing parametric tests depended on sample size (i.e. number of sites chosen from each dataset), where larger samples resulted in *p*-values less than 0.05 the regression model coefficients and the coefficients of determination (R^2) were independent of sample size. In particular, the assumption of linearity and normally distributed error terms were generally well-supported.

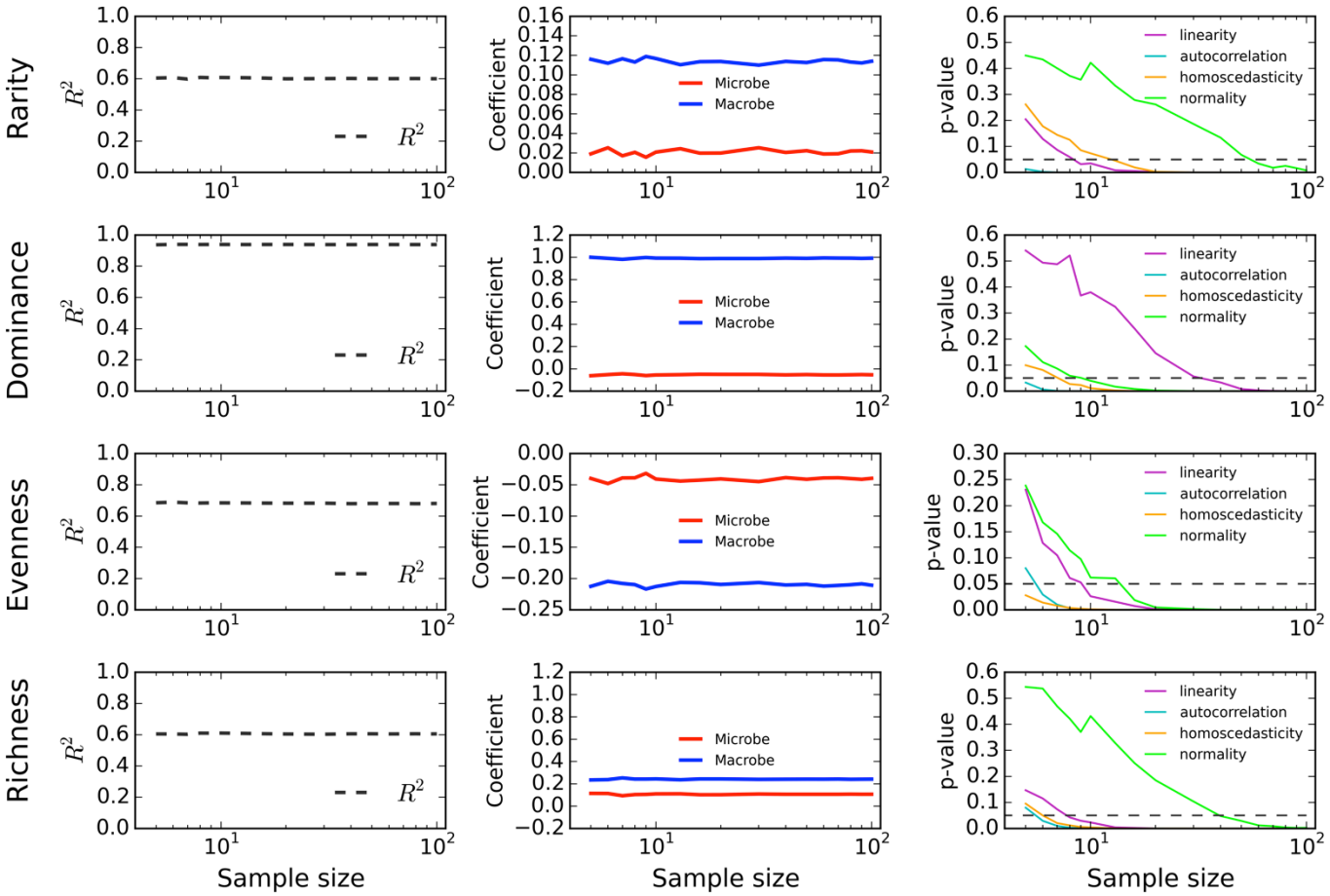


Figure S6. Testing the effect of categorical variable through random reassignment. We randomly reassigned sites to the microbe/macrobe categorical variable to reveal that identical model parameters can be obtained when the categorical variable is basically ignored. The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

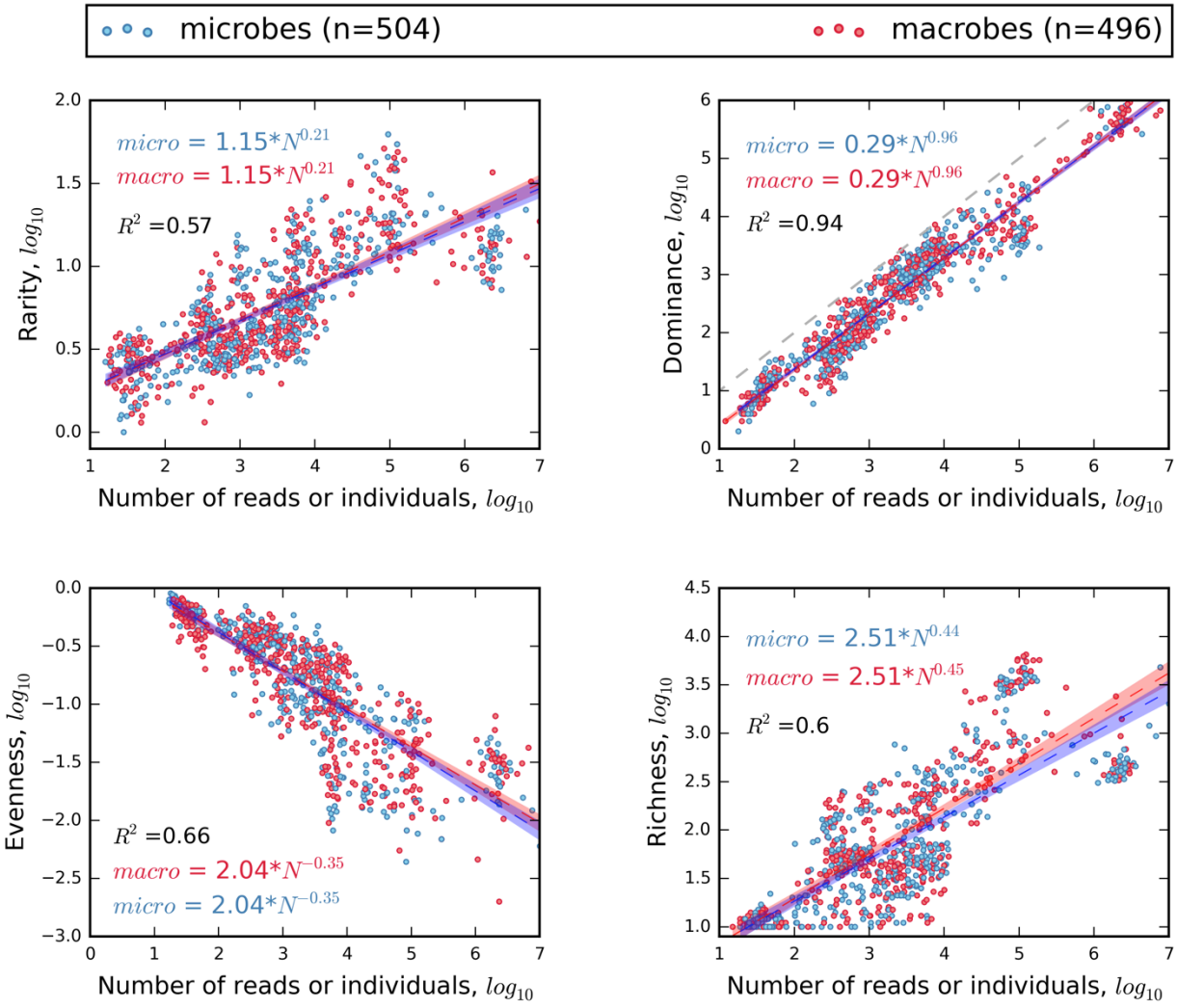
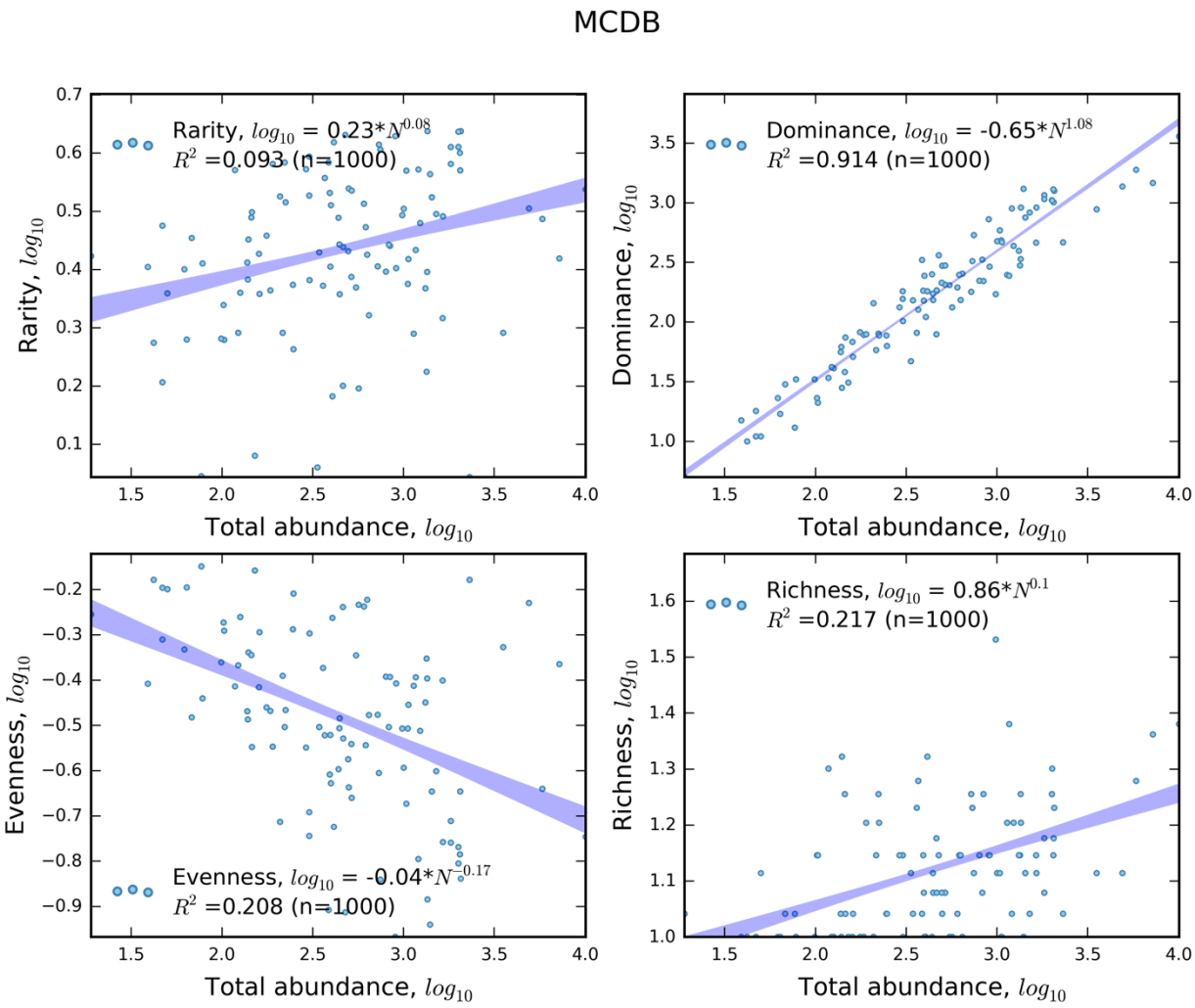
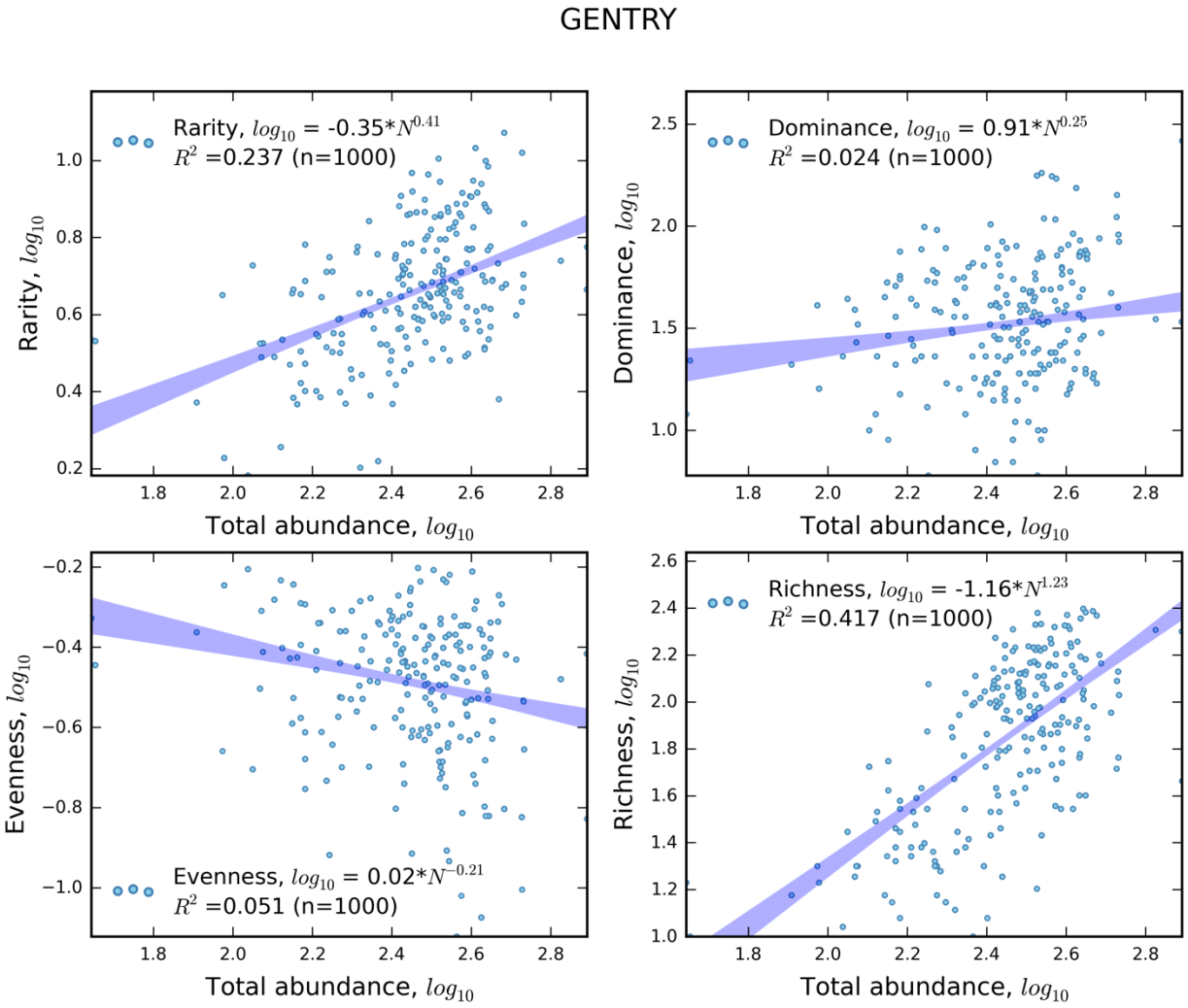


Figure S7. A-I. Results per dataset. The following figures (each with four subplots) show how aspects of diversity relate to sample abundance (*N*), i.e., the number of individual organisms or gene reads detected. The metrics are the same as those used in Fig. 1 in the main body, that is rarity (log-modulo skewness), dominance (N_{max}), Simpson's evenness metric, and observed richness (*S*). While the exact form and strength of the relationships vary, most relationships for each dataset follow the same direction, i.e., for each relationship: increasing for rarity, dominance, and richness, and decreasing evenness.

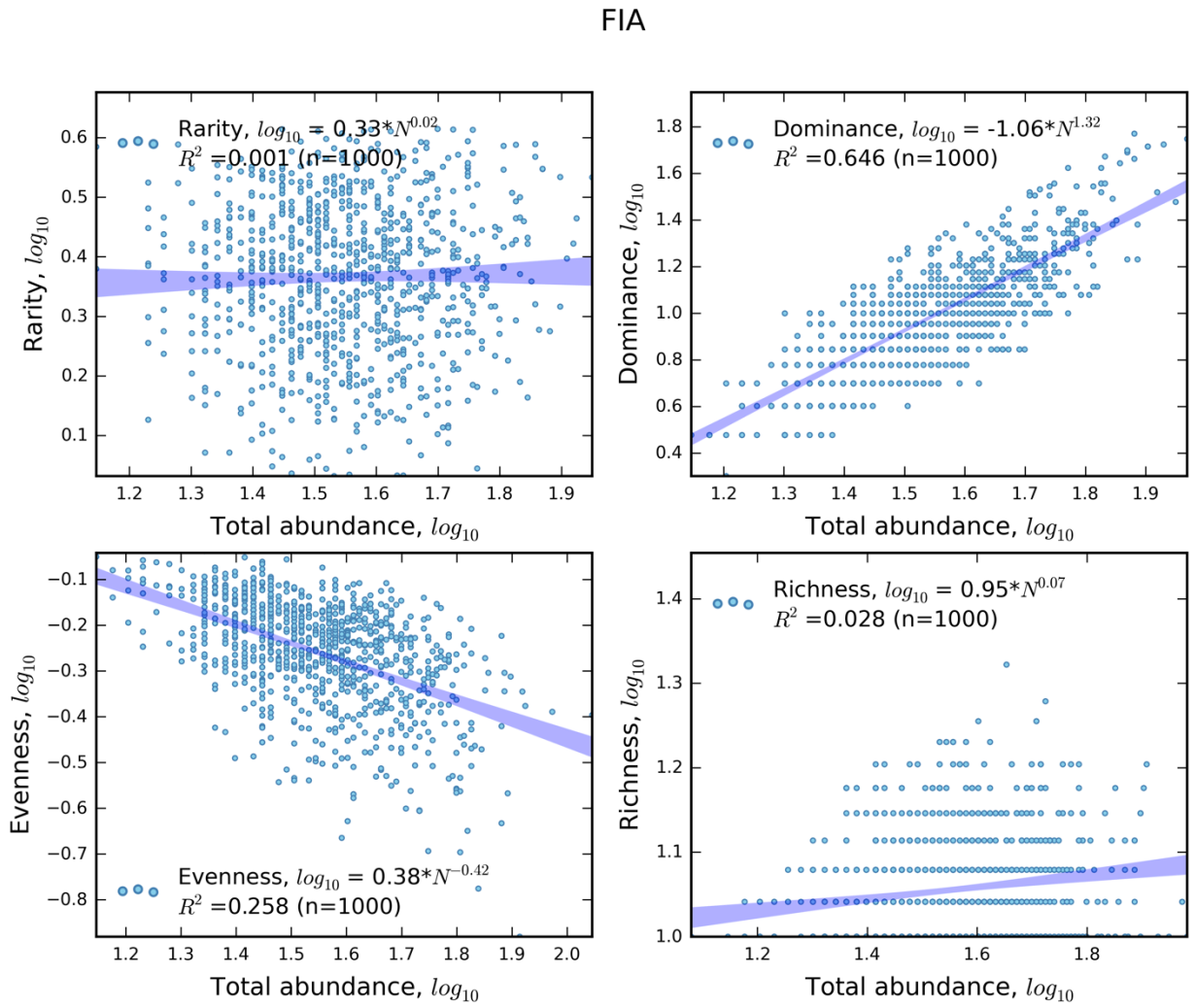
Sub-figure A. Mammal Community Database (MCDB)



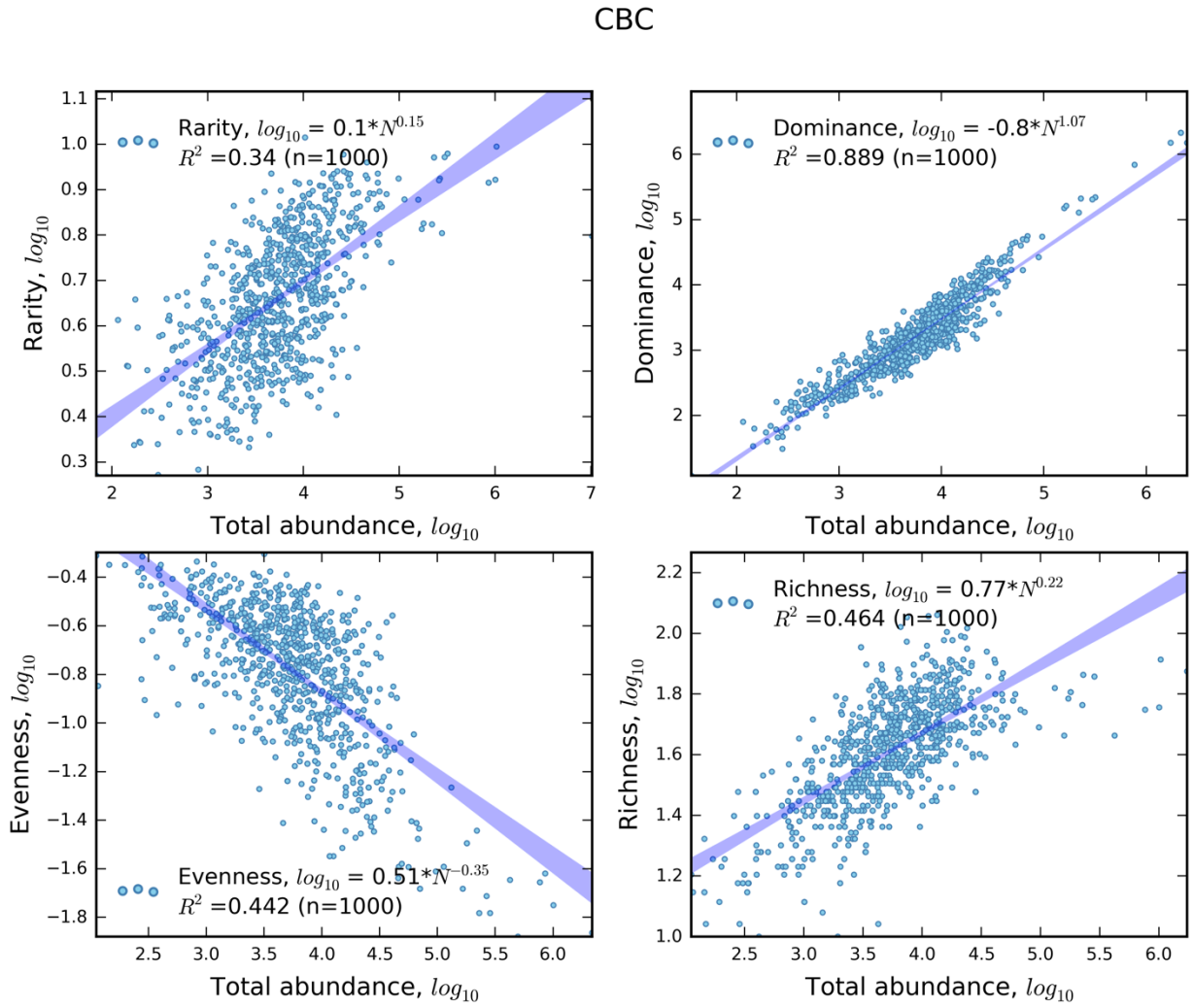
Sub-figure B. Alwyn Gentry's Forest Transects (GENTRY)



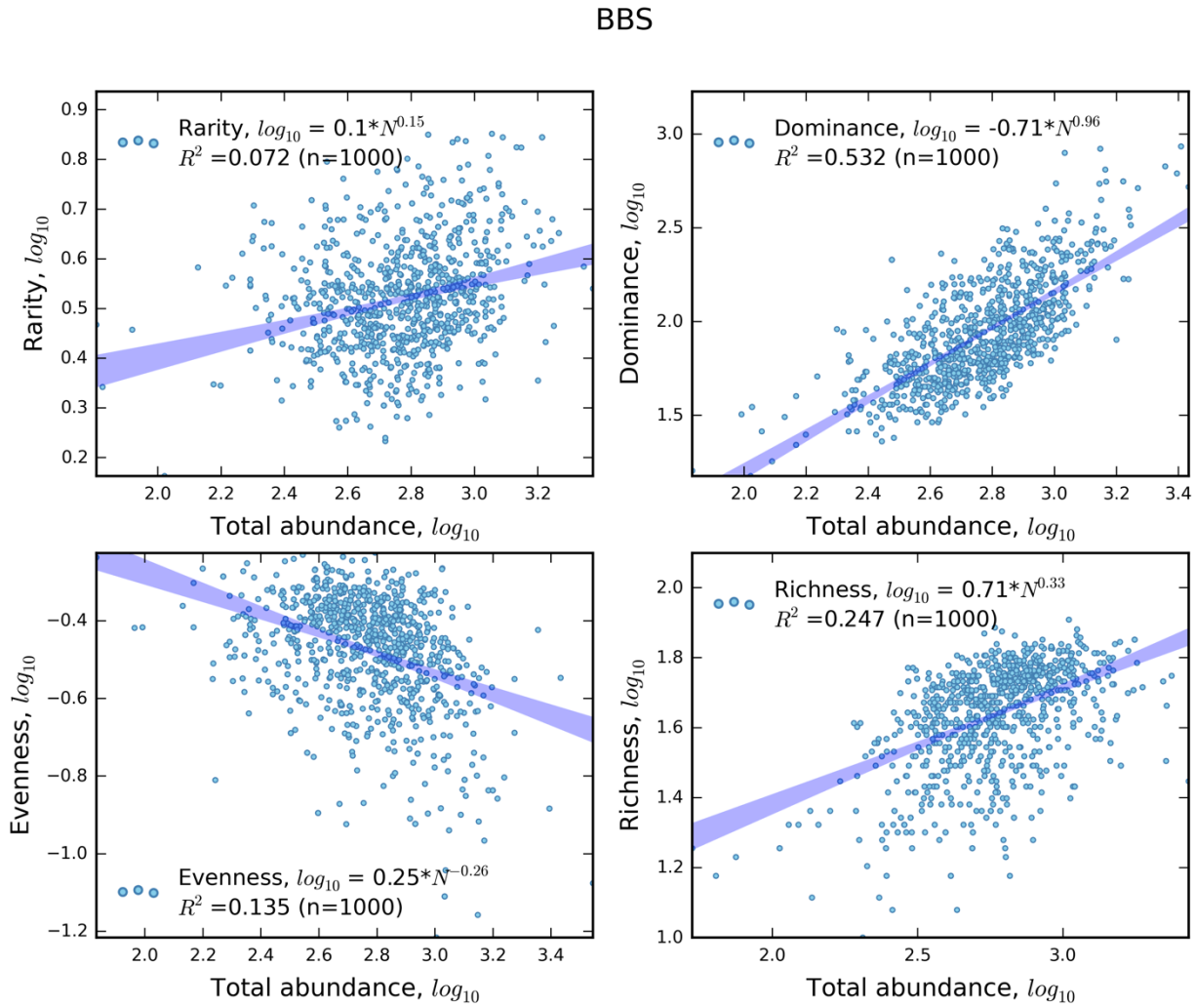
Sub-figure C. United States Department of Agriculture (USDA) Forest Inventory and Analysis dataset (FIA).



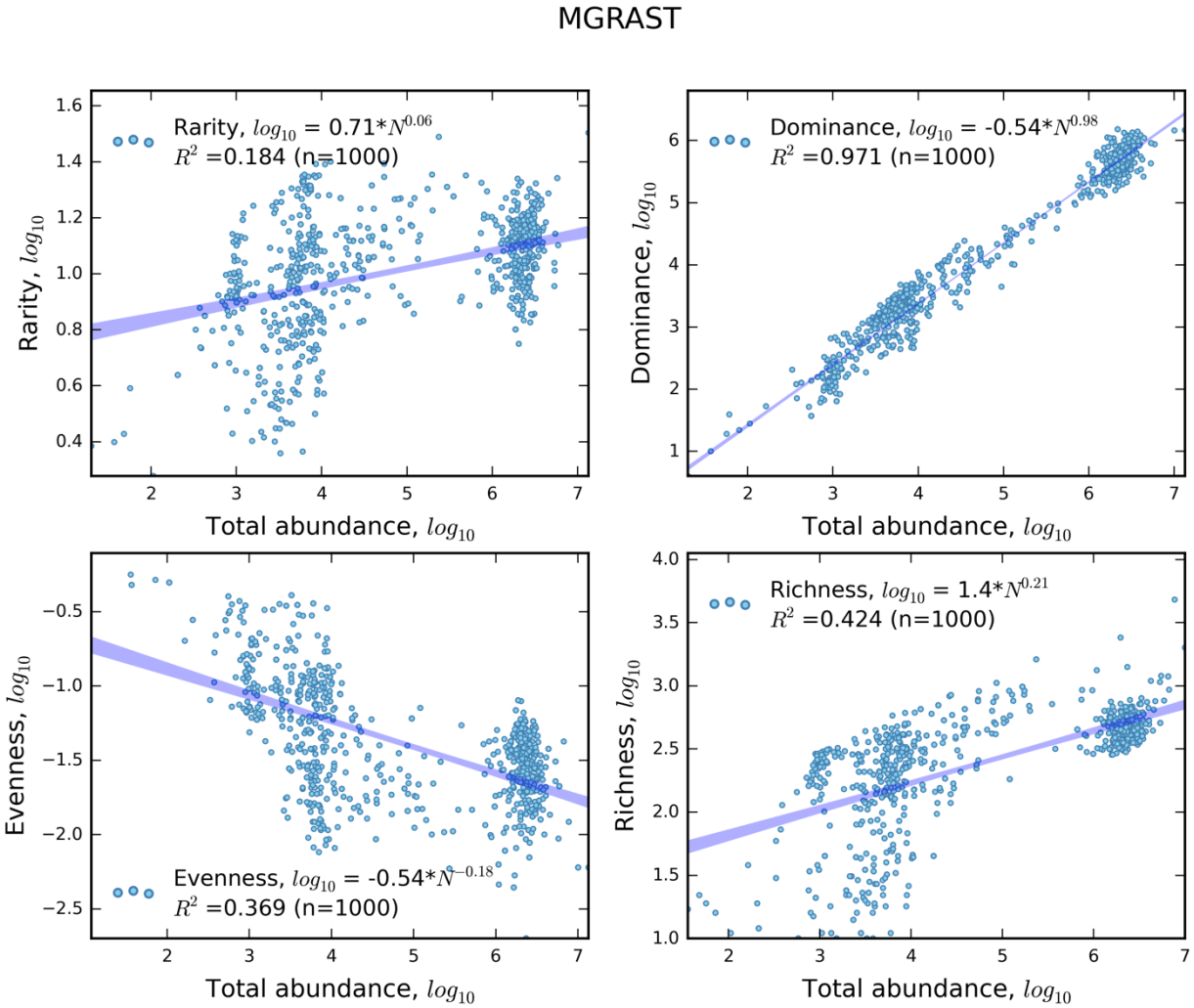
Sub-figure D. National Audubon Society's Christmas Bird Count (CBC)



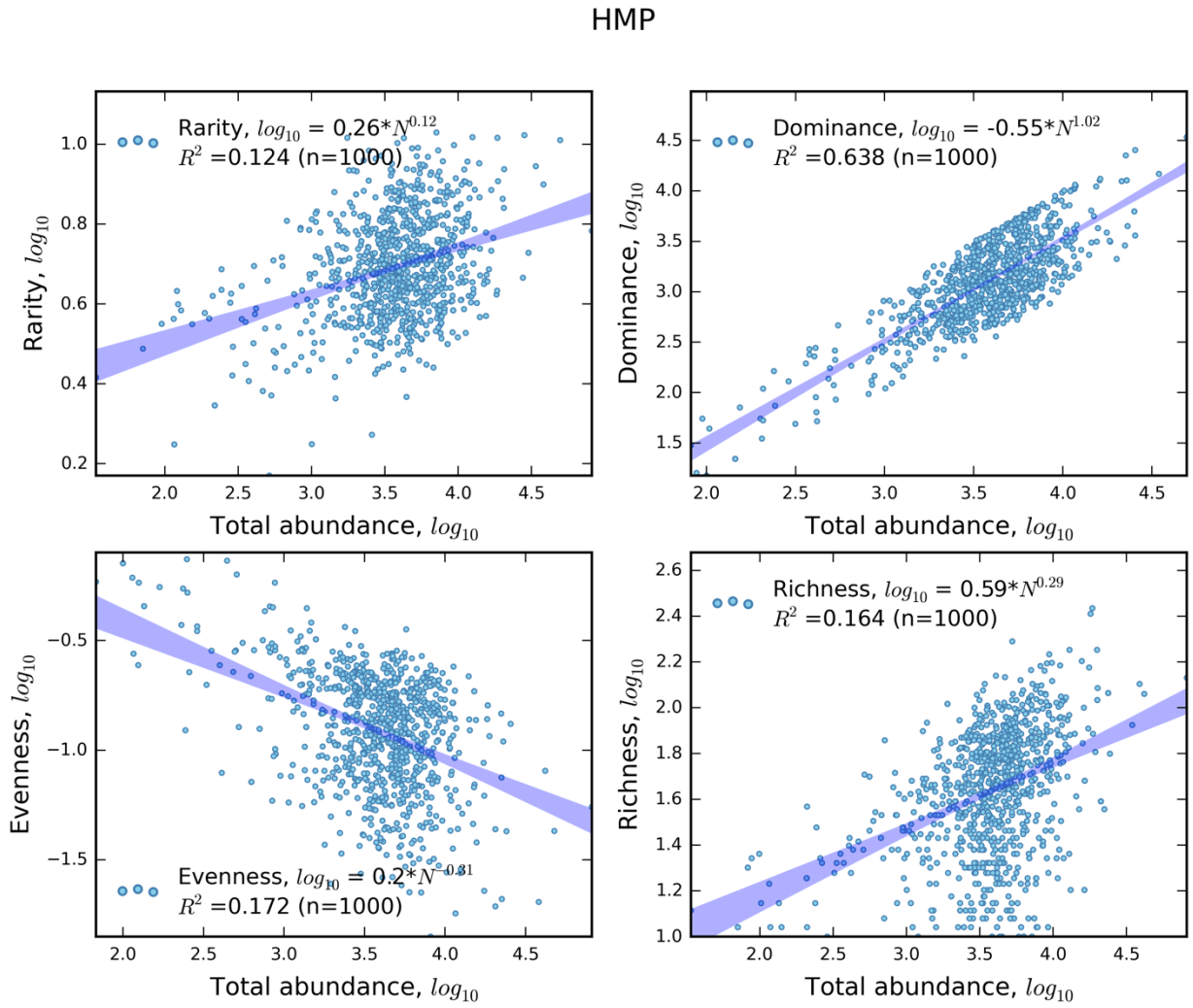
Sub-figure E. North American Breeding Bird Survey (BBS)



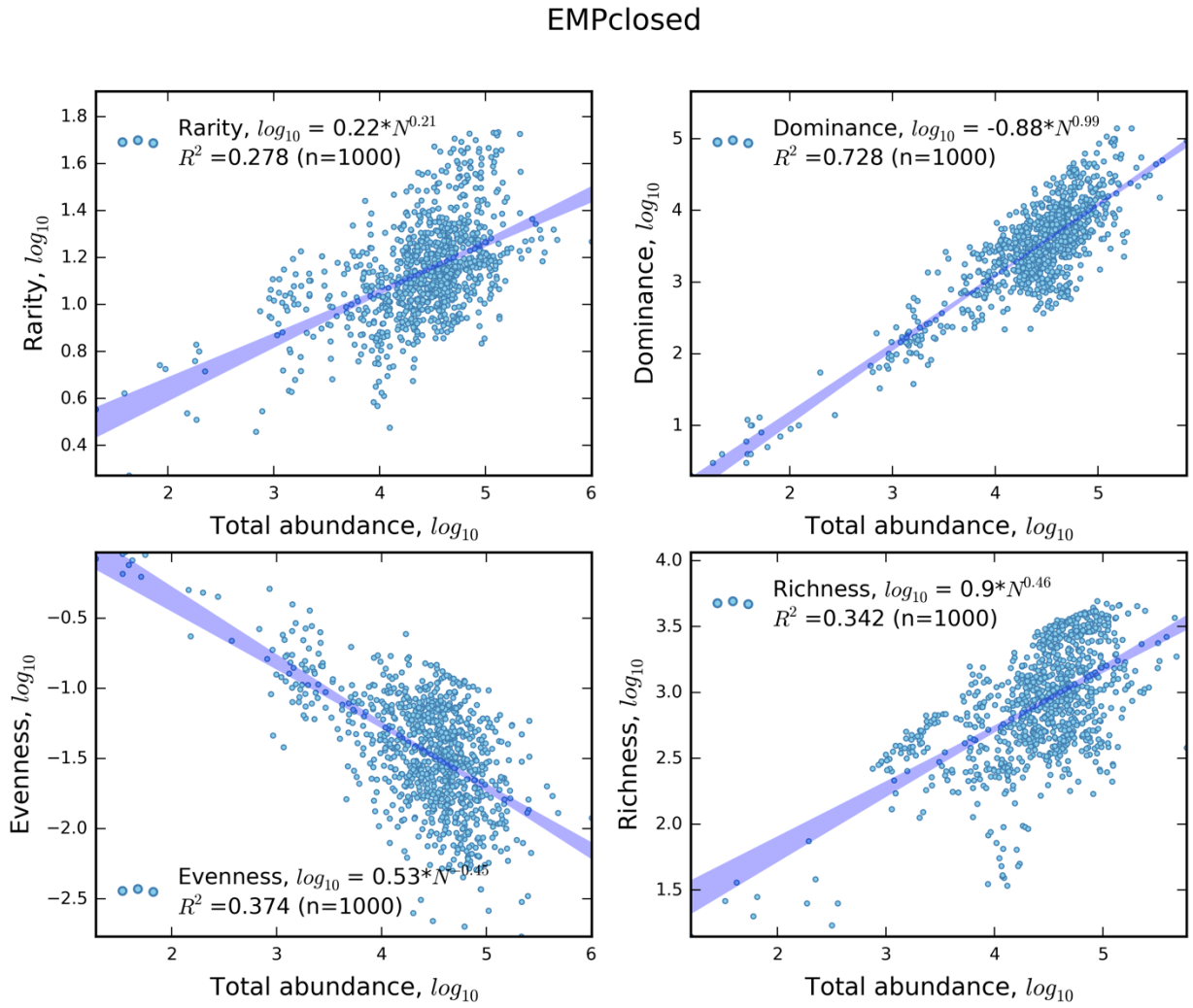
Sub-figure F. Data obtained from projects uploaded to the National Argonne Laboratories' metagenomic server MG-RAST.



Sub-figure G. Human Microbiome Project (HMP)



Sub-figure H. Earth Microbiome Project, closed reference OTU data (EMPclosed)



Sub-figure I. TARA Oceans expedition (“prokaryote” samples).

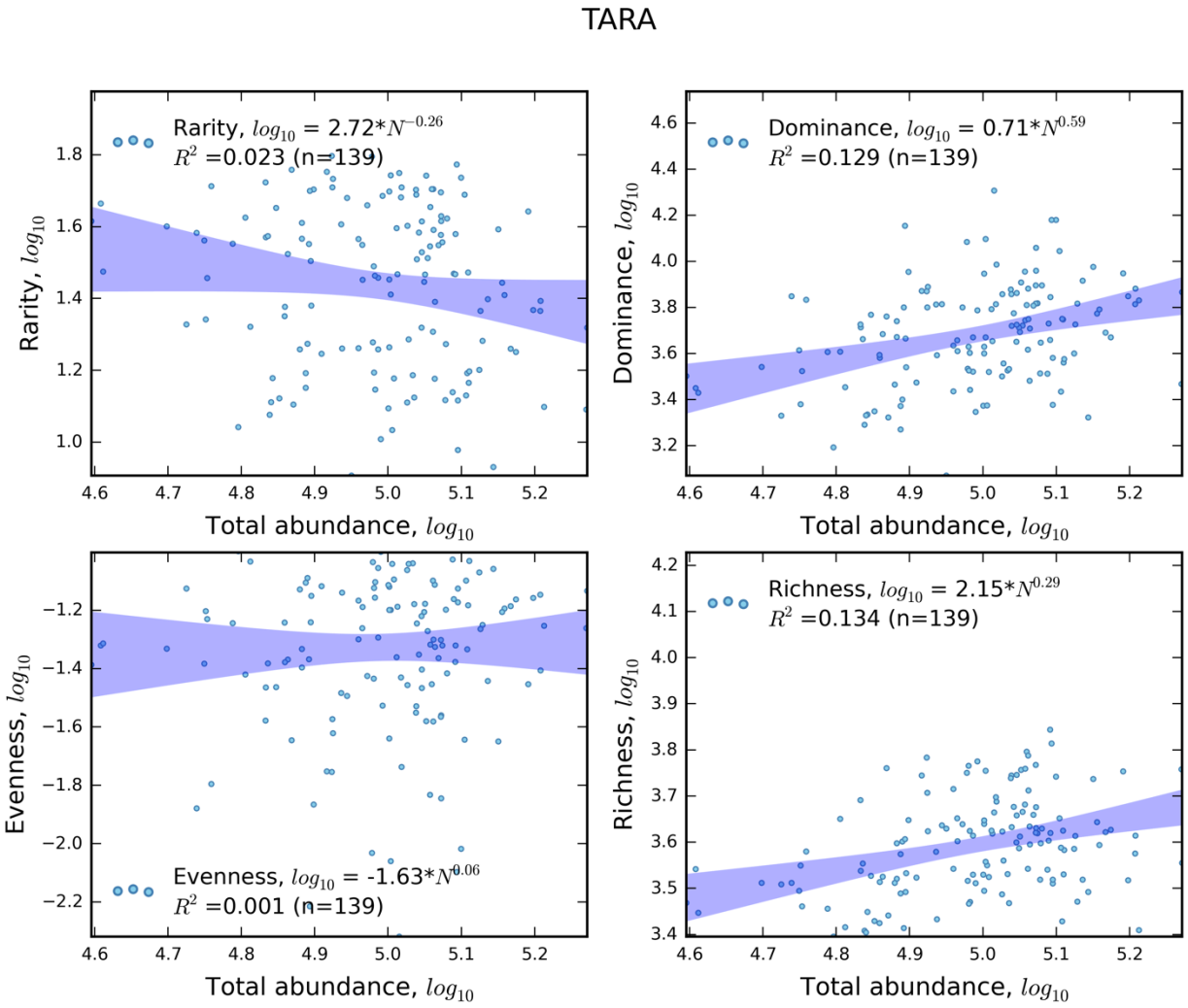


Figure S8. Binning taxa according to 95, 97, and 99 percent sequence similarity among 16S rRNA genes did not affect our results. Here, we use a subset our data from MG-RAST to show that relationships of diversity do not differ when using 95, 97, or 99% similarity. The metrics are the same as those used in Fig. 1 in the main body, that is rarity (log-modulo skewness), dominance (N_{max}), Simpson's evenness metric, and observed richness (S). The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

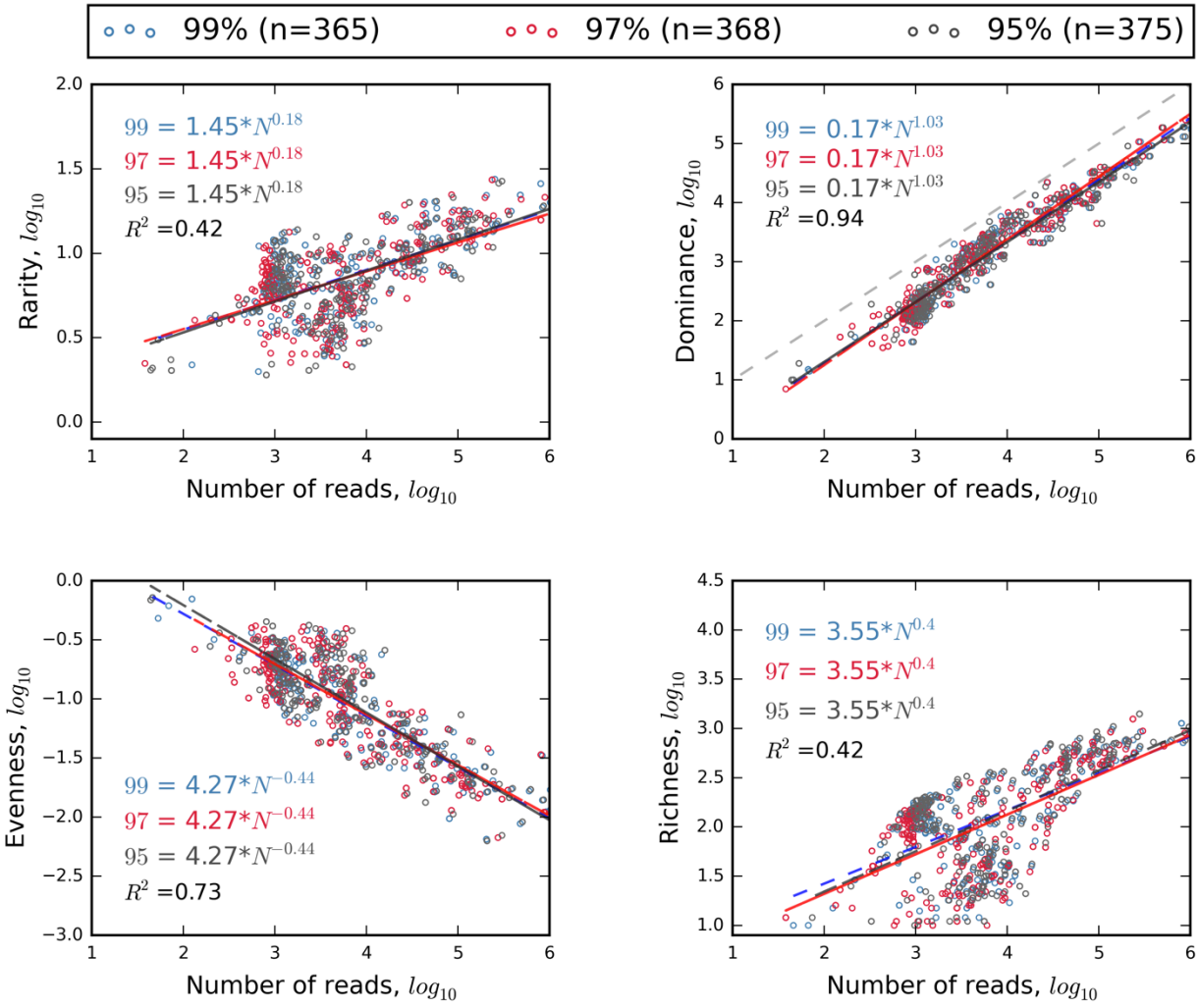


Figure S9. Including and excluding singleton taxa among microbes did not affect our results. There are some caveats associated with making microbial species assignments based on operational taxonomic units (OTUs). However, we found no substantial differences when either including or excluding microbial singletons. The metrics are the same as in Fig. 1 of the main body, that is rarity (log-modulo skewness), dominance (N_{max}), Simpson's evenness metric, and observed richness (S). The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

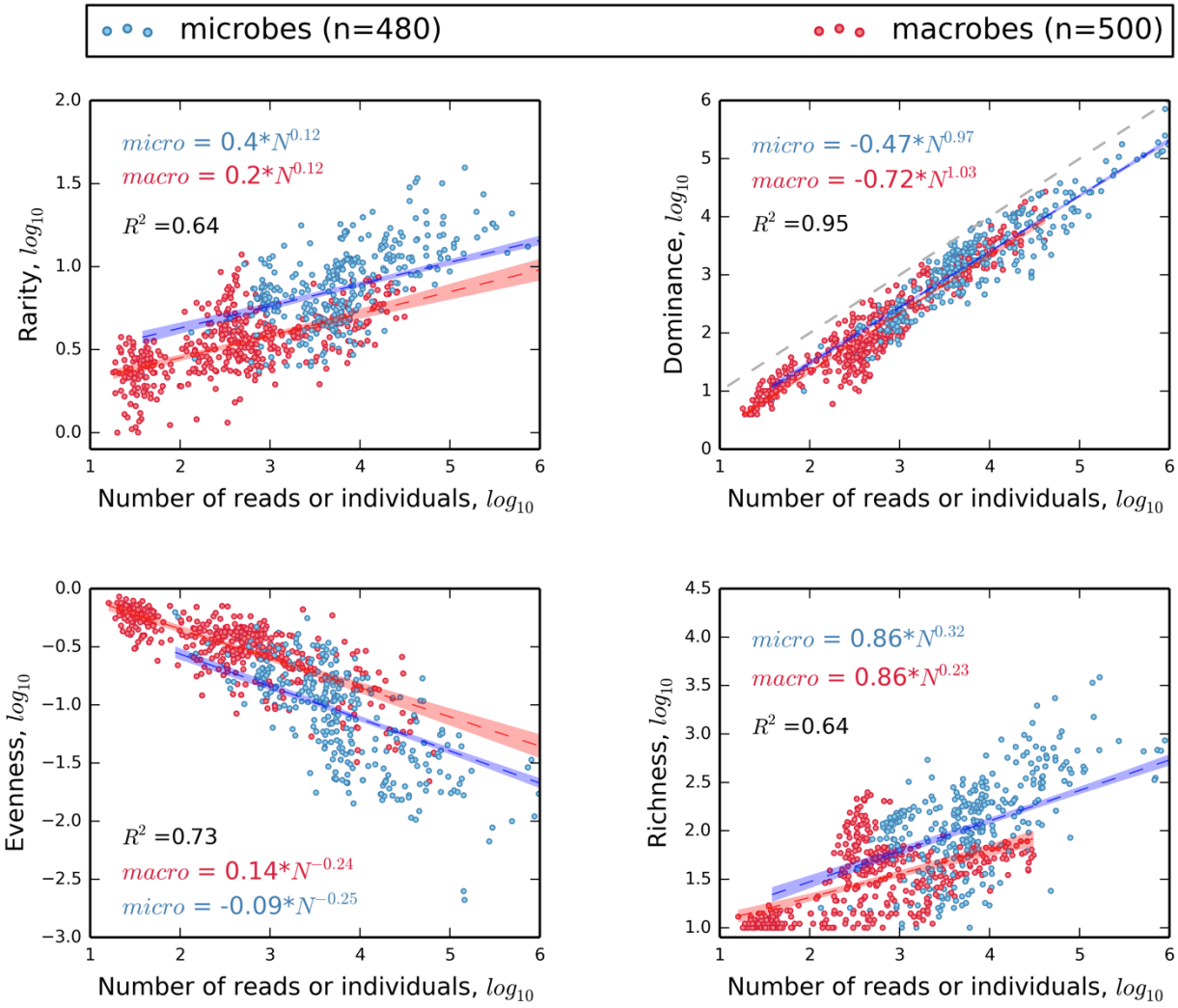


Fig S10. When using only MG-RAST data, none of the scaling relationships differ between macrobes and microbes. The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

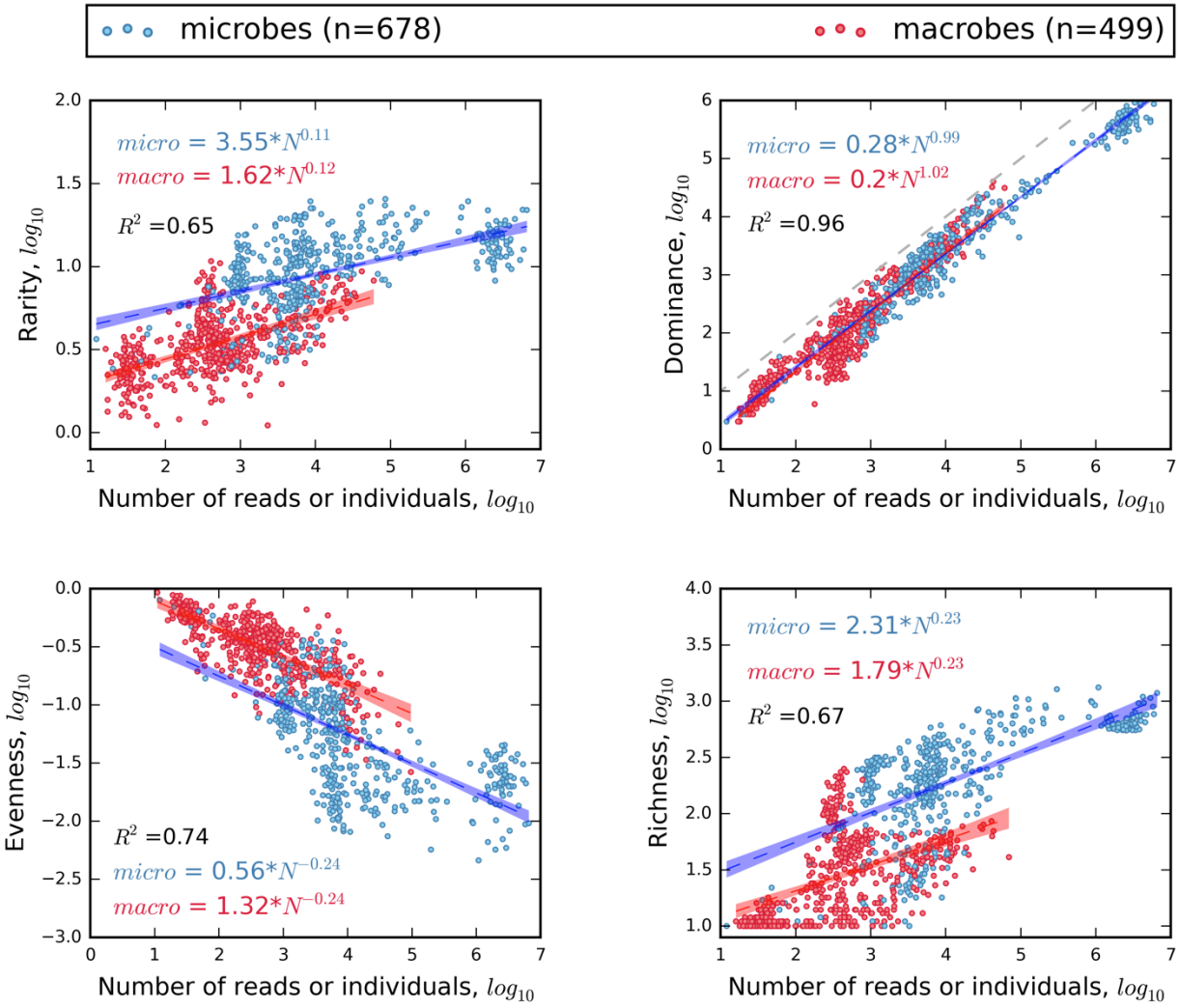


Fig S11. When using only Human Microbiome Project data, only the scaling of species evenness to total abundance (*N*) appears to differ. The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

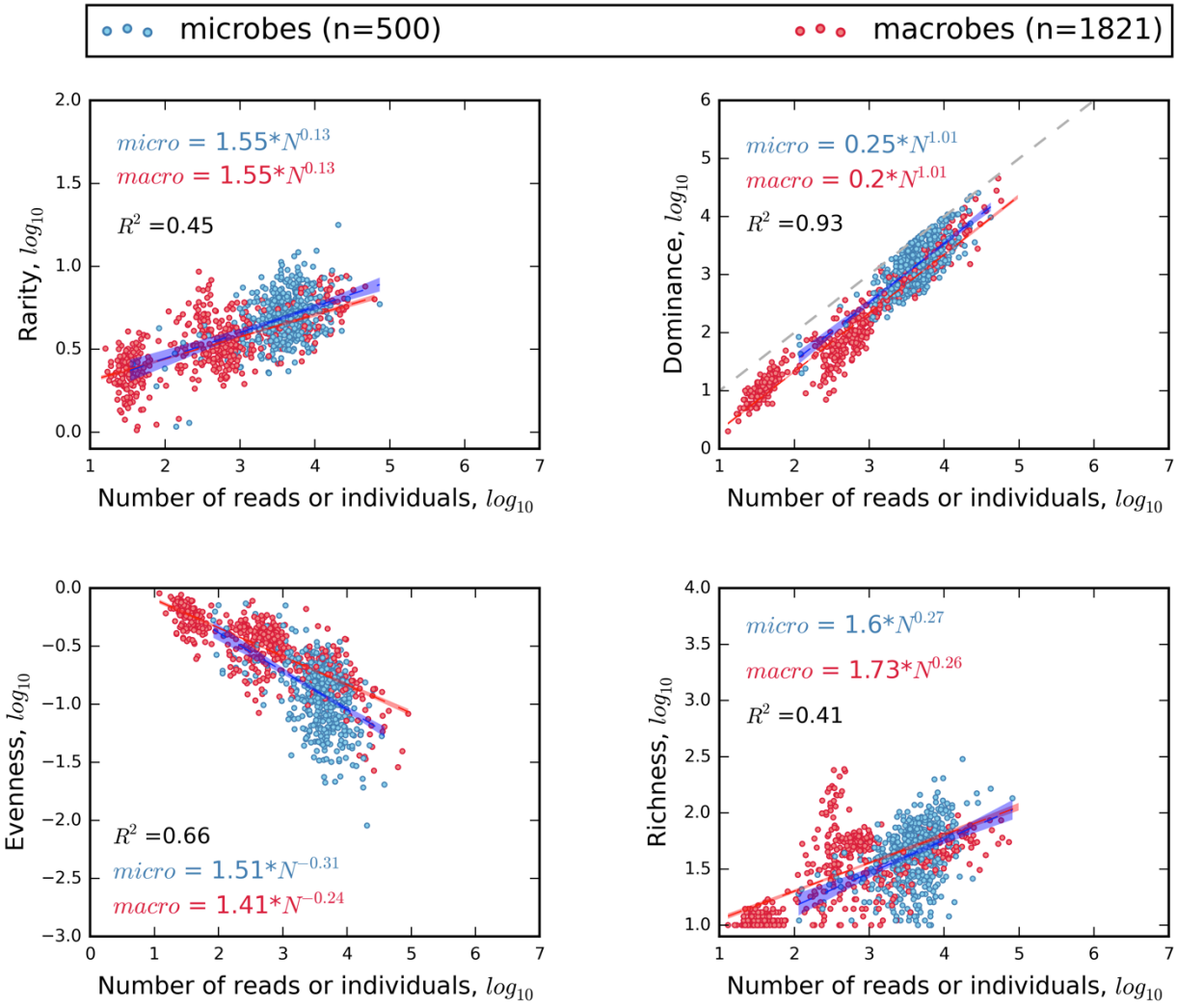


Fig S12. When using only EMP closed reference data, all scaling relationships differ between macrobes and microbes, except for the dominance relationship, which remains nearly isometric. The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

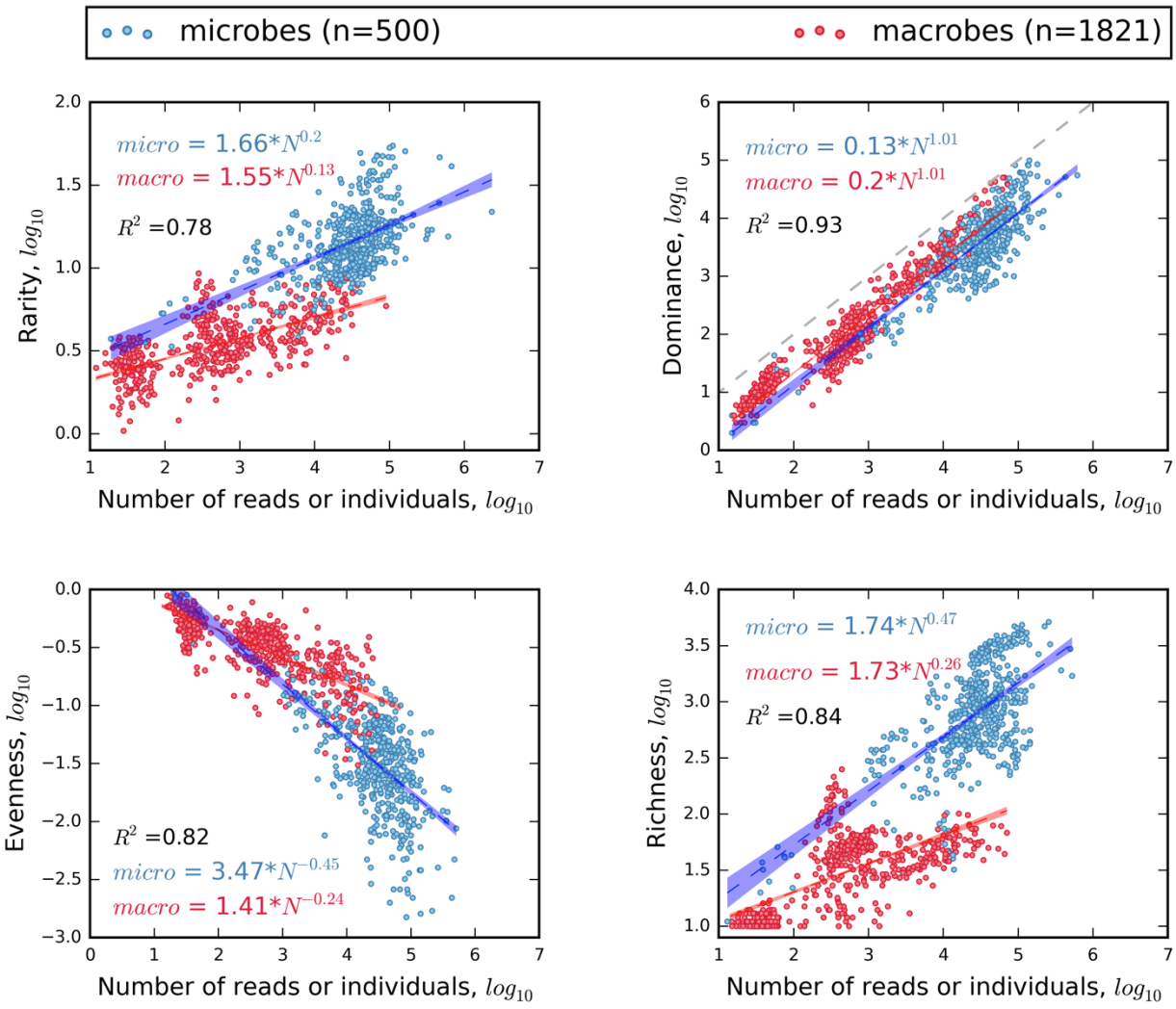


Fig S13. When using only EMP open reference data, all scaling relationships differ between macrobes and microbes, except for the dominance relationship, which remains nearly isometric. The plots of data in each subfigure represent a single random sample from microbe and macrobe data compilations. The model formulas represent average coefficient values from 10,000 random resamplings (with reassignment of the microbe/macrobe category).

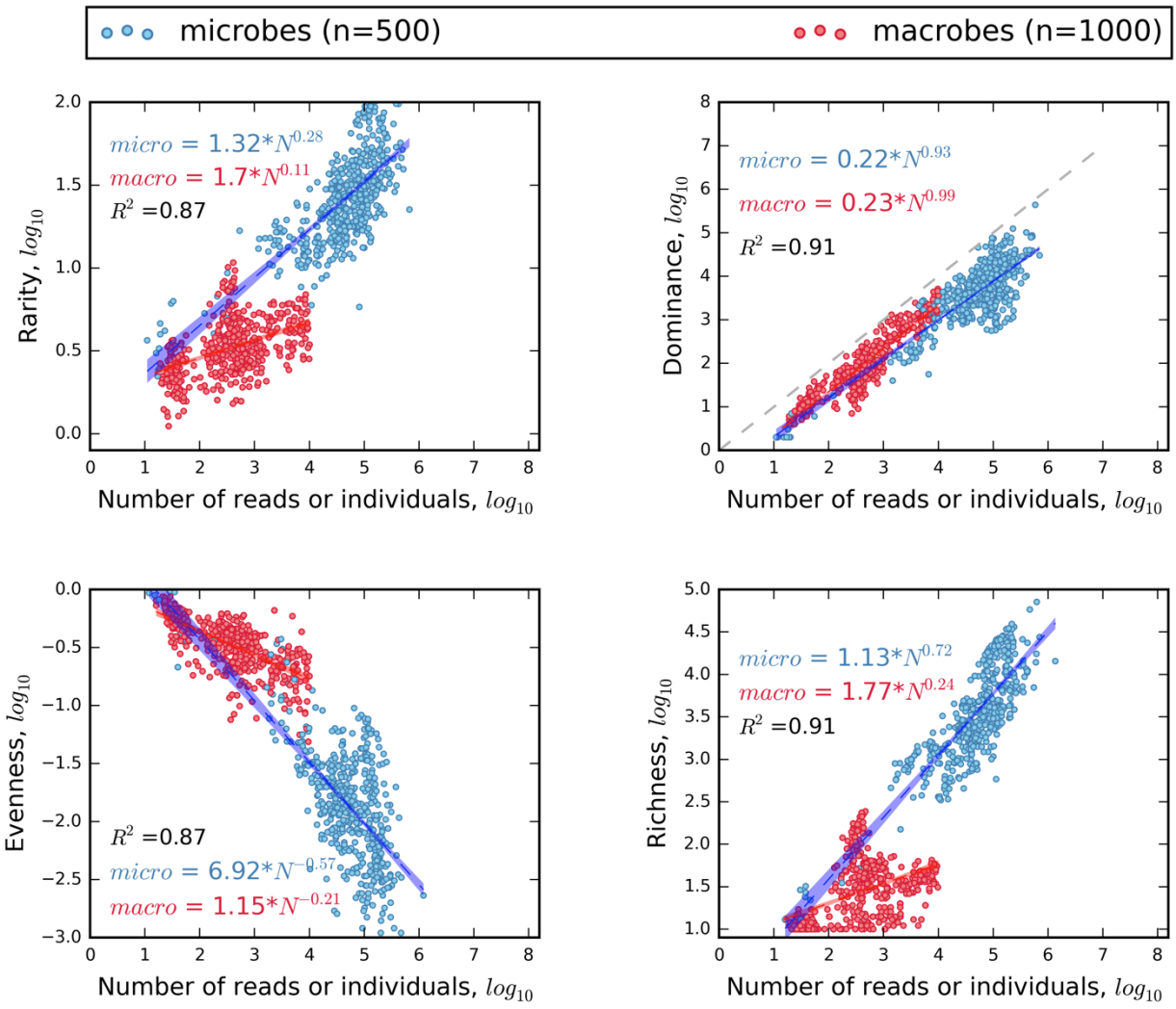


Figure S14. Flow diagram of how we used observed *N* to obtain predicted N_{max}, and then used those values to parameterize the lognormal model.

Obtaining bootstrapped predictions of *S* for a microbiome or microbial community where values of total abundance (*N*) have been reported. Below, N_{max} is the predicted abundance of the most abundant species.

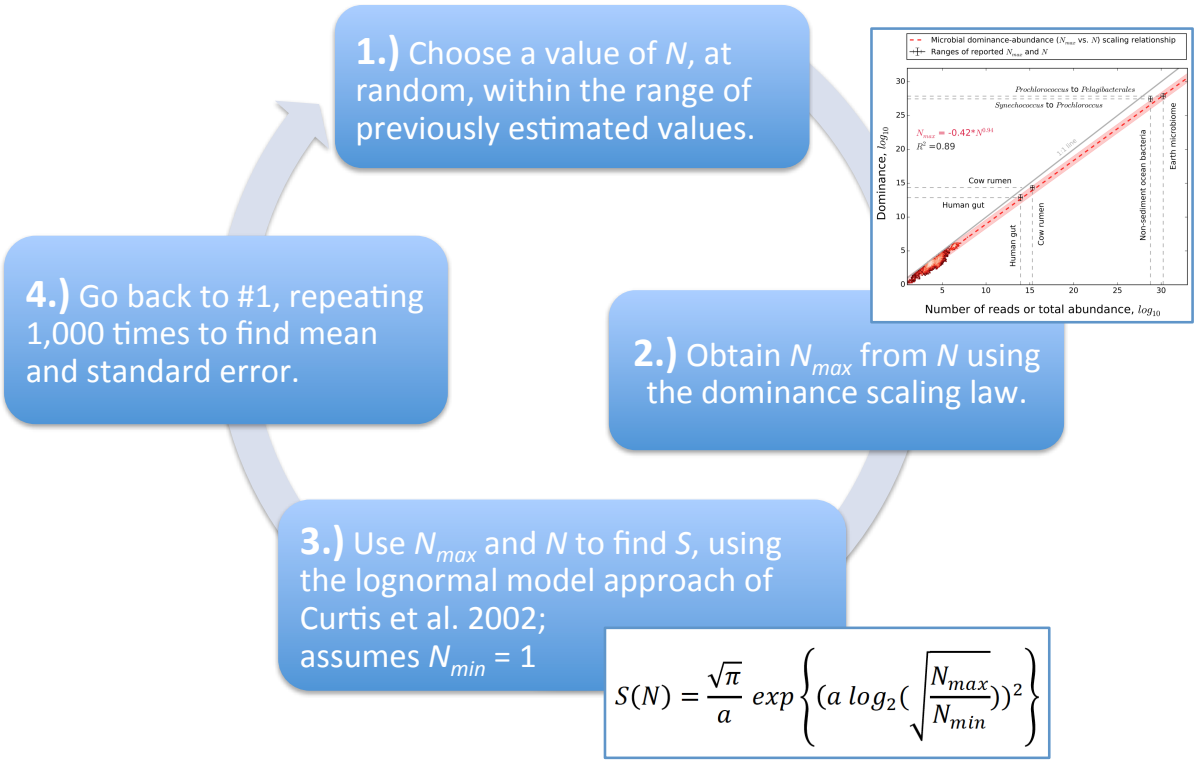


Table S1. Comparing fits of Power-law, Semi-log, Exponential and linear models. The power-law model provides the best overall fit to the data. For rarity, the power law explains nearly 20% more variation the next best model (exponential) but has comparable AIC and BIC. For dominance, the power-law model explains nearly as much variation as the linear model, but has much smaller AIC and BIC values. For evenness, the power-law explains 7% less variation than the semi-log model but has much lower AIC and BIC values. For richness, the power law has both the higher r-squared and lowest AIC and BIC values.

Rarity	R-squared	AIC	BIC
power-law	**0.552**	-306.44	-285.19
Semi-log	0.351	10227.99	10249.241
exponential	0.366	**216.78**	**238.031**
linear	0.162	3642.02	3659.61
Dominance	R-squared	AIC	BIC
power-law	0.942	**775.75**	**797.01**
Semi-log	0.205	44722.91	44744.16
exponential	0.406	4269.39	4290.64
linear	**0.976**	39399.24	39420.50
Evenness	R-squared	AIC	BIC
power-law	0.636	**846.32**	**867.57**
Semi-log	**0.707**	-3106.81	-3085.56
exponential	0.404	1586.58	1607.83
linear	0.575	-2548.41	-2527.16
Richness	R-squared	AIC	BIC
power-law	**0.572**	**1884.81**	**1906.07**
Semi-log	0.242	24918.30	24939
exponential	0.305	2610.94	2632.19
linear	0.071	25222.86	25244.11

References

1. Magurran, A. E., McGill B. J., eds. (2011). *Biological diversity: frontiers in measurement and assessment*. Vol. 12. Oxford: Oxford University Press.