Rarefy your data

Running title: Rarefy your data

Patrick D. Schloss†

† To whom corresponsdence should be addressed:

pschloss@umich.edu

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

Research article

10 ## Abstract

Importance

Introduction

- 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
- Reasons behind “rarefaction is inadmissable”
  - Weird simulation
- Alternative approaches and claims
  - sampling invariance
- Goal of this study

Results

Choice of datasets. I selected 16S rRNA gene sequence data from from 12 studies that characterized the variation in bacterial communities from diverse environments (Table 1). The specific studies 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 the standard sequence curation pipeline to generate operational taxonomic units (OTUs) using the mothur software package (1, 2). 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 (Figure S1).

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 study. Under a null model, each sample from the same study would be expected to have the same alpha diversity regardless of the sampling effort. I assessed richness without any correction, using normalized OTU counts, with estimates based on non-parametric and parametric approaches, and using rarefaction. For each study, 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 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 1B). The rarefied alpha-diversity metrics consistently demonstrated a lack of sensitivity to sampling depth.

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 sample. First, I analyzed the sensitivity of the Jaccard distance coefficient, which incorporates whether an OTU is present in each community and not their relative abundance. When calculating Jaccard distances using the uncorrected OTU counts, normalized OTU counts, relative abundances, and rarefaction only the rarefied data showed a lack of sensitivity to sampling effort (Figure 2A). 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 2B). Third, I calcualted Aitchison distances on raw OTU counts where the central log-ratio (CLR) was calculated by ignoring OTUs in samples with zero counts (robust CLR), adding a pseudocount of 1 to all OTU counts prior to calculating the CLR (one CLR), XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXX XXXXXXXXXX (n CLR), and imputing the value of zero counts (z CLR). Regardless of the approach, the Aitchison distances were all strongly sensitive to sampling effort (Figure 2C). Finally, I used the cumulative sum scaling (CSS) normalization from metagenomeSeq and variance stabilization technique (VST) from DeSeq2 prior to calculating Euclidean distances. Both approachs revealed a strong sensitivity to sampling effort (Figure 2D). Although Euclidean distances are not typically 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 a significant diffeerence 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 (PERMANOVA). First, within each study, I randomly assigned each sample to one of two treatment groups. My expectation was that approximately 5 of the 100 (5%) random tests for each comparison would yield a significant test result. Indeed, for each study and alpha and beta diversity metric and strategy for accounting for uneven sampling, approximately 5% of the tests yielded a significant result (Figure 3). Second, within each study, I assigned samples with more than the median number of sequences per sample to one treatment group and the rest to another treatment group. If there is no sensitivity to sampling effort, I would again expect that 5% of the tests would yield a significant result. In fact, only the rarefied data consistently resulted in a 5% false positive rate for alpha and beta diversity metrics (Figure 4). These results align 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, for each study samples were randomly assigned to one of two treatment groups. Samples in the first treatment group were generated by sampling from the null distribution. Samples in the second treatment group were generated by perturbing the null distribution 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. The fraction of tests that yielded a significant test was a measure of the statistical power for the test. Relative to the rarefied data, the power to detect differences in alpha and beta diversity was considerably lower for each of the strategies for handling uneven sampling effort. The power to detect differences in richness by all approaches was low (Figures 7 and 8). This was likely because the approach to generating the second community did not necessarily change the number of OTUs in each treatment group. To explore this further, in the second simulation the second treatment group was perturbed by removing 3% of the OTUs from the model. Again, the rarefied data had a considerably higher statistical power than the other approaches when measuring richness (Figure 9). Both simulations highlight 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. To assess the sensitivity of alpha and beta diversity metrics to rarefaction depth, I again used the dataset generated using the null models and rarefied them to varying depths. For the alpha diversity metrics, the value of the metrics plateaued and the coefficient of variation (i.e., the mean divided by the standard deviation) between samples remained constant as the rarefaction depth increased (Figure 10). For beta diversity metrics, the distance between samples and the coefficient of variation between samples decreased as sampling depth increased (Figure 11). These results confirm that greater sequencing depth provides a more robust estimate of the metrics.

Most studies have a high level of sequencing coverage. I calculated the Good's coverage for the observed data and found that each of the studies had a minimum coverage greater than 90% at their lowest sequencing depth (Figure 12A). This suggests that most studies have a high level of sequencing coverage. Next, I returned to the null models to ask how much sequencing effort was required to obtain higher levels of coverage. To obtain 95 and 99% coverage, an average of XX and XX-fold more sequence data was required, respectively (Figure 12B). Although it is clear that most researchers would desire greater coverage, the sequencing effort required to acheive that sequencing depth would likely limit the number of samples that could be sequenced when controlling for costs.

Evidence that not rarefying data can impact results in original studies. To assess the impact of using alternatives to rarefaction, I reassessed hypotheses from the human, mice, and peromyscus studies. These studies were selected because they originated from my research group and I know the datasets well. First, the study that published the human dataset was interested in the difference in the fecal microbial communities of patients with and without colorectal cancer. Here, I divided the patients into peope with and without screen relevant neoplasia (SRN). Among the alpha diversity metrics, the test were non-significant with similar effect sizes; however, the richness estimates obtained using parametric appraoches detected statistically significant differences in richness. Although the tests using Bray-Curtis distances were significant regardless of whether rarefaction was used, the effect sizes were smaller for the distances calculated using the CLR, VST, CSS transformations.

- Human study: Did not alter effect size or significance of alpha or beta diversity, but did result in reduced effect sizes for measures of richness and non-parametric estimators of richness; breakaway detected a difference

Mouse

Peromyscus

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

Null models. Null models were generated by randomly assigning each sequence in the study to an OTU and sample while keeping constant the number of sequences per sample and the total number of sequences in each OTU. Because the construction of the null models was a stochastic process, 100 replicates were geneated for each study.

Data availability.

Reproducible data analysis.

Acknowledgements.

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. **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.

4. **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.

5. **Beall BFN**, **Twiss MR**, **Smith DE**, **Oyserman BO**, **Rozmarynowycz MJ**, **Binding CE**, **Bourbonniere 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.

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

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

8. **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 involved in DOC, CO2, and CH4 fluxes in temperate rainforest soil. The ISME Journal **13**:950–963. doi:10.1038/s41396-018-0334-3.

9. **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.

10. **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.

11. **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.

12. **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.

13. **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.

Table 1. Summary of studies used in the analysis. For all studies, the number of sequences used from each study was rarefied to the smallest sample size. A graphical represenation of the distribution of sample sizes for each study and the samples that were removed from each study are provided in Figure S1.

Study (Ref)	Samples	Total sequences	Median sequences	Range of sequences	SRA study accession
Bioethanol (3)	95	3,970,972	16,014	3,690-356,027	SRP055545
Human (4)	490	20,828,275	32,452	10,439-422,904	SRP062005
Lake (5)	52	3,145,486	69,205	15,135-110,993	SRP050963
Marine (6)	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 (7)	111	1,545,288	12,393	4,454-33,502	SRP044050
Rainforest (8)	69	936,666	11,464	4,880-37,403	ERP023747
Rice (9)	490	22,623,937	43,399	2,777-192,200	SRP044745
Seagrass (10)	286	4,135,440	13,538	1,830-45,076	SRP092441
Sediment (11)	58	1,151,389	17,606	7,686-67,763	SRP097192
Soil (12)	18	932,563	50,487	46,622-58,935	ERP012016
Stream (13)	201	21,017,610	90,621	8,931-394,419	SRP075852

Figure 1.

Figure S1.