Figure S1. The relationship between the predicted rank-abundance and the observed rank-abundance across models and datasets. Top row: The Human Microbiome Project (HMP) as in Figure 1. Middle and bottom rows: Closed and open reference Earth Microbiome Project (EMP) datasets. SAD models are 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.

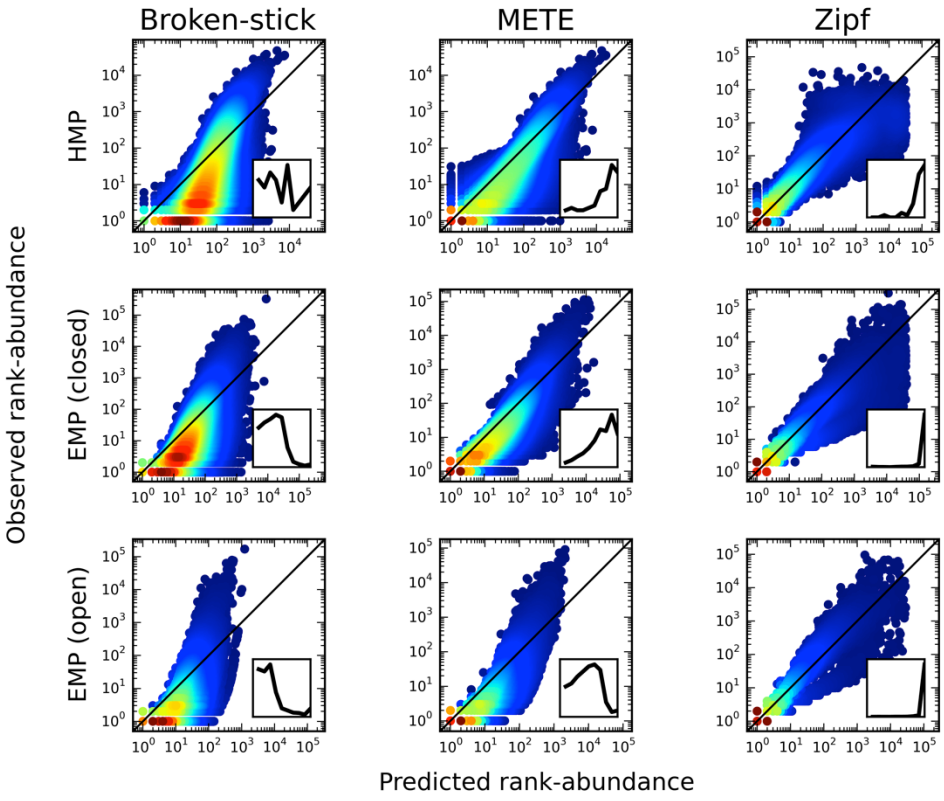

Figure S2. The relationship between the predicted rank-abundance and the observed rank-abundance across models and datasets downloaded from MG-RAST. Each row represents a different percent sequence similarity (95, 97, 99), often taken to represent species-level taxonomic units; 97% being the most common. SAD models are 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.

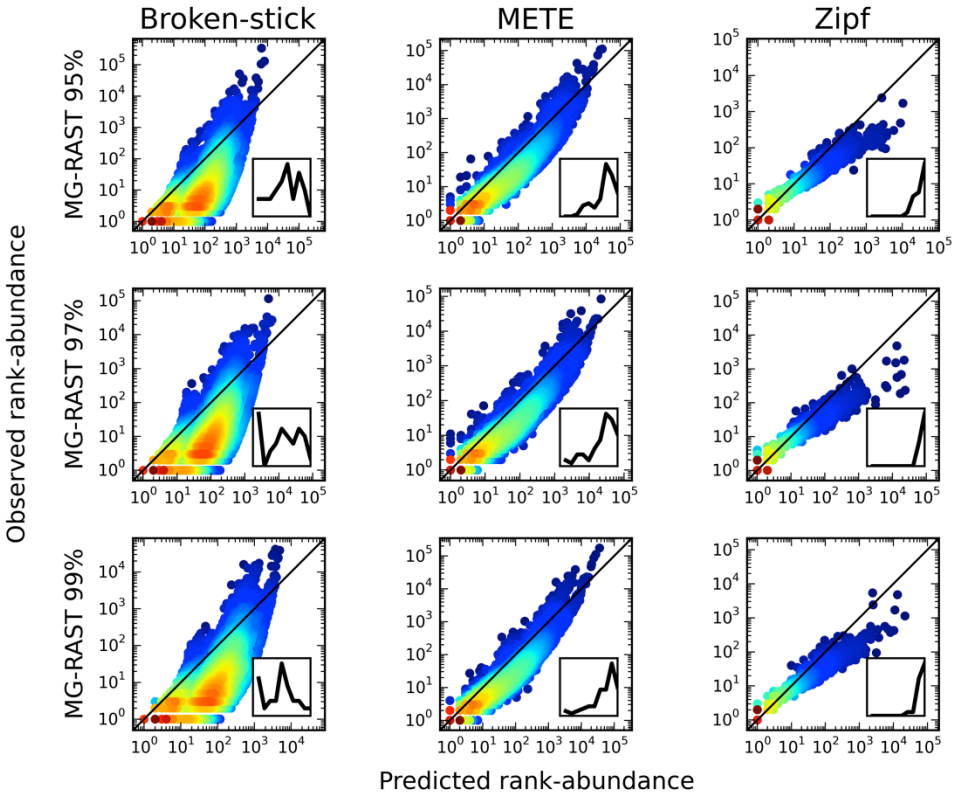

Figure S3. The relationship between the performance of a model (modified r-square based on the 1:1 line) for *closed* reference Earth Microbiome Project (EMP) data and the state variables of total numbers of individual reads (N_0), total number of taxa (S_0), and average abundance of reads among taxa (N_0/S_0).

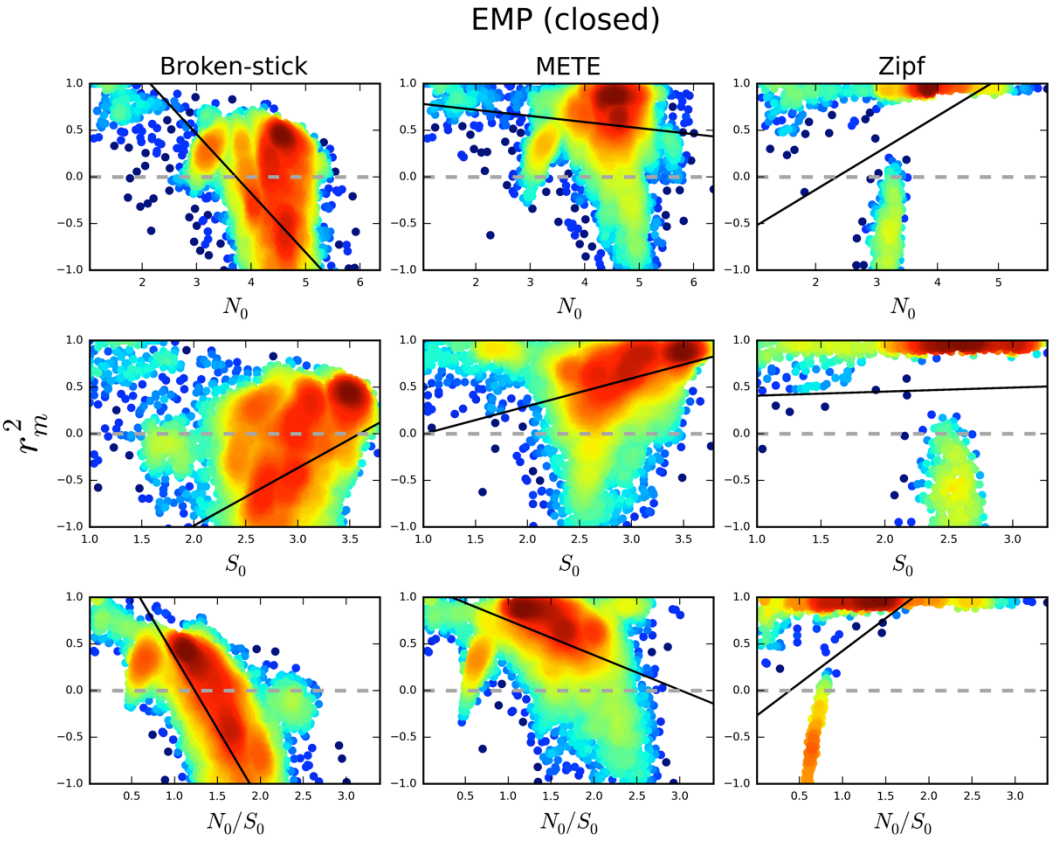

Figure S4. The relationship between the performance of a model (modified r-square based on the 1:1 line) for *open* reference Earth Microbiome Project (EMP) data and the state variables of total numbers of individual reads (N_0), total number of taxa (S_0), and average abundance of reads among taxa (N_0/S_0).

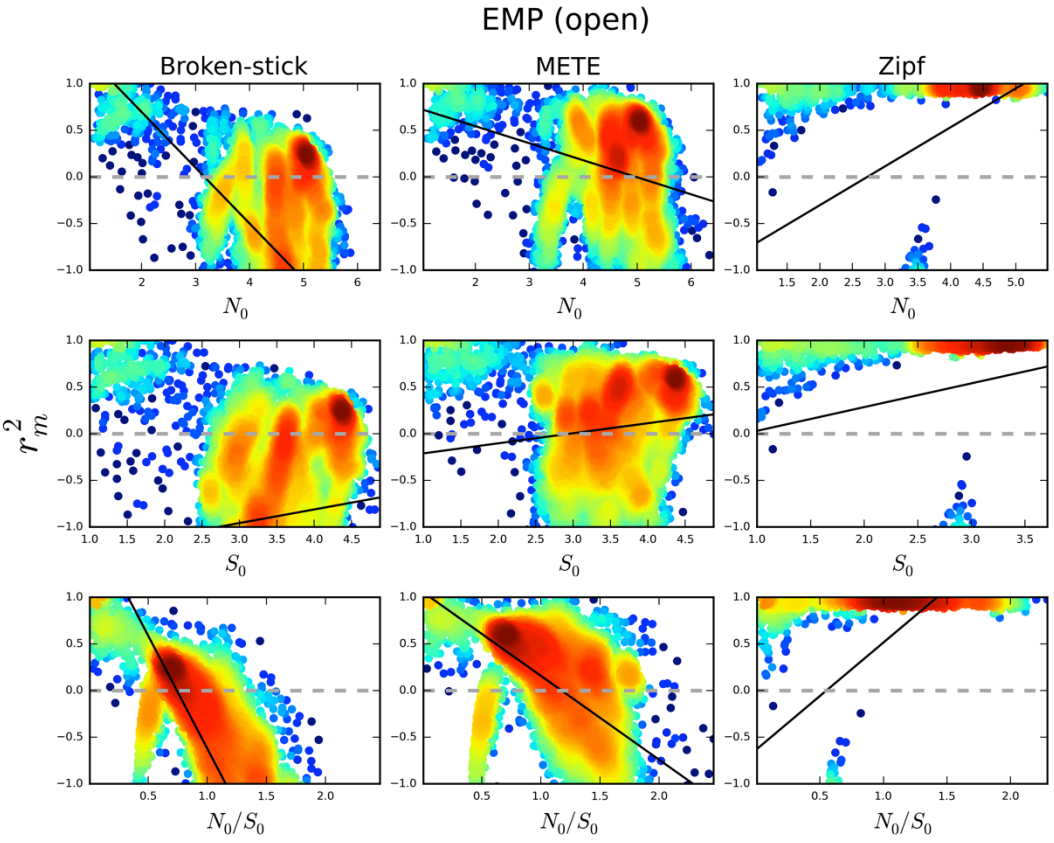

Figure S5. The relationship between the performance of a model (modified r-square based on the 1:1 line) for all MG-RAST datasets at the 97% sequence similarity cutoff and the state variables of total numbers of individual reads (N_0), total number of taxa (S_0), and average abundance of reads among taxa (N_0/S_0).

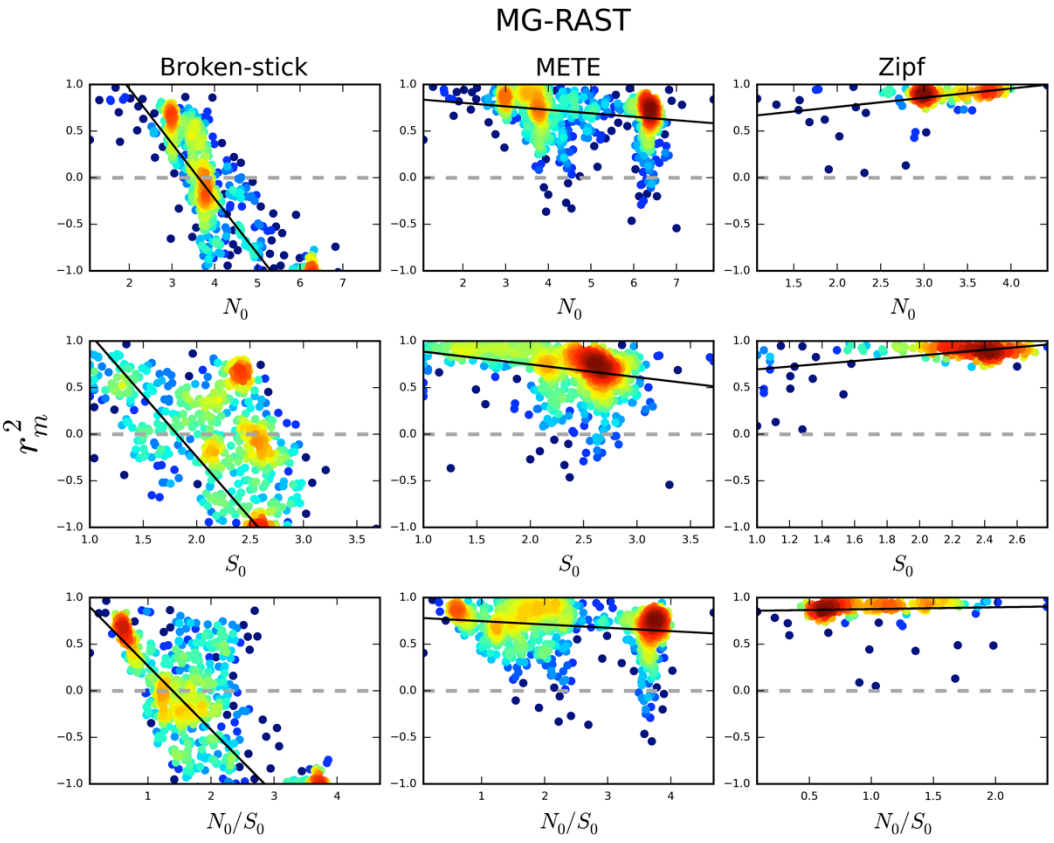

Figure S6. The relationship between the performance of a model (modified r-square based on the 1:1 line) for a subset of MG-RAST datasets (BOVINE, CHU, CATLIN, LAUB) at the 95% sequence similarity cutoff and the state variables of total numbers of individual reads (N_0), total number of taxa (S_0), and average abundance of reads among taxa (N_0/S_0).

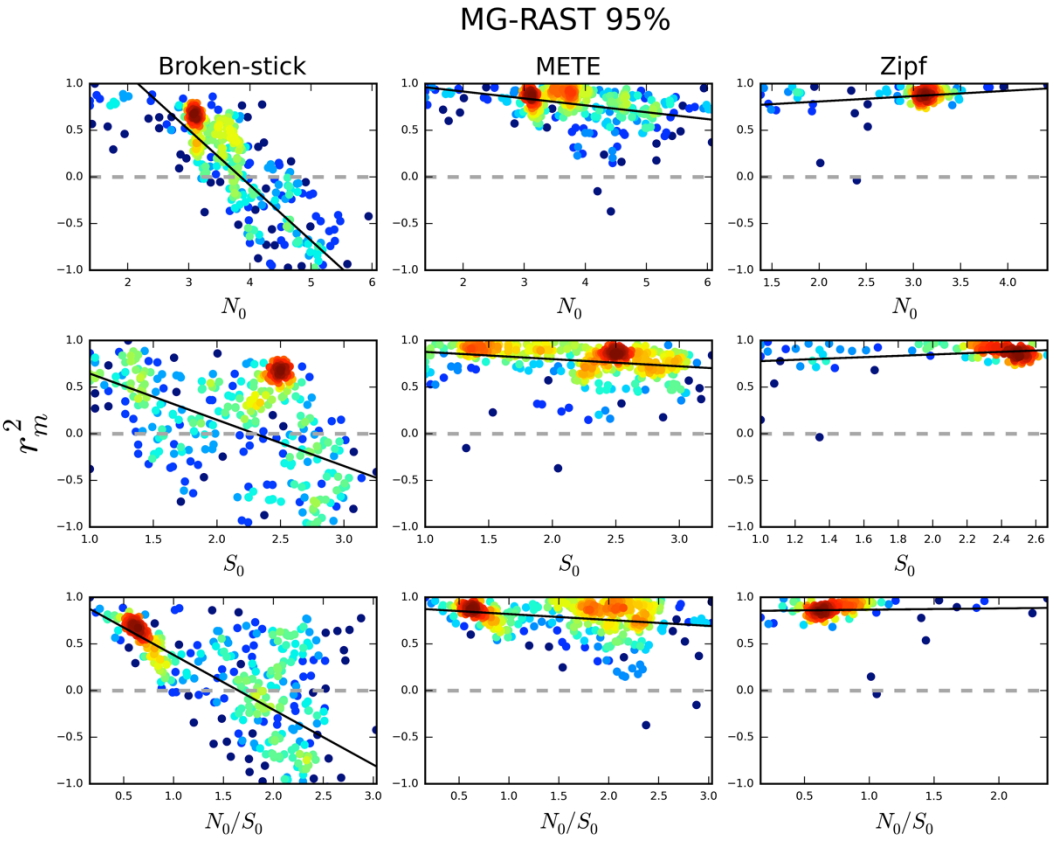

Figure S7. The relationship between the performance of a model (modified r-square based on the 1:1 line) for a subset of MG-RAST datasets (BOVINE, CHU, CATLIN, LAUB) at the 97% sequence similarity cutoff and the state variables of total numbers of individual reads (N_0), total number of taxa (S_0), and average abundance of reads among taxa (N_0/S_0).

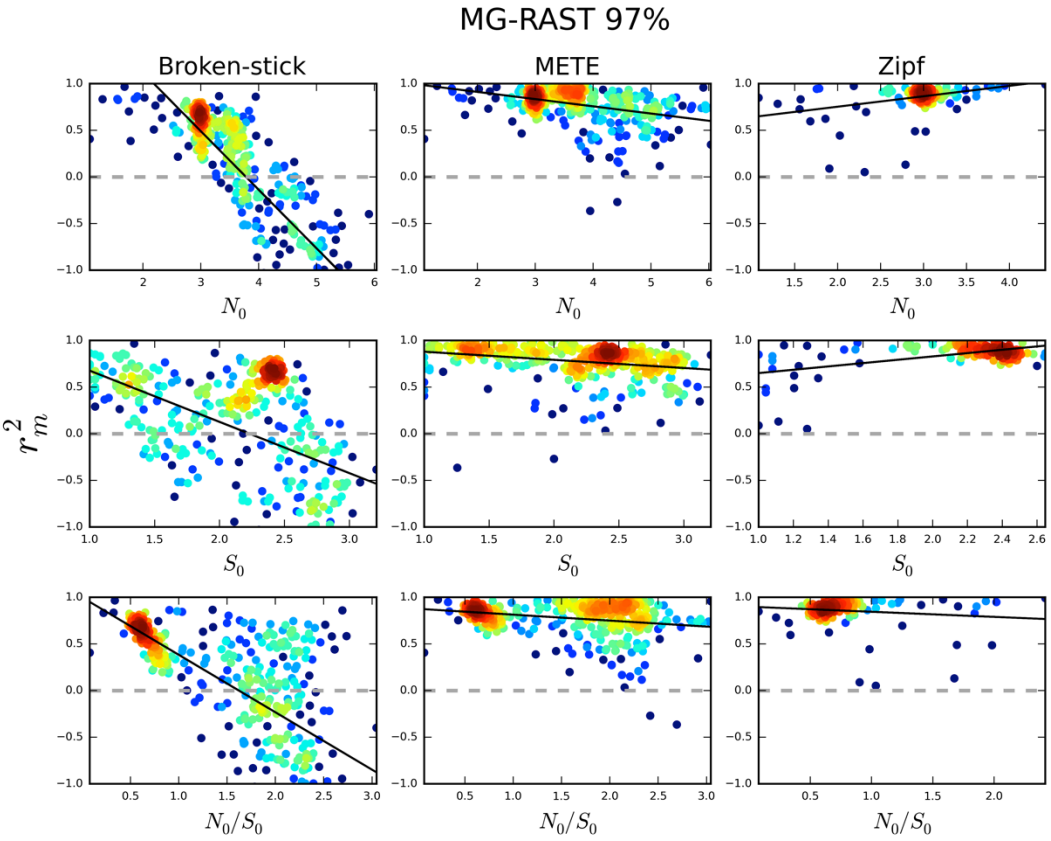

Figure S8. The relationship between the performance of a model (modified r-square based on the 1:1 line) for a subset of MG-RAST datasets (BOVINE, CHU, CATLIN, LAUB) at the 99% sequence similarity cutoff and the state variables of total numbers of individual reads (N_0), total number of taxa (S_0), and average abundance of reads among taxa (N_0/S_0).

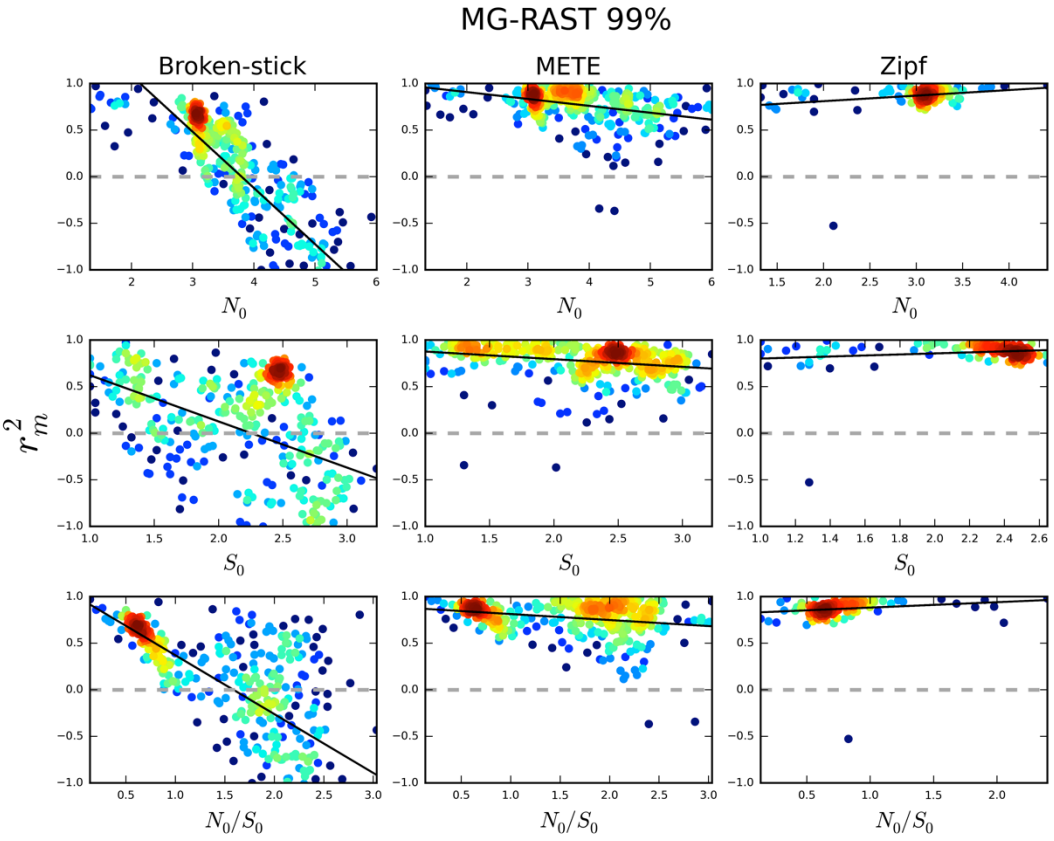

Figure S9. The relationship between the maximum likelihood estimate of the Zipf distribution (α) and the total number of individuals (N_0) (left column) and the abundance of the most abundant taxon (N_{max}) (center column) for the Human Microbiome Project (HMP), open and closed reference datasets for the Earth Microbiome Project (EMP), and all MG-RAST datasets where taxa are clustered at the 97% sequence similarity cutoff. Right column: kernel density curves reveal a strong mode of maximum likelihood values for the Zipf distribution (α) centered between -2 and -1.5.

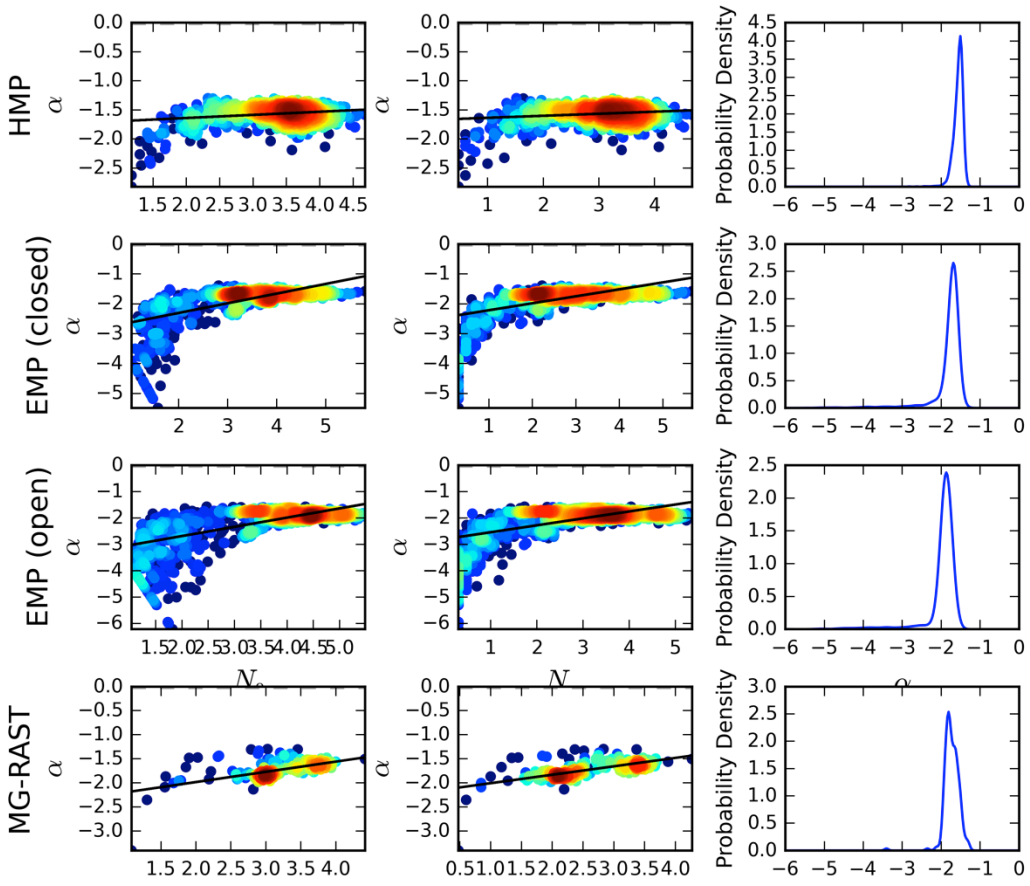

Figure S10. The relationship between the maximum likelihood estimate of the Zipf distribution (α) and the total number of individuals (N_0) (left column) and the abundance of the most abundant taxon (N_{max}) (center column) for a subset of MG-RAST datasets (BOVINE, CHU, CATLIN, LAUB) where taxa are clustered at the 97% sequence similarity cutoff. Right column: kernel density curves reveal a strong mode of maximum likelihood values for the Zipf distribution (α) centered between -2 and -1.5.