Rarefy your data

Running title: Rarefy your data

Patrick D. Schloss[†]

[†] To whom corresponsdence should be addressed:

5 pschloss@umich.edu

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

Research article

Abstract

Importance

Introduction

https:// edepot.wur.nl/ 547087. "Log- ratio PCA is designed to give results that are library size- independent. However, as we demonstrated mathematically and with examples based on simulated and real data, log- ratio PCA becomes library size- dependent, if there are many infrequent taxa (many zeroes) and library sizes differ largely. In this situation, the row centring used in log- ratio PCA brings an effect of r (the row mean of the log- transformed counts) in the clr- transformed matrix. Note that this effect is irrespective of whether or not these infrequent taxa are genuine or due to sequencing noise or allocation error. This library size dependence is unexpected in the sense that, after applying the clr, the transformed matrix is free of the effect of the row totals for strictly positive data ($y_{ij}>0$ for all i and j). We additionally demonstrate that library size variability causes a loss in power in dextecting an effect of x with log- ratio RDA. If there is additionally a correlation between treatment and the library size, the type 1 error for detecting the effect of x can be seriously inflated."

https:// www.nature.com/ articles/ s41522-020-00160-w. "An important characteristic of a feature table is that it is typically sparse, sometimes as many as ~90% are zero entries21, which creates a challenge for analyzing rare taxa. A quick and simple strategy to deal with excess zeros is to add a small positive constant (e.g. 1) called pseudo-count14,22 to each cell of the feature table. The addition of a pseudo-count becomes necessary when using methods of analysis that require log transformation of the observed counts. Even though adding a pseudo-count is simple and widely used, the choice of the pseudo-count is ad hoc. Studies have shown that differential abundance or clustering results could be sensitive to the choice of pseudo count23,24. Although different values of pseudo counts have been discussed in the literature23,24,25,26, to the best of our knowledge, there is no consensus on how to choose the optimal value. Other strategies involve modeling zero counts by some probability models21,27."

https://academic.oup.com/bioinformatics/article/34/16/2870/4956011   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6755255/

- Motivation

    - Problem of uneven sampling effort

- What is rarefaction? History, reason for rarefaction

    - Repeated down sampling of datasets to a common number of observations to calculate the average value to ascertain the expected value of a metric for the metric under study; typically richness

- **–** Control for unneven sampling effort

- **–** Methods vary in their sensitivity to uneven sampling

- **–** Compositional data analysis

45 • Reasons behind "rarefaction is inadmissable"

- **–** Weird simulation

• Alternative approaches and claims

- **–** sampling invariance

• Goal of this study

50 This analysis included 16S rRNA gene sequence data from from 12 studies that characterized the variation in bacterial communities from diverse environments (Table 1). The original studies generated the sequence data by pooling separate PCR products that were generated by amplifying the V4 region of the 16S rRNA gene from the bacterial DNA in multiple samples. Because pooling equimolar quantities of DNA is frought with difficulties, it was common to observe wide variation in the number of sequences in each sample (Table 55 1 and Figure S1).

Results

Without rarefaction, metrics of alpha diversity are sensitive to sampling effort. To test the sesitivity of various approaches of measuring alpha diversity to sampling effort, I generated null models for each dataset. Under a null model, each community from the same dataset would be expected to have the 60 same alpha diversity regardless of the sampling effort. I measured the richness of the communities in each dataset without any correction, using scaled ranked subsampling (SRS) normalized OTU counts, with estimates based on non-parametric and parametric approaches, and using rarefaction (e.g. Figure S2). For each dataset, all of the approaches, except for rarefaction, showed a strong correlation between richness and the number of sequences in the sample (Figure 1A). Next, I assessed diversity using the 65 Shannon diversity index and the inverse Simpson diversity index without any correction, using normalized OTU counts, and rarefaction; I also used a non-parametric estimator of Shannon diversity. The correlation between sampling depth and the diversity metric was not as strong as it was for richness and the inverse Simpson diversity values were less sensitive than the Shannon diversity values; however, the correlation to the rarefied diversity metrics were the lowest for all of the metrics and studies (Figure 1A). The rarefied 70 alpha-diversity metrics consistently demonstrated a lack of sensitivity to sampling depth.

4

Without rarefaction, metrics of beta diversity are sensitive to sampling effort. To test the sesitivity of various approaches of measuring beta diversity to sampling effort, I used the same null models used for studying the sensitivity of alpha diversity. Under a null model, the ecological distance between any pair of samples would be the same regardless of the difference in the number of sequences observed in each sam-

75 ple (e.g., Figure S3). First, I calculated the Jaccard distance coefficient between all pairs of communities within a dataset. The Jaccard distance coefficient is the fraction of OTUs that are unique to either community and does not account for the abundance of the OTUs. Jaccard distances were calcualted using the uncorrected OTU counts, with rarefaction, relative abundances, and following normalization using cumulative sum scaling (CSS) and SRS. Only the rarefied distances showed a lack of sensitivity to sampling effort

80 (Figure 1B). Second, I analyzed the sensitivity of the Bray-Curtis distance coefficient, which is a popular metric that incorporates the abundance of each OTU. Similar to what I observed with the Jaccard coefficient, only the rarefied data showed a lack of sensitivity to sampling effort (Figure 1B). Third, I calcualted the Euclidean distance on raw OTU counts where the central log-ratio (CLR) was calculated (i.e., Aitchison distances) by ignoring OTUs in samples with zero counts (Robust CLR), adding a pseudocount of one to all

85 OTU counts prior to calculating the CLR (One CLR), adding a pseudocount of one divided by the total number of sequences obtained for the community (Nudge CLR), and imputing the value of zero counts (Zero CLR). The Aitchison distances were all strongly sensitive to sampling effort (Figure 1B). Finally, I used the variance-stabilization transformation (VST) from DeSeq2 prior to calculating Euclidean distances. Again, there was a strong sensitivity to sampling effort (Figure 1B). Although Euclidean distances are not typically

90 used on raw or rarefied count data in ecology, rarefied Euclidean distances were not sensitive to sampling effort. Across each of the beta diversity metrics and approaches used to account for uneven sampling effort and sparsity, rarefaction was the least sensitive approach to differences in sampling effort.

Rarefaction limits the detection of false positives when sampling effort and treatment group are confounded. Next, I investigated the impact of the various strategies and metrics on falsely detecting

95 a significant difference using the the same communities generated from the null model in the analysis of alpha and beta diversity metrics. To test for differences in alpha and beta diversity I used the non-parametric Wilcoxon test and non-parametric permutation-based multivariate analysis of variance (PER-MANOVA). First, I employed an unbiased null treatment model to measure the false detection rate, which should not have meaningfully differed from 5%. Indeed, for each dataset and alpha and beta diversity metric

100 and strategy for accounting for uneven sampling, approximately 5% of the tests yielded a significant result (Figure 2). Second, I employed a biased null treatment model where the treatment group was completely confounded with the number of sequences in each sample. Under this model, only the rarefied data con-

5

sistently resulted in a 5% false positive rate for alpha and beta diversity metrics (Figure 2). These results aligned with the observed sensitivity of alpha and beta diversity metrics to sampling effort and underscore the value of rarefaction.

Rarefaction preserves the statistical power to detect differences between treatment groups. To assess the impact of different approaches to control for uneven sampling effort I performed two additional simulations. In the first simulation, I implemented a skewed abundance distribution model to create two treatment groups for each datasets that were each populated with half of the samples each with the same number of sequences as the samples had in the observed data. The power to detect differences in richness between the two simulated treatment groups by all approaches was low (Figure 4A). This was likely because the approach for generating the perturbed community did not necessarily change the number of OTUs in each treatment group. Regardless, the simulations testing differencse in richness using the Rice and Stream datasets had the greatest power when the richness data were rarefied. To explore this further, a richness-adjusted community model was created by removing 3% of the OTUs from a null model model. As suggested by the first simulation, the rarefied richness data had a higher statistical power than the other approaches when measuring richness (Figure 5). The simulations testing the power to detect differences in Shannon diversity also showed that rarefied data performed other methods (Figure 4A). When testing for differences in the Inverse Simpson diversity index the the difference between rarefaction and the other methods was negligible (Figure 4A). For tests of beta diversity I found that rarefaction was the most reliable approach to maintain statistical power to detect differences between two communities. Among the tests using the Jaccard and Bray-Curtis metrics, raw count data and CSS normalized data had little power relative to rarefied, relative abundance, and SRS normalized data. The differences in power between rarefied, relative abundance, and SRS normalized data was small, but if there were differences, the power obtained using rarefied data was greater than the other methods. Among the tests using Euclidean distances, using raw counts and CLR and DeSeq2 transformed data had little power relative to the distances calcualted using rarefied and relative abundance data. This power-based analysis of the simulated communities using different methods of handling uneven sample sizes demonstrated the value of rarefaction for preserving the statistical power to detect differences between treatment groups for measures of alpha and beta diversity.

Increased rarefaction depth reduces intra-sample variation in alpha and beta diversity. Once concern with rarefying communities is the perceived loss of sequencing information when more a large fraction of data appears to be removed when the community with the greatest sequencing depth is rarefied to the size of the community with the least (e.g., 99% with the Bioethanol dataset). To assess the sensitivity of alpha and beta diversity metrics to rarefaction depth, I again used the dataset generated using the null models,

but rarefied each community to varying sampling depths (Figure 6). The richness values increased with sampling effort as rare OTUs would continue be detected. In contrast, the Shannon diversity and Bray-Curtis values plateaued with increased sampling effort. This result was expected since increased sampling would lead to increased precision in the measured abundance of OTUs. Next, I measured the coefficient of variation (i.e., the mean divided by the standard deviation) between samples for richness, Shannon diversity, and Bray-Curtis distances. Although the richness values appeared to increased unbounded with smapling effort, the coefficient of variation for each dataset was relatively stable. In general, the coefficient of variation increased slightly with sampling depth only to decline once smaller samples were removed from the analysis at higher sampling depths. Interestingly, the coefficient of variation between Shannon diveristy values decreased towards zero with increased sampling effort and the coefficient of variation between Bray-Curtis distances tended to increased. Regardless, the coefficients of variation were relatively small.

Most studies have a high level of sequencing coverage. To explore the concern over loss of sequencing depth further, I calculated the Good's coverage for the observed data. The median coverage for each dataset ranged between 89.4 and 99.8% for the Seagrass and Human datasets, respectively (Figure 7). When I rarefied each dataset to the size of the smallest community in the dataset, with the exception of the Seagrass, Rice, and Stream datasets, the median coverage for the rarefied communities was still greater than 90%. These results suggest that most studies had a level of sequencing coverage that aligned with the diversity of the communities. Next, I used the null model for each dataset to ask how much sequencing effort was required to obtain higher levels of coverage. To obtain 95 and 99% coverage, an average of 2.70 and 101.20-fold more sequence data was estimated to be required than was required to obtain 90% coverage, respectively (Figure 7). As suggested by the simulated coverages curve in Figure 7, these estimates are conservative. Regardless, the sequencing effort required to acheive higher sequencing depth would likely limit the number of samples that could be sequenced when controlling for costs. Although it may be disconcerting to rarefy to a sequencing depth that is considerably lower than that obtained for the best sequenced community in a dataset, sequencing coverage for many environments is probably adequate even at the lower sequencing depth. Of course, the results above have demonstrated that rarefaction is necessary to avoid problems with making inferences.

Discussion

- Rarefy your data
- Problems with recommended methods...

  – Many recommended methods are borrowed from gened expression analysis

  – Meaning of zeroes in data - structural vs. below limit of detection

- Factors that determine what number of sequences to rarefy to

- Need better methods of pooling libraries that result in more even distribution of sequences across samples

- Rarefy your data

Materials and Methods

Choice of datasets. The specific studies used in this study were selected because their data was publicly accessible through the Sequence Read Archive, the original investigators sequenced the V4 region of the 16S rRNA gene using paired 250 nt reads, and my previous familiarity with the data. The use of paired 250 nt reads to sequence the V4 region resulted in a near complete two-fold overap of the V4 region resulting in high quality contigs with a low sequencing error rate (1). These data were processed through a standardized sequence curation pipeline to generate operational taxonomic units (OTUs) using the mothur software package (1, 2). OTUs were identified using the OptiClust algorithm to cluster sequences together that were not more than 3% different from each other (3).

Null community model. Null community models were generated such that within a dataset the number of sequences per sample and the number of sequences per OTU across all samples within the dataset were the same as was observed in original. This model effectively generated statistical samples of a single community so that there should not have been a statistical difference between the samples. This model implemented by randomly assigning each sequence in the dataset to an OTU and sample while keeping constant the number of sequences per sample and the total number of sequences in each OTU. This is a similar approach to that of the IS algorithm described by Ulrich and Gotelli (4). Because the construction of the null models was a stochastic process, 100 replicates were geneated for each dataset.

Null treatment model. I created an unbiased and biased treatment model. In the unbiased model, samples were randomly assigned to one of two treatment groups. In the biased treatment model, samples that had more than the median number of seqeunces for a dataset were assigned to one treatment group and the rest were assigned to a second treatment group. Comparison of any diversity metric across the two treatment groups should have only yielded a significant result in 5% of the simulations when testing under a Type I error (i.e., α) of 0.05.

Skewed abundance community model. In the skewed abundance community model, communities were

randomly assigned to one of two simulated treatment groups. Communities in the first treatment group were generated by calculating the relative abundance of each OTU across all samples and using those values as the probability of sampling each OTU. This probability distribution was sampled until each sample had the same number of sequences that it did in the observed data. Samples in the second treatment group were generated by adjusting the relative abundances of the OTUs in the firs treatment group by increasing the relative abundance of 10% of the OTUs by 5%. These values were determined after empirically searching for conditions that resulted in a large fraction of the randomizations yieleding a significant result across most of the studies. Sequences were sampled from the skewed community community until each sample had the same number of sequences that it did in the observed data. Under the skewed abundance community model each sample represented a statistical sampling of two communities such that there should not have been a statistically significant difference within a treatment group, but there was between the treatment groups. Because the construction of the skewed abundance community model was a stochastic process, 100 replicates were geneated for each dataset.

Richness-adjusted community model. In the richness-adjusted community model, communities were randomly assigned to one of two simulated treatment groups. Communities in the first treatment group were generated by calculating the relative abundance of each OTU across all samples and using those values as the probability of sampling each OTU. This probability distribution was sampled until each sample had the same number of sequences that it did in the observed data. Samples in the second treatment group were generated by removing 3% of the OTUs from the dataset and recalculating the relative abundance of the remaining OTUs. Sequences were sampled from the richness-adjusted community distrubtion until each sample had the same number of sequences that it did in the observed data. Under the richness-adjusted community model each sample represented a statistical sampling of two communities such that there should not have been a statistically significant difference within a treatment group, but there was between the treatment groups. Because the construction of the richness-adjusted community model was a stochastic process, 100 replicates were geneated for each dataset.

Test of statistical significance. Statistical comparisons of alpha diversity metrics across the simulated treatment groups were performed using the non-parametric two-sample Wilcoxon test as implemented in R. This test was selected because the alpha-diversity metrics tended to not be normally distributed and each dataset required a different transformation to normalize the data. Comparisons of beta diversity metrics were performed using the adonis2 function from the vegan (v.2.6.2) R package (5). The adonis2 function implements a non-parametric multivariate analysis of variance using distances matrices as described by McArdle and Anderson (6). Throughout this study I used 0.05 as the threshold for asssessing statistical

9

significance.

Power analysis. The parameters used to design the skewed abundance and richness-adjusted community models were set to impose a known effect size when using rarefied community data. The statistical power to detect these differences was determined by calculating the p-value for each of 100 replicate simulated set of samples from each dataset using the various alpha and beta diversity metrics. The percentage of tests that yielded a significant P-value was considered the statistical power (i.e., 1 minus the Type II error) to detect the difference.

Alpha diversity calculations. Various strategies for handling uneven sampling effort were evaluated to identify the best approach for calculating community richness and Shannon and inverse Simpson diversity indeices. Raw OTU counts were used as input to calculate sample richness and Shannon and inverse Simpson diversity using mothur (2, 7). Shannon diversity was calculated as

$$H_{shannon} = -\sum_{i=1}^{S_{obs}} \frac{n_i}{N} ln \frac{n_i}{N}$$

The Simpson diversity was calculated as

$$D_{simpson} = \frac{\sum_{i=1}^{S_{obs}} n_i (n_i - 1)}{N (N - 1)}$$

The inverse Simpson diversity was calculated as $1/D_{simpson}$. In both formulae, n_i is the number of sequences in OTU i and N is the number of sequences in the sample. Rarefaction of richness, Shannon diversity and Inverse Simpson diversity values were carried out in mothur. Briefly, mothur calculates each value on a random draw of the same number of sequences from each sample and obtains a mean value based on 1,000 random draws. Scaled ranked subsampling (SRS) was used to normalize OTU counts to the size of the smallest sample in each dataset using the SRS R package (v.0.2.3)(8). Normalized OTU counts were used to calculate sample richness and Shannon and inverse Simpson diversity values using mothur. Data normalized by cumulative sum scaling (CSS) were not reported for alpha-diversity values since the relative abundances of the features do not change with the normalization procedure (9). The non-parametric bias-corrected Chao1 and ACE richness estimators (10) and a non-parametric estimator of the Shannon diversity (11) were calculated using raw OTU counts with mothur. Parametric estimates of sample richness were calculated using the breakaway (BA) R package (v.4.7.9)(12). The current analysis reports both the results from running default model selection procedure and the Poisson model. The default model selection returned either the Kemp, Negative Binomial, or Poisson models. Relative abundance data

were not used to calculate alpha diversity metrics since the richness and evenness does not change from the raw data when dividing each sample by the total number of sequences in the sample.

255 **Beta diversity calculations.** Similar to the alpha diversity calculations, multiple approaches were used to control for uneven sampling effort and calculate beta diversity. Raw and OTU counts were used for input to calculate the Jaccard, Bray-Curtis, and Euclidean dissimilarity indices using the vegdist function from the vegan R package (v.2.6.2)(5). The Jaccard index was calculated as

$$D_{Jaccard} = 1 - \frac{S_{AB}}{S_A + S_B - S_{AB}}$$

where S_A and S_B were the number of OTUs in samples A and B and S_{AB} was the number of OTUs shared
260 between the two samples. The Bray-Curtis index was calculated as

$$D_{Bray-Curtis} = 1 - \frac{\sum_{i=1}^{S_T} |n_{A,i} - n_{B,i}|}{N_A + N_B}$$

where $n_{A,i}$ and $n_{B,i}$ are the number of sequences observed in OTU i from samples A and B, respectively. N_A and N_B are the total number of sequences in samples A and B, respectively. S_T is the total number of OTUs observed between the two samples. The Euclidean distance was calculated as

$$D_{Euclidean} = 1 - \sqrt{\sum_{i=1}^{S_T} (n_{A,i} - n_{B,i})^2}$$

These metrics were calculated using the relative abundance of each OTU using the vegdist function from
265 vegan. The relative abundance was calculated as the number of sequences in the OTU (e.g., $n_{A,i}$) divided by the total number of sequences in the sample (e.g., N_A).

Rarefied beta-diversity values were calculated using the avgdist function in vegan. Briefly, vegan's avgdist function calculates each pairwise dissimilarity index after obtaining a random draw of the same number of sequences from each sample. After obtaining 100 random draws it returns the mean value.

270 Three approaches were taken to normalize the number of sequences across samples within a dataset. Scaled ranked subsampling (SRS) and cumulative sum scaling (CSS) were used to normalize raw OTU counts using the SRS (v.0.2.3) and metagenomeSeq (v.1.36.0) R packages, respectively (8, 9). The normalized counts were then used to calculate Jaccard and Bray-Curtis dissimilarity indices. Finally, the variance-stabilization transformation (VST) as implemented in the DESeq2 (v.1.34.0) R package was used

11

to normalize the data as described by McMurdie and Holmes (13, 14). Because the VST approach generates negative values, which are incompatible with calculating Jaccard and Bray-Curtis dissimilarity values, Euclidean distances were calcualted instead.

Raw OTU counts were used to calculate centered log ratio values for each OTU, which were then used to calcualte Euclidean distances; such distances are commonly refered to as Aitchison distances. Centered log-ratio (CLR) abundances are calculated as:

$$clr\left(n_j\right) = \left[\ln\frac{x_{ij}}{g(x_j)}, ..., \ln\frac{x_{S_T j}}{g(x_j)}\right]$$

where x_{ij} was the number of sequences observed for OTU i in sample j and $g()$ was the geometric mean x_j was the counts of the S_T OTUs in sample j. Because the geometric mean is zero if any OTU is absent from a sample, the CLR is undefined when there are unobserved OTUs in a sample. To overcome this problem, I attempted a four approaches. The first, Zero CLR, imputed the value of the zeroes based on the observed data using the zCompositions (v.1.4.0.1) R package (15). The second, One CLR, added a pseudocount of 1 to the abundance of all OTUs (9, 16). The third, Nudge CLR, added a pseudocount of 1 divided by the total number of sequences in a sample to each OTU in the sample (16, 17). The final approach, Robust CLR, removed unobserved OTUs prior to calcualting the CLR (18).

Analysis of sequencing coverage. To assess the level of sequencing coverage I calcualted Good's coverage (C_{Good}) using mothur:

$$C_{Good} = 100\% \times \left(1 - \frac{n_1}{N_T}\right)$$

where n_1 is the number of OTUs with only one sequence in the sample and N_T is the total number of sequences in the sample. Good's coverage was calculated using (i) the observed OTU counts for each sample and dataset, (ii) following rarefaction (1,000 iterations) of the observed OTU counts to the size of the smallest sample in each dataset, and (iii) after rarefying (1,000 iterations) the null community distribution.

Reproducible data analysis. A complete reproducible workflow written in Snakemake (v.7.15.2) and conda (v.4.12.0) computational environment can be obtained from the GitHub hosted git repository for this project (https://github.com/SchlossLab/Schloss_Rarefaction_XXXXX_2022). This paper was written in R markdown (v.2.16) with the kableExtra (v.1.3.4) package. The mothur (v.1.47.0) and R (4.1.3) software packages were used for all analyses with extensive use of functions in the tidyverse metapackage (v.1.3.1). A preliminary version of this analysis was presented as the Rarefaction video series on the Riffomonas Project

YouTube channel (https://www.youtube.com/playlist?list=PLmNrK_nkqBpJuhS93PYC-Xr5oqur7IlWf).

305

References

1. **Kozich JJ**, **Westcott SL**, **Baxter NT**, **Highlander SK**, **Schloss PD**. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and environmental microbiology **79**:5112–5120.

2. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Horn DJV**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**:7537–7541. doi:10.1128/aem.01541-09.

3. **Westcott SL**, **Schloss PD**. 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere **2**. doi:10.1128/mspheredirect.00073-17.

4. **Ulrich W**, **Gotelli NJ**. 2010. Null model analysis of species associations using abundance data. Ecology **91**:3384–3397. doi:10.1890/09-2157.1.

5. **Oksanen J**, **Simpson GL**, **Blanchet FG**, **Kindt R**, **Legendre P**, **Minchin PR**, **O'Hara RB**, **Solymos P**, **Stevens MHH**, **Szoecs E**, **Wagner H**, **Barbour M**, **Bedward M**, **Bolker B**, **Borcard D**, **Carvalho G**, **Chirico M**, **De Caceres M**, **Durand S**, **Evangelista HBA**, **FitzJohn R**, **Friendly M**, **Furneaux B**, **Hannigan G**, **Hill MO**, **Lahti L**, **McGlinn D**, **Ouellette M-H**, **Ribeiro Cunha E**, **Smith T**, **Stier A**, **Ter Braak CJF**, **Weedon J**. 2022. Vegan: Community ecology package.

6. **McArdle BH**, **Anderson MJ**. 2001. FITTING MULTIVARIATE MODELS TO COMMUNITY DATA: A COMMENT ON DISTANCE-BASED REDUNDANCY ANALYSIS. Ecology **82**:290–297. doi:10.1890/0012-9658(2001)082[0290:fmmtcd]2.0.co;2.

7. **Magurran AE**. 2004. Measuring biological diversity. Wiley.

8. **Beule L**, **Karlovsky P**. 2020. Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): Application to microbial communities. PeerJ **8**:e9593.

9. **Paulson JN**, **Stine OC**, **Bravo HC**, **Pop M**. 2013. Differential abundance analysis for microbial marker-gene surveys. Nature Methods **10**:1200–1202. doi:10.1038/nmeth.2658.

10. **Chao A**, **Chiu C-H**. 2016. Species richness: Estimation and comparison. Wiley StatsRef: statistics reference online **1**:26.

11. **Chao A**, **Shen T-J**. 2003. Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. Environmental and ecological statistics **10**:429–443.

12. **Willis A**, **Bunge J**. 2015. Estimating diversity via frequency ratios. Biometrics **71**:1042–1049.

13. **Love MI**, **Huber W**, **Anders S**. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**. doi:10.1186/s13059-014-0550-8.

14. **McMurdie PJ**, **Holmes S**. 2014. Waste not, want not: Why rarefying microbiome data is inadmissible. PLoS Computational Biology **10**:e1003531. doi:10.1371/journal.pcbi.1003531.

15. **Palarea-Albaladejo J**, **Martin-Fernandez J**. 2015. zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems **143**:85–96. doi:10.1016/j.chemolab.2015.02.019.

16. **Lin H**, **Peddada SD**. 2020. Analysis of microbial compositions: A review of normalization and differential abundance analysis. npj Biofilms and Microbiomes **6**. doi:10.1038/s41522-020-00160-w.

17. **Costea PI**, **Zeller G**, **Sunagawa S**, **Bork P**. 2014. A fair comparison. Nature Methods **11**:359–359. doi:10.1038/nmeth.2897.

18. **Martino C**, **Morton JT**, **Marotz CA**, **Thompson LR**, **Tripathi A**, **Knight R**, **Zengler K**. 2019. A novel sparse compositional technique reveals microbial perturbations. mSystems **4**. doi:10.1128/msystems.00016-19.

19. **Li Q**, **Heist EP**, **Moe LA**. 2015. Bacterial community structure and dynamics during corn-based bioethanol fermentation. Microbial Ecology **71**:409–421. doi:10.1007/s00248-015-0673-9.

20. **Baxter NT**, **Ruffin MT**, **Rogers MAM**, **Schloss PD**. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine **8**. doi:10.1186/s13073-016-0290-3.

21. **Beall BFN**, **Twiss MR**, **Smith DE**, **Oyserman BO**, **Rozmarynowycz MJ**, **Binding CE**, **Bourbon-niere RA**, **Bullerjahn GS**, **Palmer ME**, **Reavie ED**, **Waters LMK**, **Woityra LWC**, **McKay RML**. 2015. Ice cover extent drives phytoplankton and bacterial community structure in a large north-temperate lake: Implications for a warming climate. Environmental Microbiology **18**:1704–1719. doi:10.1111/1462-2920.12819.

22. **Henson MW**, **Pitre DM**, **Weckhorst JL**, **Lanclos VC**, **Webber AT**, **Thrash JC**. 2016. Artificial sea-water media facilitate cultivating members of the microbial majority from the gulf of mexico. mSphere **1**. doi:10.1128/msphere.00028-16.

23. **Baxter NT**, **Wan JJ**, **Schubert AM**, **Jenior ML**, **Myers P**, **Schloss PD**. 2014. Intra- and interindi-vidual variations mask interspecies variation in the microbiota of sympatric peromyscus populations. Applied and Environmental Microbiology **81**:396–404. doi:10.1128/aem.02303-14.

24. **Levy-Booth DJ**, **Giesbrecht IJW**, **Kellogg CTE**, **Heger TJ**, **D'Amore DV**, **Keeling PJ**, **Hallam SJ**, **Mohn WW**. 2018. Seasonal and ecohydrological regulation of active microbial populations in-volved in DOC, CO2, and CH4 fluxes in temperate rainforest soil. The ISME Journal **13**:950–963. doi:10.1038/s41396-018-0334-3.

25. **Edwards J**, **Johnson C**, **Santos-Medellín C**, **Lurie E**, **Podishetty NK**, **Bhatnagar S**, **Eisen JA**, **Sundaresan V**. 2015. Structure, variation, and assembly of the root-associated microbiomes of rice. Proceedings of the National Academy of Sciences **112**:E911–E920. doi:10.1073/pnas.1414592112.

26. **Ettinger CL**, **Williams SL**, **Abbott JM**, **Stachowicz JJ**, **Eisen JA**. 2017. Microbiome succession during ammonification in eelgrass bed sediments. PeerJ **5**:e3674. doi:10.7717/peerj.3674.

27. **Graw MF**, **DAngelo G**, **Borchers M**, **Thurber AR**, **Johnson JE**, **Zhang C**, **Liu H**, **Colwell FS**. 2018. Energy gradients structure microbial communities across sediment horizons in deep marine sediments of the south china sea. Frontiers in Microbiology **9**. doi:10.3389/fmicb.2018.00729.

28. **Johnston ER**, **Rodriguez-R LM**, **Luo C**, **Yuan MM**, **Wu L**, **He Z**, **Schuur EAG**, **Luo Y**, **Tiedje JM**, **Zhou J**, **Konstantinidis KT**. 2016. Metagenomics reveals pervasive bacterial populations and reduced community diversity across the alaska tundra ecosystem. Frontiers in Microbiology **7**. doi:10.3389/fmicb.2016.00579.

29. **Hassell N**, **Tinker KA**, **Moore T**, **Ottesen EA**. 2018. Temporal and spatial dynamics in microbial community composition within a temperate stream network. Environmental Microbiology **20**:3560–3572. doi:10.1111/1462-2920.14311.

365

Tables

Table 1. Summary of studies used in the analysis. For all studies, the number of sequences used from each dataset was rarefied to the smallest sample size. A graphical represenation of the distribution of sample sizes for each dataset and the samples that were removed from each dataset are provided in Figure

S1.

Dataset (Ref)	Samples	Total sequences	Median sequences	Range of sequences	SRA study accession
Bioethanol (19)	95	3,970,972	16,014	3,690-356,027	SRP055545
Human (20)	490	20,828,275	32,452	10,439-422,904	SRP062005
Lake (21)	52	3,145,486	69,205	15,135-110,993	SRP050963
Marine (22)	7	1,484,068	213,091	132,895-256,758	SRP068101
Mice (1)	348	2,785,641	6,426	1,804-30,311	SRP192323
Peromyscus (23)	111	1,545,288	12,393	4,454-33,502	SRP044050
Rainforest (24)	69	936,666	11,464	4,880-37,403	ERP023747
Rice (25)	490	22,623,937	43,399	2,777-192,200	SRP044745
Seagrass (26)	286	4,135,440	13,538	1,830-45,076	SRP092441
Sediment (27)	58	1,151,389	17,606	7,686-67,763	SRP097192
Soil (28)	18	932,563	50,487	46,622-58,935	ERP012016
Stream (29)	201	21,017,610	90,621	8,931-394,419	SRP075852

Figures
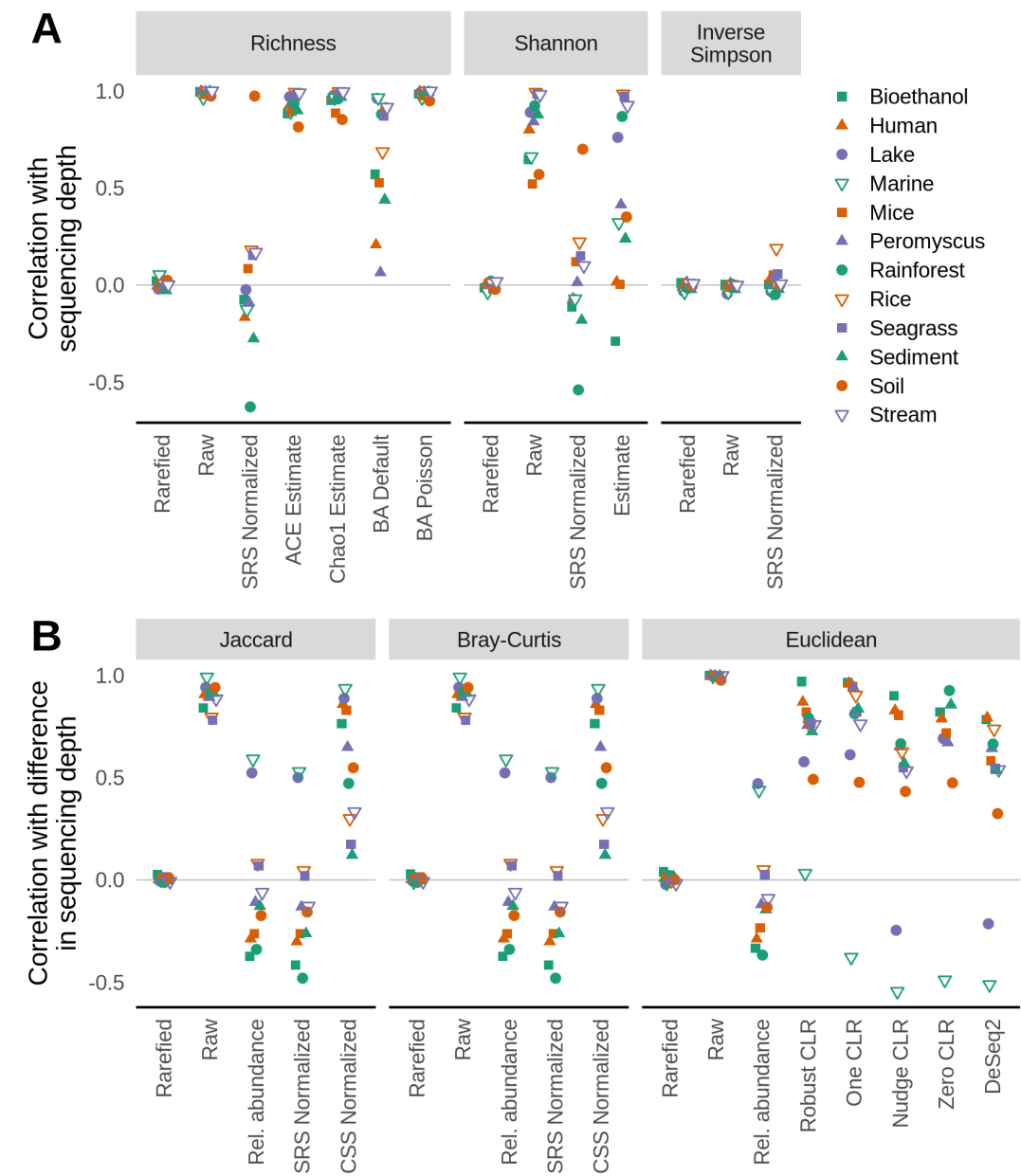
Figure 1. Rarefaction eliminates the correlation between sequencing depth and alpha diversity (A) and between differences in sampling depth and beta (B) diversity metrics when using null community models. Examples of the relationship between different metrics and methods for controlling for uneven sequencing effort are provided in Figures S2 and S3 for alpha and beta diversity metrics, respectively. Each point represents the mean of 100 random null community models; the standard deviation was smaller than the size of the plotting symbol.
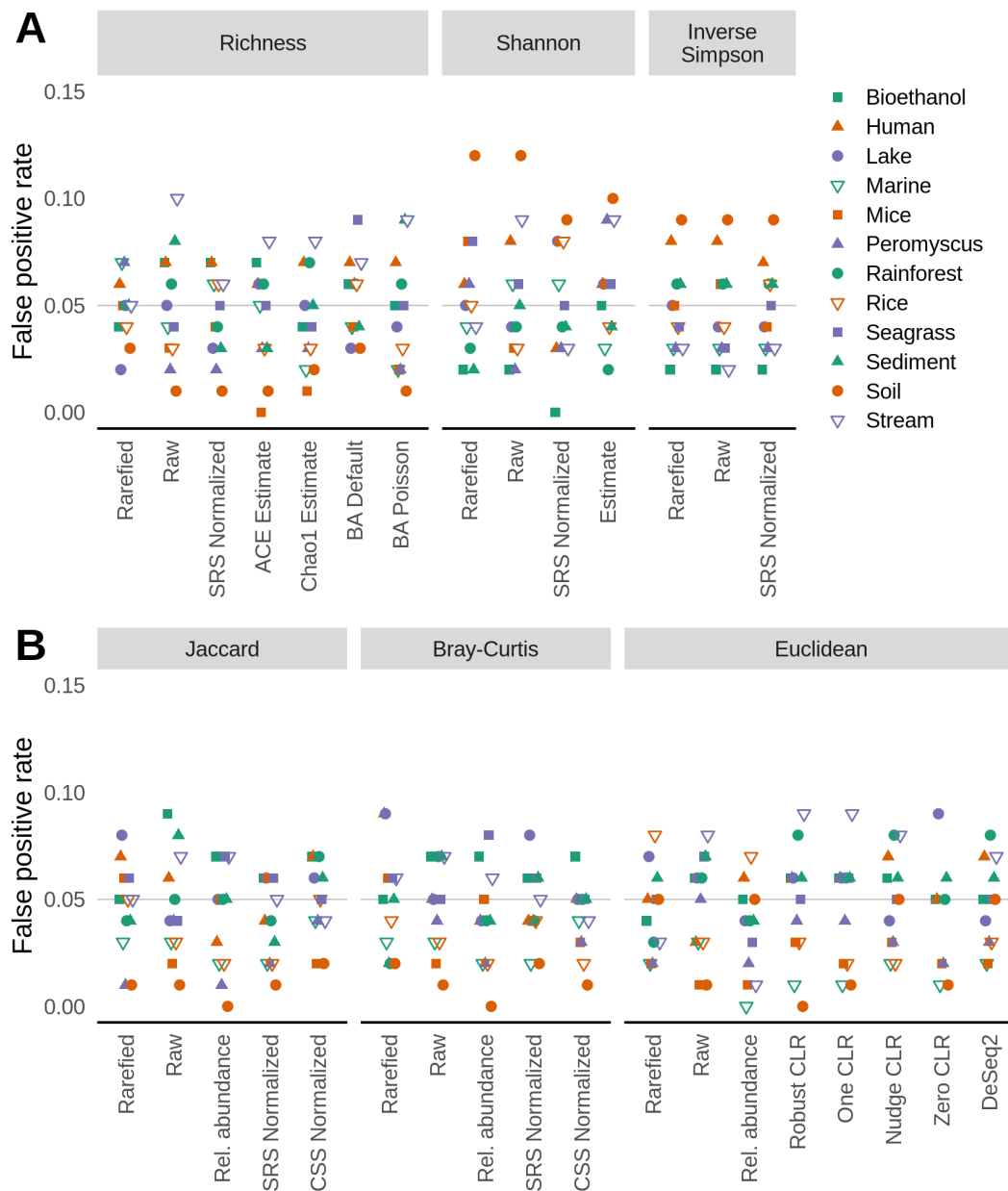
Figure 2. The risk of falsely detecting a difference between treatment groups drawn from a null model does not meaningfully vary from 5%, regardless of approach for controlling for uneven sequencing depth. Samples were randomly assigned to different treatment groups. To calculate the false detection rate, datasets were regenerated 100 times and differences in alpha diversity were tested using a Wilcoxon test (A) and differences in beta diversity were tested using PERMANOVA (B) at a 5% threshold. The false positive rate was the number of times a dataset yeilded a significant result.
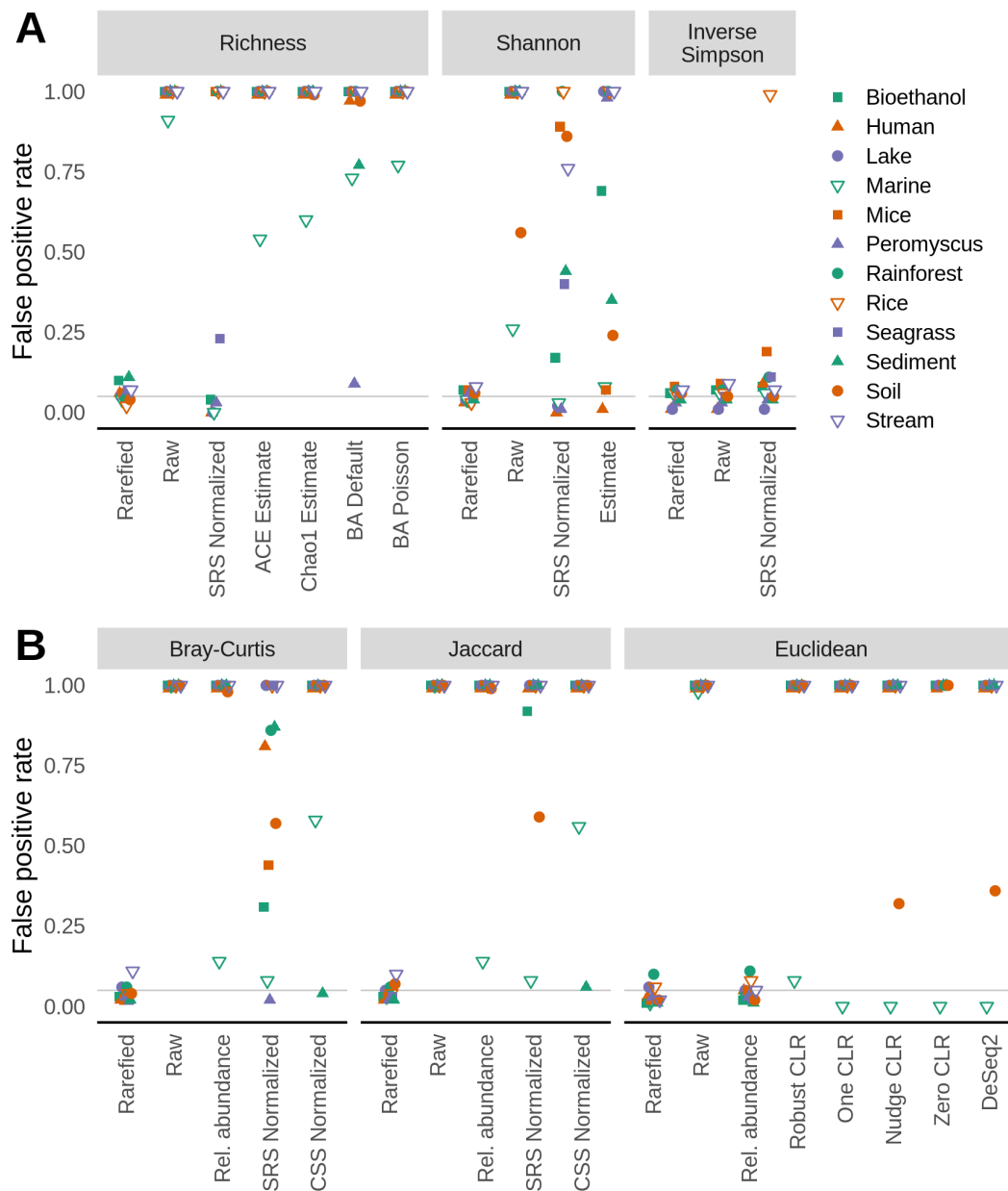
Figure 3.The risk of falsely detecting a difference between treatment groups drawn from a null model does not meaningfully vary from 5% when data are rarefied when sequencing depth is confounded with treatement group. Samples were assigned to different treatment groups based on whether they were above the median number of sequences for each dataset. To calculate the false detection rate, datasets were regenerated 100 times and differences in alpha diversity were tested using a Wilcoxon test (A) and differences in beta diversity were tested using PERMANOVA (B) at a 5% threshold. The false positive rate was the number of times a dataset yeilded a significant result.
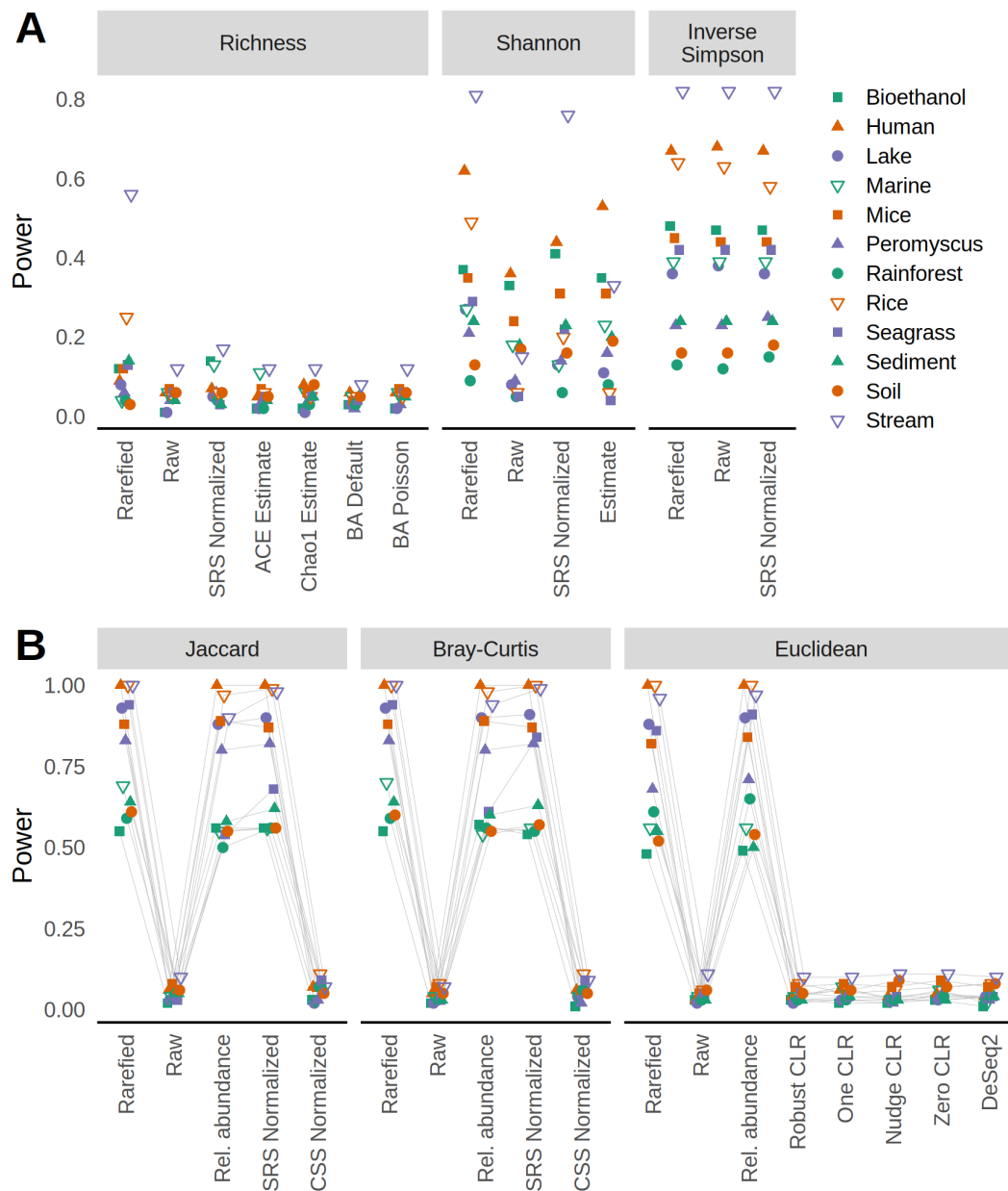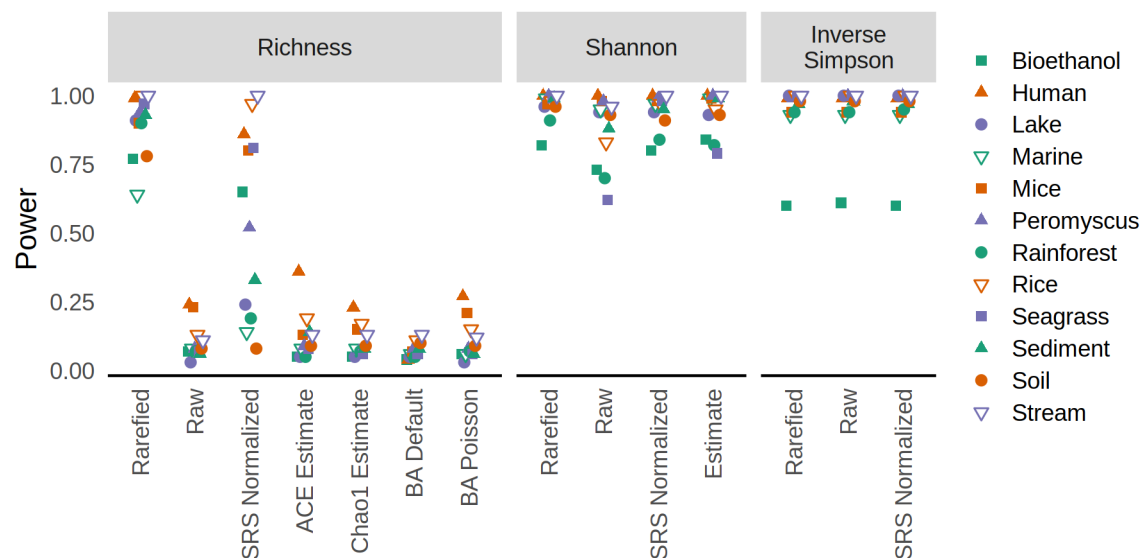
Figure 4. The ability to detect true differences in treatment groups for alpha (A) and beta (B) diversity metrics is greatest when communities differing in the relative abundance of their OTUs are rarefied. For each dataset samples were randomly assigned to one of two community distributions where the abundance of OTUs differed. To calculate the power for each datasets, datasets were regenerated 100 times and differences in alpha diversity were tested using a Wilcoxon test (A) and differences in beta diversity were tested using PERMANOVA (B) at a 5% threshold. The power was the number of times a dataset yielded a significant result.

Figure 5. The ability to detect true differences in treatment groups for alpha diversity metrics is greatest when communities differing in richness are rarefied. For each dataset samples were randomly assigned to one of two community distributions where one distribution contained a subset of OTUs found in the other. To calculate the power for each dataset, datasets were regenerated 100 times and differences in alpha diversity were tested using a Wilcoxon test (A) and differences in beta diversity were tested using PERMANOVA (B) at a 5% threshold. The power was the number of times a dataset yielded a significant result.

23
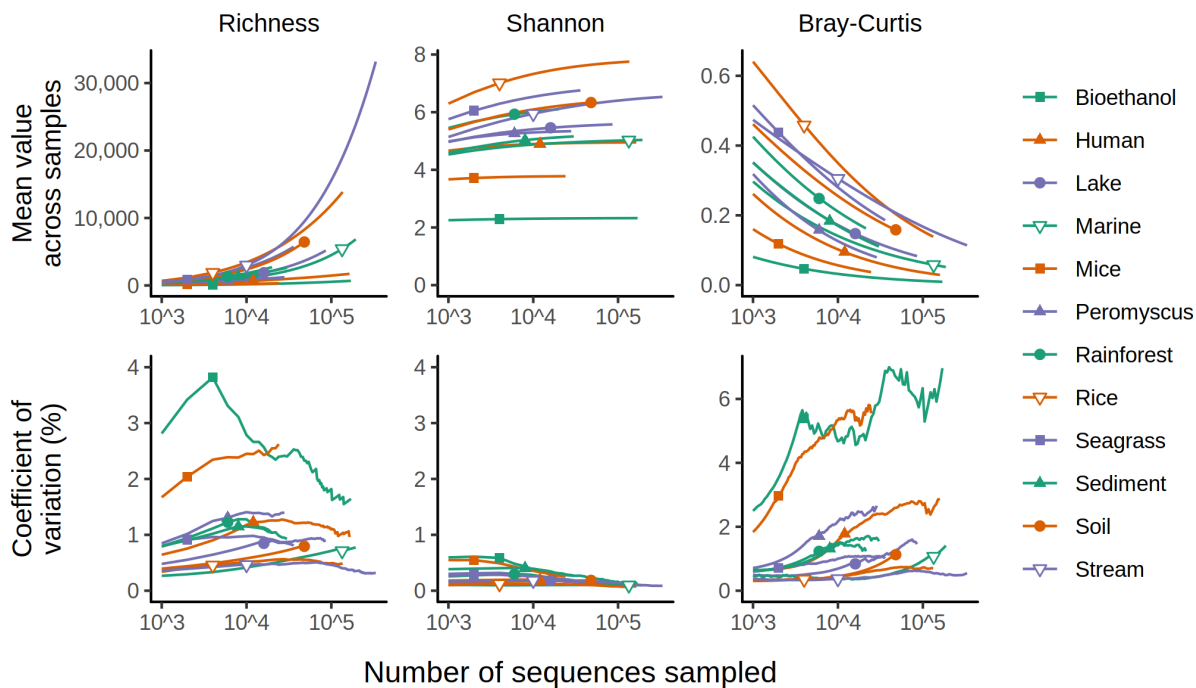
Figure 6. The mean and coefficient of variation for rarefied richness, shannon diversity, and Bray-Curtis dissimilarity vary with sequencing depth. For each dataset, a null community distribution was created and samples were created to have the same sequencing depth as they did originally. The placement of the plotting symbol indicates the size of the smallest sample. Results are only shown for sequencing depths where a dataset had 5 or more samples.
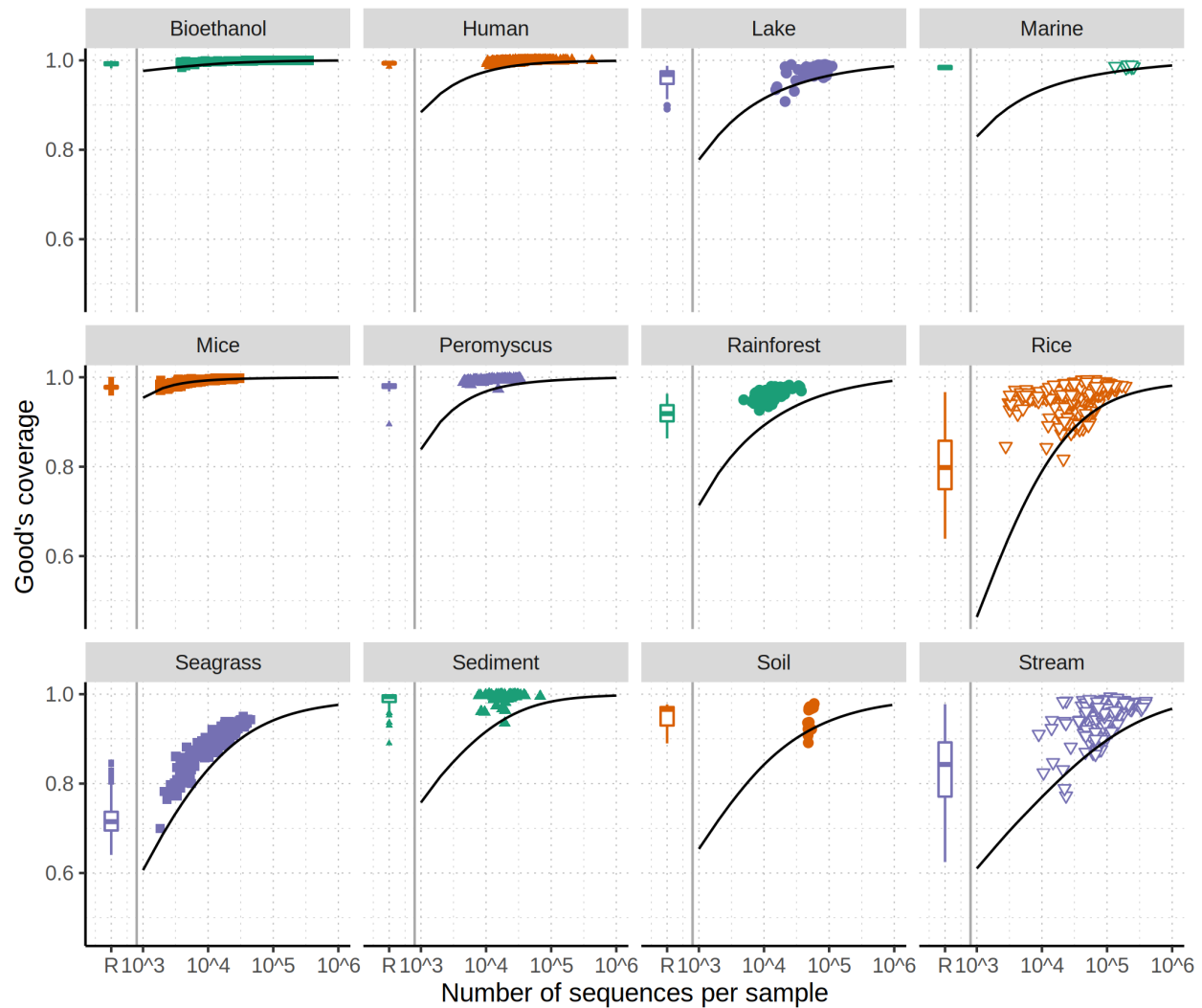
Figure 7. Most datasets are sequenced to a level that provides a high level of coverage. Each plotting symbol represents the observed Good's coverage for a different sample in each dataset. The smoothed line indicates the simulated coverage for varying levels of sampling effort when a null community is generated from the observed data. The box and whisker plot indicates the range of coverage values when the observed commmunity data were rarefied to the size of the least sequenced sample.
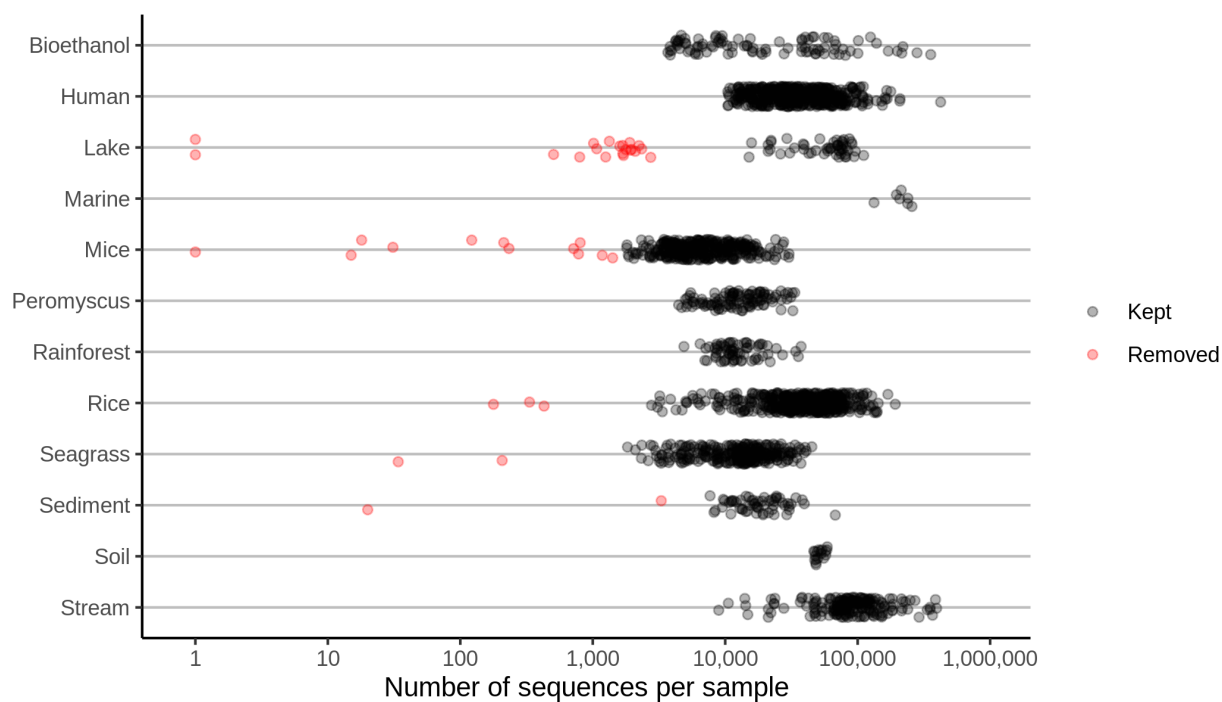
Figure S1. The number of sequences observed in each sample for each dataset included in this analysis generally varied by 10 to 100-fold. The threshold for specifying the number of sequences per sample varied by dataset and was determined based on identifying natural breaks in the data.
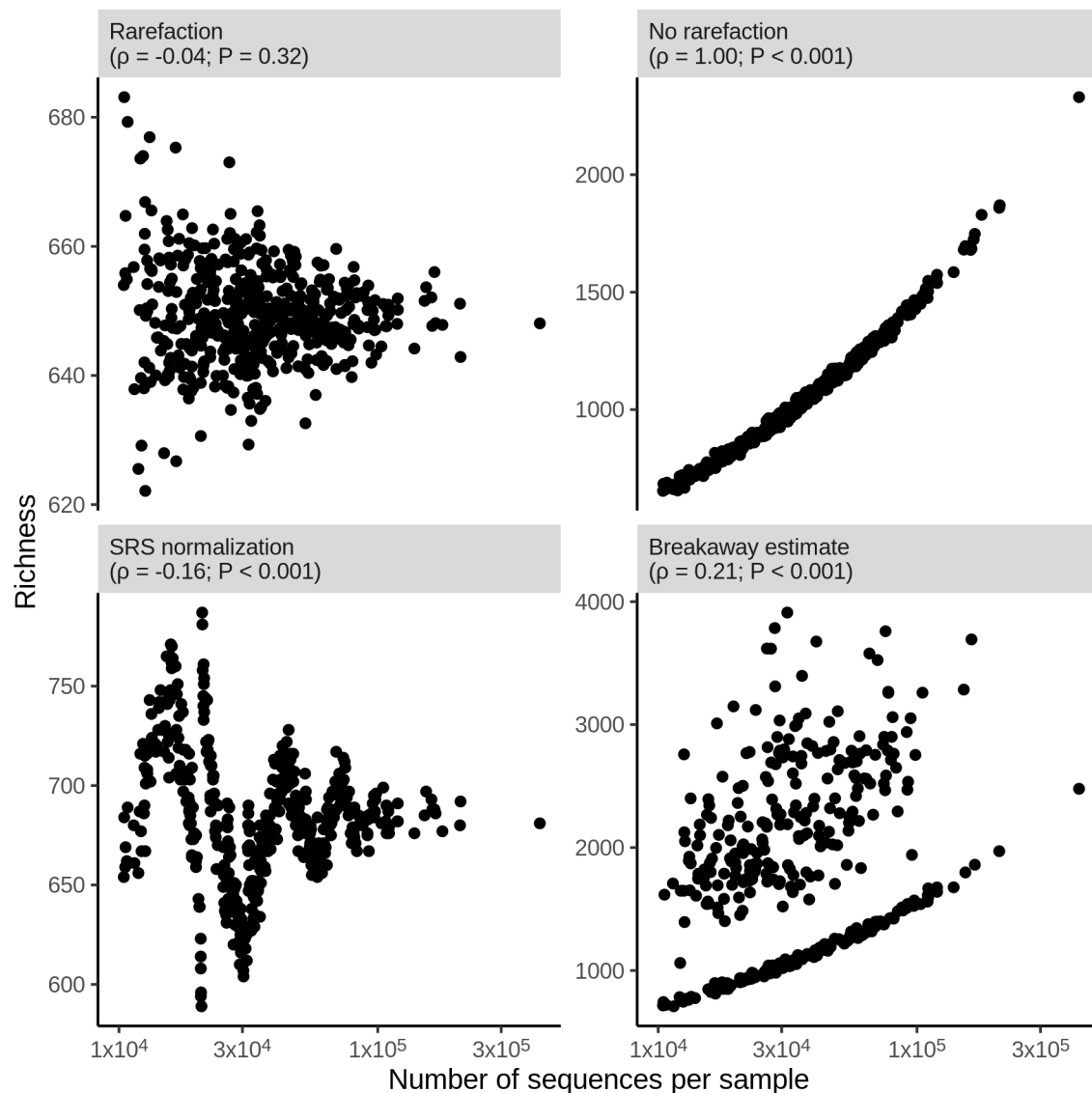
Figure S2. Examples of the richness in each of the 490 samples that were generated for one randomization of the null model using the human dataset. The x-axis indicates the number of sequences in each of the samples prior to each method's appraoch of controlling for uneven sampling effort. The Spearman correlation coefficient (ρ) and test of whether the coefficient was significantly different from zero are indicated for each panel.
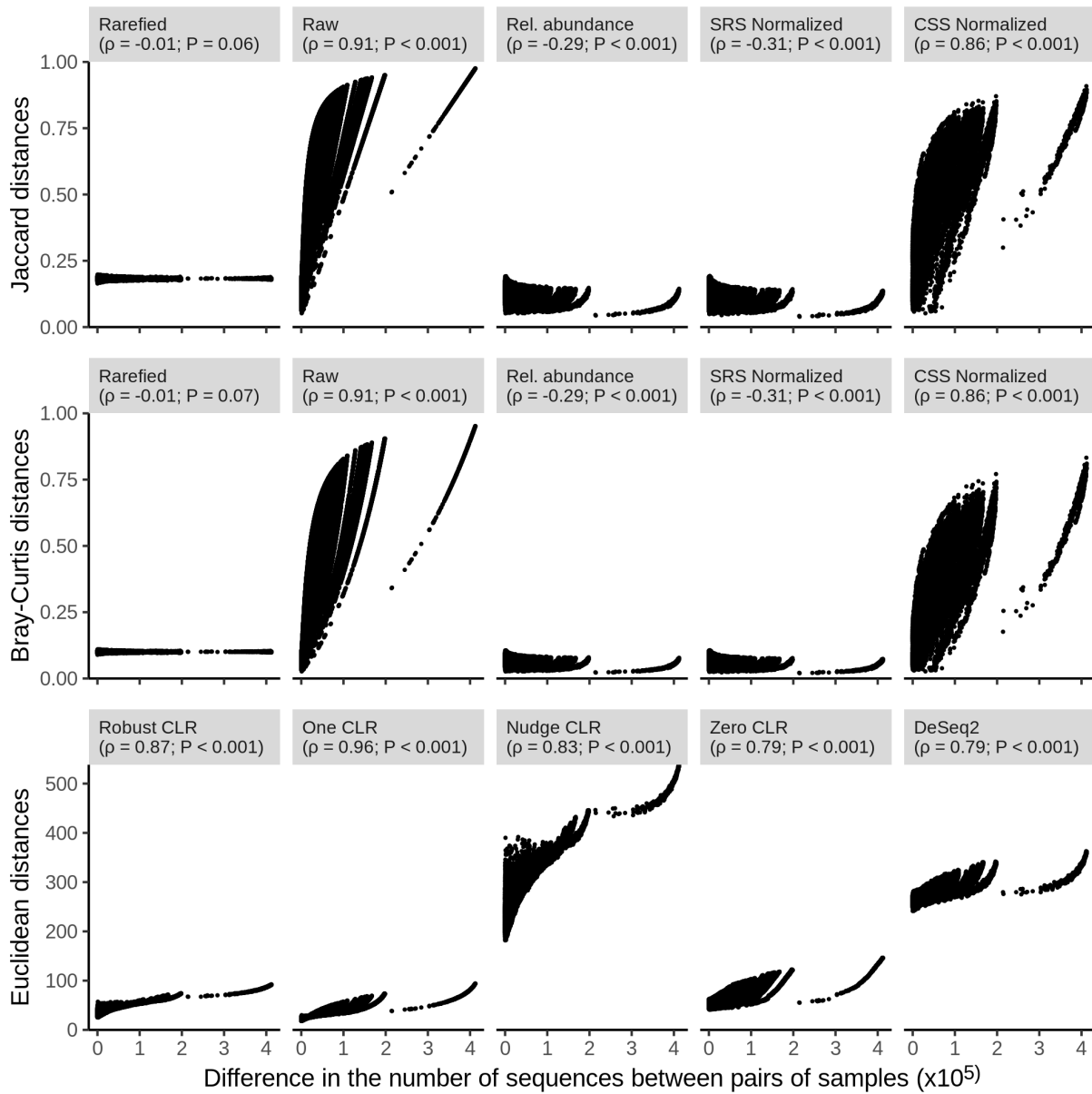
27

Figure S3. Examples of differences in beta diversity in each of the 490 samples that were generated for one randomization of the null model using the human dataset. The x-axis indicates the difference in the number of sequences in each of the samples that went into calcualting the pairwise distance prior to each method's appraoch of controlling for uneven sampling effort. The Spearman correlation coefficient (ρ) and test of whether the coefficient was significantly different from zero are indicated for each panel.