



RESEARCH ARTICLE

Methods for normalizing microbiome data: An ecological perspective

Donald T. McKnight¹  | Roger Huerlimann¹  | Deborah S. Bower^{1,2} |
Lin Schwarzkopf¹ | Ross A. Alford¹ | Kyall R. Zenger¹

¹College of Science and Engineering, James Cook University, Townsville, Queensland, Australia

²School of Environmental and Rural Science, University of New England, Armidale, New South Wales, Australia

Correspondence

Donald T. McKnight
Email: donald.mcknight@my.jcu.edu.au

Handling Editor: Simon Jarman

Abstract

1. Microbiome sequencing data often need to be normalized due to differences in read depths, and recommendations for microbiome analyses generally warn against using proportions or rarefying to normalize data and instead advocate alternatives, such as upper quartile, CSS, edgeR-TMM, or DESeq-VS. Those recommendations are, however, based on studies that focused on differential abundance testing and variance standardization, rather than community-level comparisons (i.e., beta diversity). Also, standardizing the within-sample variance across samples may suppress differences in species evenness, potentially distorting community-level patterns. Furthermore, the recommended methods use log transformations, which we expect to exaggerate the importance of differences among rare OTUs, while suppressing the importance of differences among common OTUs.
2. We tested these theoretical predictions via simulations and a real-world dataset.
3. Proportions and rarefying produced more accurate comparisons among communities and were the only methods that fully normalized read depths across samples. Additionally, upper quartile, CSS, edgeR-TMM, and DESeq-VS often masked differences among communities when common OTUs differed, and they produced false positives when rare OTUs differed.
4. Based on our simulations, normalizing via proportions may be superior to other commonly used methods for comparing ecological communities.

KEYWORDS

Bray–Curtis, community comparisons, diversity, evenness, ordination, principal coordinates analysis, simulation

1 | INTRODUCTION

Using high-throughput sequencing to examine microbial communities has become a common practice. These techniques are, however, not without their pitfalls, and it is important for researchers to use the most appropriate analytical methods for answering the ecological questions at hand. One common pitfall stems from the fact that sequencing results in variable numbers of reads per sample. These differences in read depth often need to be corrected prior to analyses, and many methods have been proposed for normalizing data.

Two of the oldest and most intuitive methods are (a) transforming the data to proportions by dividing the reads for each operational taxonomic unit (OTU) in a sample by the total number of reads in that sample (also known as Total Sum Normalization [TSS]) and (b) rarefying the data by randomly subsampling each sample to the lowest read depth of any sample. In recent years, however, both methods have been heavily criticized. Proportions are criticized because they do not account for heteroskedasticity (Weiss et al., 2017) and result in spurious correlations when comparing the abundance of specific OTUs relative to other OTUs (Jackson, 1997). Rarefying is

criticized because it discards potentially useful data (McMurdie & Holmes, 2014; but see Weiss et al., 2017). Further, several studies have documented that proportions and rarefied data perform poorly in differential abundance testing and often have high type I error rates (Bullard, Purdom, Hansen, & Dudoit, 2009; Dillies et al., 2013; McMurdie & Holmes, 2014; Weiss et al., 2017). As a result, other methods have been proposed and have rapidly gained popularity. These methods include, upper quantile normalization (UQ; Bullard et al., 2009), CSS normalization implemented in the R package METAGENOMES (Paulson, Stine, Bravo, & Pop, 2013), a variance stabilizing transformation implemented in the R package DESeq2 (hereafter referred to as DESeq-VS; Love, Huber, & Anders, 2014), and a trimmed mean of M-values normalization implemented in the R package edgeR (hereafter referred to as edgeR-TMM; Robinson, McCarthy, & Smyth, 2010; McCarthy, Chen, & Smyth, 2012).

Several studies have contrasted the effectiveness of these normalization methods, generally favouring CSS, DESeq-VS, and edgeR-TMM; however, they have usually judged the methods based on how well they standardized the within-sample variance across samples, whether they allowed data to cluster in ordination plots, and how well they performed in differential abundance testing (Bullard et al., 2009; Dillies et al., 2013; Lin et al., 2016; McMurdie & Holmes, 2014; Paulson et al., 2013; Weiss et al., 2017). By those metrics, proportions and rarefying perform poorly, which has often led to blanket recommendations against using them. From an ecological perspective, however, there are additional performance measures that are important to consider. Specifically, it is valuable to determine whether these methods produce accurate comparisons among entire communities (i.e., beta diversity), rather than simply whether specific OTUs differ.

The Bray–Curtis (BC) dissimilarity metric is one of the most easily interpreted and widely used methods for comparing communities, particularly in microbiome analyses. It can be used as a stand-alone measure of dissimilarity, as well as providing dissimilarity matrices that are used for constructing ordination plots and making statistical comparisons among sets of communities (e.g., PERMANOVAs). Bray–Curtis dissimilarities, and most other distance and dissimilarity measures, do not require equal variances, and there is good reason to think that standardizing the variance prior to calculating BC would distort patterns, rather than clarifying them. Therefore, this paper will first discuss the ecological reasons why transforming to proportions

or rarefying may be the most suitable methods for transforming ecological data prior to calculating distance or dissimilarity measures, then it will provide both real and simulated data to illustrate the concepts. It is important to note that while we will focus on BC scores throughout this paper, our arguments and conclusions also apply to other community comparison metrics that incorporate abundance.

1.1 | The importance of fully standardizing reads

The first potential pitfall of transformation methods such as UQ, CSS, edgeR-TMM, and DESeq-VS is that, unlike proportions and rarefying, they do not guarantee that the number of reads will be equal across samples. This is problematic, because measures like BC are affected by differences in read depths, sometimes in unintuitive ways. For example, consider the four hypothetical samples in Figure 1; S1 and S2 are samples from the same community, but S2 has twice the read depth of S1. As a result, the BC between them is 0.333, even though they are from the same community and should have a BC of 0. Furthermore, the community from which S3 was sampled is only slightly different from that of S1, whereas S4's community differs strongly from S1's. Nevertheless, because S3 and S4 both have twice the read depth of S1, the BC for both samples is 0.333 when compared to S1. Indeed, when comparing two samples where the read depth of one sample is twice that of the other, the BC will always be a minimum of 0.333 (it will be exactly 0.333 if the number of reads for each individual OTU is also equal to or greater than the number of reads for that OTU in the other sample). Thus, the differences in read depths have rendered the community-level comparisons among these samples meaningless, and even misleading. Therefore, the fact that many normalization methods do not guarantee standardized read depths raises serious concerns about their applicability for community-level comparisons.

1.2 | The importance of species evenness

The diversity of a community can be partitioned into species richness (i.e., the number of species present) and species evenness (i.e., the relative abundance of the species present). Evenness (and its inverse, dominance) is an important aspect of diversity (Hillebrand, Bennett, & Cadotte, 2008; Stirling & Wilsey, 2001; Wilsey, Chalcraft, Bowles, & Willig, 2005) that has strong effects on community function and

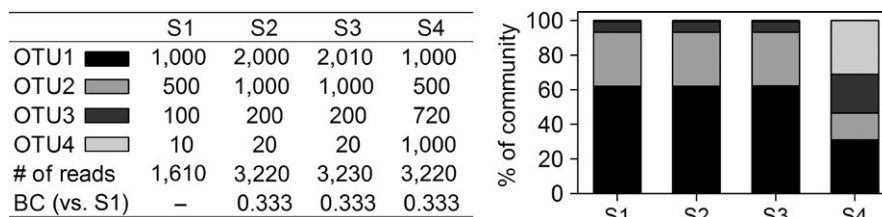


FIGURE 1 Samples (S1–S4) from four hypothetical communities illustrating the potential problems that arise when samples have different numbers of reads. The data are shown both as a table of raw read counts and a stacked bar plot. The bar plot illustrates the fact that S1, S2, and S3 are nearly identical after accounting for read depth, whereas S4 is distinct. Nevertheless, all samples have the same BC when compared to S1. BC = Bray–Curtis dissimilarity between S1 and the sample in a given column

stability (Ghazoul, 2006; Hillebrand & Cardinale, 2004; Wittebolle et al., 2009), resistance to invasion (Wilsey & Polley, 2002), and the influence of species richness on community functions (Hillebrand et al., 2007). Therefore, species evenness is an important consideration when comparing communities.

Nevertheless, many normalization methods (e.g., UQ, CSS, edgeR-TMM, and DESeq-VS) focus on standardizing the within-sample variance across samples (i.e., forcing each sample to have the same distribution of reads; Dillies et al., 2013; Lin et al., 2016). For some statistical tests, such as most methods for differential abundance testing, having the same variance in each sample is important, but it is potentially problematic when comparing entire communities, because variance and evenness are tightly linked. A highly even community (i.e., a community where all the members are roughly equally abundant) will also have a low variance (i.e., there will be a low variance within the community because all the OTUs will be present in similar numbers); whereas a community with low evenness (i.e., a community where a few members dominate) will have a high variance. Therefore, by standardizing the variance across samples, these methods suppress differences in species evenness.

Consider, for example, two communities, each of which consist of the same 100 OTUs, but one has high evenness and the other has very low evenness. These communities will differ greatly in their variances, but that difference in variances is not only important; it is the critical distinction between those communities and standardizing the variance would mask that crucial difference.

1.3 | Dominant species vs rare species

The next potential problem is that methods like UQ, CSS, edgeR-TMM, and DESeq-VS employ log transformations as part of their mechanism for standardizing variances (generally a log base 2 with a plus one pseudocount). The purpose behind this is to reduce the effect of highly abundant OTUs so that the effects of rare OTUs can be seen. With the exception of CSS, these methods originated for RNA-seq data where reducing the effect of dominant genes is vital to detect differences among rare genes; however, its utility for community data is less clear. Although rare members of an ecological community often perform important functions (Fuhrman, 2009; Pedrós-Alió, 2006), the dominant members tend to drive the bulk of community functionality (Cottrell & Kirchman, 2003; Fuhrman, 2009; Zhang, Jiao, Cottrell, & Kirchman, 2006). Therefore, reducing the importance of dominant OTUs and amplifying the importance of rare OTUs may give a misleading picture of the differences among communities.

Consider, for example, the hypothetical communities in Figure 2. S5 and S6 are nearly identical, whereas S7 clearly differs from S5, and those similarities and differences are conveyed by the BC values in the raw data. After log transforming the data, however, the difference between S5 and S6 (based on BC) increases, while the difference between S5 and S7 is greatly reduced. Indeed, based on the log-transformed data, one would incorrectly conclude that S7 is the community that is most similar to S5. This erroneous result arises

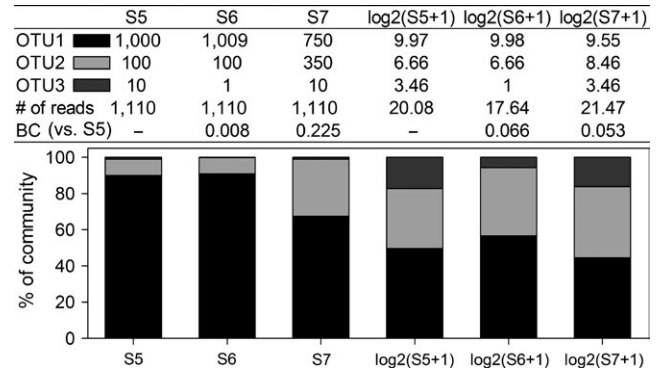


FIGURE 2 Samples (S5–S7) from three hypothetical communities illustrating the potential problems that arise from log transforming community data. The samples are shown with and without a $\log_2(x + 1)$ transformation, and the data are shown both as a table of raw read counts and a stacked bar plot. The bar plot illustrates the fact that the log transformation increases the importance of rare operational taxonomic units (OTUs) which decreasing the importance of common OTUs, ultimately suppressing the differences between S5 and S7 and exaggerating the differences between S5 and S6. BC = Bray–Curtis dissimilarity between S5 and the sample in a given column (for the log-transformed data, the comparisons were made with the log-transformed S5)

from the fact that the log transformation amplified the slight differences between S5 and S6 for OTU3, while suppressing the large differences between S5 and S7 for OTU1 and OTU2.

2 | MATERIALS AND METHODS

2.1 | Mouse gut microbiomes

To examine the potential problems with normalization methods, we applied several different transformations to a mouse gut microbiome dataset (Turnbaugh et al., 2009), previously used in the paper proposing CSS (Paulson et al., 2013; available in the METAGENOMESEQ package). We normalized the data using proportions, rarefying (performed in the PHYLOSEQ package; McMurdie & Holmes, 2013), UQ (performed in the EDGE R package), CSS (performed in the METAGENOMESEQ package), edgeR-TMM, and DESeq-VS (with “blind” set to False). UQ, CSS, edgeR-TMM, and DESeq-VS generally apply a \log_2 transformation with a pseudocount of 1 as the final step, but because we were also interested in the effects of log transformations, we normalized the data with and without the log transformation for each method (including proportions, rarefied data, and the original [true] data). The choice of pseudocount affects the log-transformed results, and, for results to be comparable, it is important for the scale of the pseudocount relative to the total number of reads to be similar across normalization methods (Costea, Zeller, Sunagawa, & Bork, 2014). Therefore, for proportions, UQ and edgeR-TMM, the normalized results were multiplied by 10,000 prior to the log transformation, and for CSS the results were multiplied by 1,000 prior to log transformation (which is standard for CSS). Scaling the results

by a constant value does not affect the BC results for the normalized data prior to the log transformation, but it does affect the BC value following the log transformation, and scaling by these values was necessary for the log-transformed data to be comparable across methods (Costea et al., 2014).

For each normalization method, we examined the spread of the data (i.e., maximum number of reads per sample, minimum numbers of reads, mean number of reads per sample, and per cent difference between the maximum and minimum number of reads) to see how well the methods standardized the read depths across samples. Additionally, to test how accurately the methods performed for BC comparisons, we identified 81 pairs of samples in which the per cent difference between the read depth for the original (non-normalized) data was <0.5%. Because those samples were extremely similar in read depth, they were comparable without normalizing. Therefore, we calculated BC dissimilarities within each pair of samples for the original data (without normalizing), and we considered those comparisons to be the true results. Then, we calculated BC dissimilarities for each pair using each normalization method and compared the results with the results from the original data. For each method, we normalized the entire dataset prior to subsetting to these pairs, and each sample was compared to the sample with the closest read depth. The package *VEGAN* (Oksanen et al., 2017) was used for all BC calculations. Metadata for the pairs of samples we analysed, as well as additional analyses comparing samples of different diet types are available in Supporting Information.

2.2 | Simulated data

To further compare the results of different normalization methods, we wrote a simulation in R to conduct a mock microbiome study involving two populations. Briefly, the simulation took a distribution of OTUs and randomly sampled from it to form an initial distribution for population 1 (consisting of an amount of DNA per OTU). Then, for each OTU in that distribution, it randomly selected a number from a normal distribution with a user-defined mean (hereafter called the mean dissimilarity) and a SD of 0.3 times that mean. It then multiplied the DNA yield for that OTU by that number and randomly added or subtracted the resulting amount of DNA. This produced a second initial distribution that was used to form population 2 (it could also be set so that only OTUs in a given percentile [based on the amount of DNA in the distribution for population 1] varied between the two distributions). A similar procedure was then used to generate 10 individuals in each population, based on the two distributions (each individual was a microbiome sample). The amount of DNA was then standardized (as occurs in real studies) and each sample was “read” by randomly sampling from it (with replacement). The number of reads per sample was randomly selected from a user-defined range.

Next, the data were normalized using each method as described in the “Mouse gut microbiomes” section, and for each method, the simulation returned the maximum and minimum read depths for the 20 simulated samples, as well as the *p*-value and *R*² value for a linear regression between read depth and BC (mean per sample based

TABLE 1 Read depths for the mouse gut microbiome dataset based on different normalization methods

	Original	Proportions	Rarefied	UQ	CSS	edgeR-TMM	DESeq-VS	Original log	Proportions log	Rarefied log	UQ log	CSS log	edgeR-TMM log	DESeq-VS log
Max	5,808	10,000	848	39,352	22,512	26,408	5,878	1,081	1,825	404	2,504	1,719	1,843	1,126
Min	848	10,000	848	3,535	4,871	5,644	882	342	891	192	511	825	842	359
Mean	2,270	10,000	848	11,579	9,340	10,196	2,332	647	1,282	318	1,313	1,238	1,285	675
SD	654	0	0	6,662	2,097	2,254	662	129	170	39	411	154	189	134
% diff	85.4	0.0	0.0	91.0	78.4	78.6	85.0	68.4	51.2	52.5	79.6	52.0	54.3	68.2
max-min														
% diff pairs	0.2	0.0	0.0	69.8	16.0	16.2	0.4	12.4	12.6	12.5	35.1	7.1	12.8	12.4

Notes. % diff max-min = the per cent difference between the maximum and minimum read depth, % diff pairs = the mean per cent difference in read depth between the 81 pairs of samples where, prior to normalization, the per cent difference in read depth was <0.5% (i.e., after normalization, the per cent difference was calculated for each pair, then averaged across pairs). The “Original” column shows the data prior to any normalization. For the pairs of samples where read depths were similar beforehand, most normalization methods actually increased the differences between samples.

on comparisons to all other samples). Additionally, it performed a PERMANOVA between the two populations via the package VEGAN (Oksanen et al., 2017). Finally, it returned the BC between the first

individual in each population. All of these calculations were also performed on the original, standardized samples prior to sequencing. These standardized samples all had the same amount of DNA (with slight rounding errors) and represented the true communities (they will be referred to as “original” throughout). Thus, they provided a baseline for testing how well the methods performed. Although standardizing DNA yields prior to sequencing is a component of real studies, in simulations, it is mathematically equivalent to transforming to proportions; therefore, to ensure that this did not bias our results in favour of proportions, we also conducted several tests where the baseline points of comparison were the raw samples (prior to standardization for sequencing) with a UQ, CSS, edgeR-TMM, or DESeq-VS normalization. These tests did not alter our results and are presented and discussed in Supporting Information.

We used this simulator to simulate 200 iterations each for all combinations of the following conditions: mean dissimilarity between populations = 0, 0.2, 0.4, 0.6, 0.8 (when mean dissimilarity = 0, the two populations were formed from the same distribution); range of possible read depths = 5,000–15,000 and 1,000–20,000; OTUs that varied between population starting distributions = all, top 10% [i.e., only the OTUs in the 90th percentile and above based on DNA yield in the population 1 distribution], and the bottom 30%.

We used a variation of that simulator to examine the effect of normalization methods on clustering in ordination plots. It constructed populations as above, but it simply returned principal coordinates analysis (PCoAs) based on BC for each normalization method.

We used several metrics to judge the performance of the normalization methods. First, we compared their ability to standardize read depths by examining the per cent difference in read depths between the sample with the highest read depth and the sample with the lowest read depth within each iteration. Next, we examined the accuracy of the BC estimates by constructing scatter plots comparing the BC estimates from normalized data to the BC estimates from the original communities. We expected normalization methods that accurately reflected the original communities to have little variation between the original and normalized BC values (i.e., a high R^2), slopes close to 1, and intercepts close to 0. We also compared the results of the PERMANOVAs, correlations between read depth and BC, and PCoAs, with the expectation that methods appropriate for

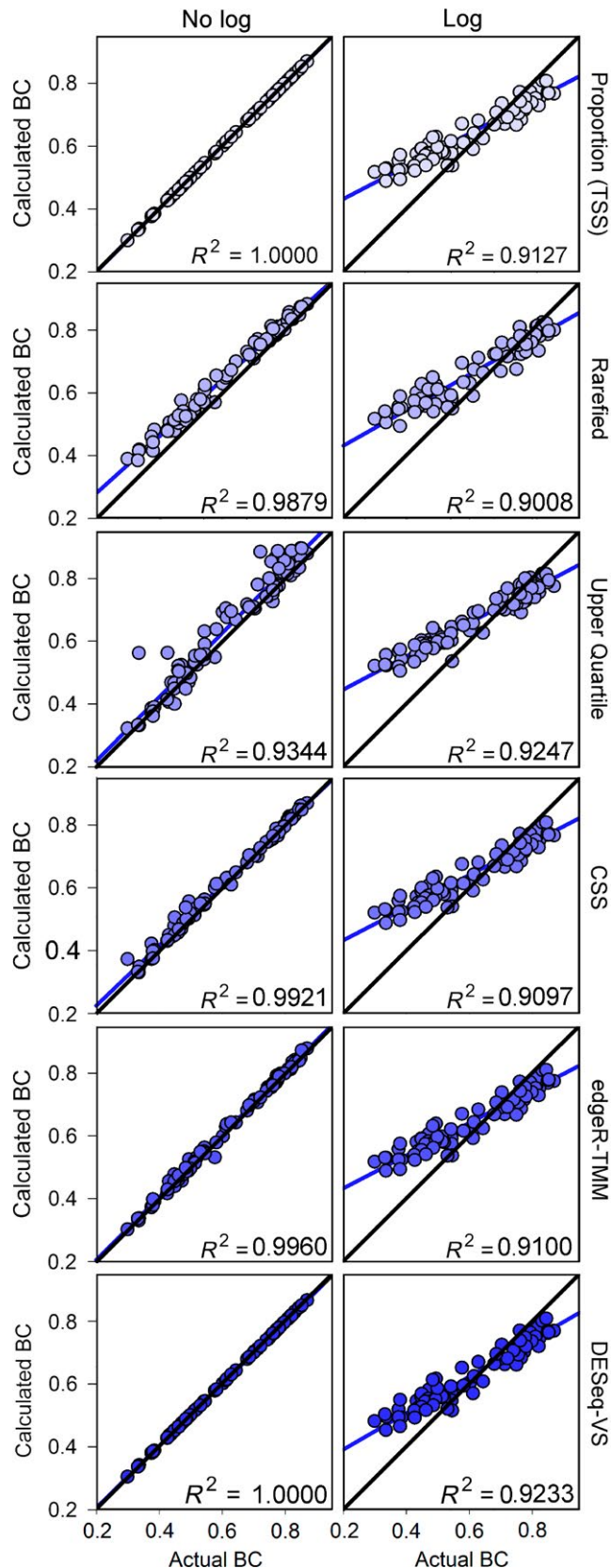


FIGURE 3 Correlations between the Bray–Curtis dissimilarities for the original (non-normalized [true]) data and the Bray–Curtis dissimilarities following normalization. Black lines show a slope of 1 and intercept of 0. These data are from the mouse gut microbiome dataset, and only the 81 pairs of samples where the per cent difference between read depths was <0.5% for the original data are shown (all data were used during the normalization step). It should be noted that DESeq-VS has the option of doing transformations “blind” (i.e., without incorporating *a priori* knowledge about groups) or with *a priori* knowledge. For this dataset, the results were highly inaccurate if *a priori* information was used. Therefore, we presented the results without *a priori* information here, and the results with *a priori* information are available in Supporting Information

comparing communities should yield results that are similar to the results from the original communities.

3 | RESULTS

3.1 | Mouse gut microbiomes

All normalization methods except for proportions and rarefying performed poorly in terms of their ability to standardize the read depth across samples (Table 1). In every case (except proportions and rarefied data), the sample with the deepest read depth had over twice the number of reads as the sample with the lowest read depth. Additionally, for the 81 pairs of samples that had similar read depths before standardization, all methods that did not involve a log transformation produced BC dissimilarities that correlated closely with the BC estimates from the untransformed data, though rarefied data had a slightly inaccurate slope and the UQ data had more variation than the other methods (Figure 3). After applying the log transformation, however, the results for all methods had increased variation in the relationship between the original and normalized BC values, the slopes of the regressions deviated strongly from 1, and the intercepts deviated from 0.

3.2 | Simulated data

All normalization methods except proportions and rarefying performed poorly in terms of their ability to standardize the read depth across samples (Table 2). For the log-transformed data, when the read depths varied from 1,000–20,000, the mean per cent differences between the sample with the deepest and shallowest read depth per iteration were 25.9, 49.9, 37.0, and 42.2 for UQ, CSS, edgeR-TMM, and DESeq-VS, respectively. Further, for every method except proportions and rarefying, there were frequently undesirable correlations between the number of reads and mean BC (Figure 4). This was particularly true for the log-transformed data and for simulations that had a wide range of read depths prior to normalizing.

Similarly, for the comparisons between the BC of the original communities and the BC of the normalized data, proportions had both the tightest correlation and the slope that most closely matched a slope of 1 (Figure 5). The other methods (particularly CSS) had increased levels of variation in the relationship between original and normalized data. All methods performed poorly following the log transformation, resulting in increased variation and slopes that deviated strongly from 1, especially when only the bottom 30% of OTUs varied between the initial distributions.

The PERMANOVAs showed that when the variation in initial read depth was low (5,000–15,000), all methods were roughly equally powerful, prior to the log transformation (rarefied data had a slight loss of power), and their results closely matched the results of the original data (i.e., the real communities; Figure 4). This was true even when only the top 10% or bottom 30% of OTUs varied in the initial distributions. Results were similar when the variation in initial read depth was higher (1,000–20,000);

TABLE 2 Mean (SD) per cent differences between the maximum and minimum read depth per iteration for the simulated data

	Original	Proportions	Rarefied	UQ	CSS	edgeR-TMM	DESeq-VS	Original log	Proportions log	Rarefied log	UQ log	CSS log	edgeR-TMM log	DESeq-VS log
5,000–15,000	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	44.8 (11.5)	47.8 (6.6)	36.9 (7.5)	36.7 (7.0)	18.4 (7.2)	21.3 (7.1)	22.3 (7.4)	12.7 (2.8)	26.7 (4.7)	20.3 (6.9)	22.1 (5.8)
1,000–20,000	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	46.9 (10.6)	72.0 (7.4)	48.9 (8.0)	57.1 (7.7)	17.3 (6.9)	29.0 (7.5)	24.4 (7.4)	25.9 (8.0)	49.9 (7.9)	37.0 (7.8)	42.2 (7.5)

Note. For each iteration, the per cent difference was calculated, and these are the means across iterations. 5,000–15,000 and 1,000–20,000 indicate the range of possible read depths prior to normalization.

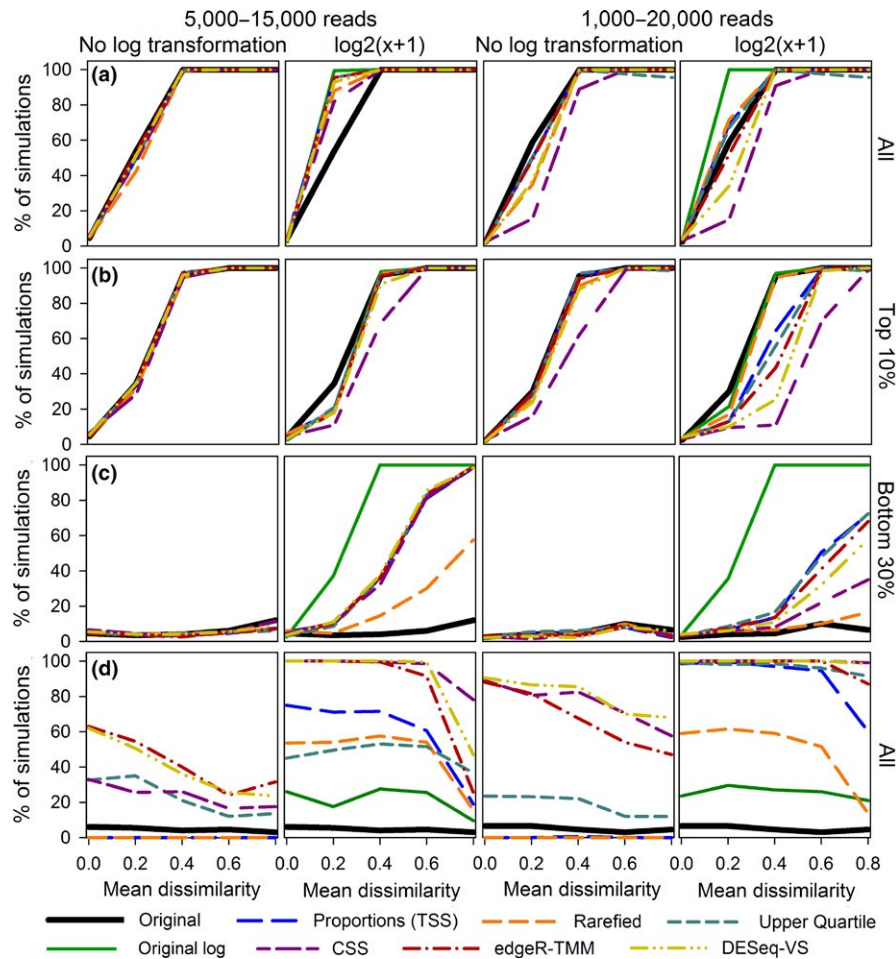


FIGURE 4 Simulation results. (rows a–c) The per cent of iterations (out of 200) where a PERMANOVA returned a significant difference ($\alpha = 0.05$) between the populations. (row d) The per cent of iterations (out of 200) where there was a significant correlation ($\alpha = 0.05$) between read depth and mean Bray–Curtis dissimilarity (mean per individual). These are spurious correlations that indicate a failure of the normalization method. Mean dissimilarity = the setting for the difference between the distributions from which the populations were constructed (0 = identical distributions, 0.8 is highly dissimilar), All = all operational taxonomic units (OTUs) were allowed to vary between the two distributions on which the populations were based, Top 10% = only the OTUs in the 90th percentile and above (based on DNA yield for population 1's distribution) varied between distributions, Bottom 30% = only the OTUs in the 30th percentile and below varied. The thick black "Original" line shows the results for the real communities without a log transformation (even in the $\log_2(x+1)$ columns, where it serves as a point of comparison); whereas the green "Original log" line shows those data following a $\log_2(x+1)$ transformation

however, there was a slight loss of power across methods (particularly for CSS); proportions, UQ, and edgeR-TMM performed the best.

In contrast, when the data were log transformed, they did not closely match the results of the original data (Figure 4). When the variation in read depth was low and all OTUs varied between starting communities, all log-transformed methods had a high rate of false positives compared to the original data (i.e., they detected differences in the communities that were not apparent in the original data). When the variation in read depth was higher, the results were varied and proportions, rarefied data, and UQ had false positives, while CSS, edgeR-TMM, and DESeq-VS had reduced power. When only the top 10% of OTUs varied in the initial distributions, all log-transformed methods had reduced power, and when only the bottom 30% of OTUs varied, all methods had high rates of false positives (except rarefied data when variability in read depth was

high). For the top 10% data, the results were exaggerated when the variation in read depth was high, and for the bottom 30% data, the results were exaggerated when the variation in read depth was low. After log transforming the original data, they showed similar patterns to the normalization methods, but the patterns were often exaggerated.

The PCoAs revealed similar patterns (Figure 6; Supporting Information). All the methods generally performed reasonably well prior to a log transformation (with proportions and rarefied data most closely matching the original communities). Once the data were log transformed, however, the results often differed strongly from the results of the original data. When all OTUs varied in the initial distributions and the mean dissimilarity between the populations was set to a low value (e.g., 0.2), log-transformed data frequently showed clusters that were not evident in the original data. This was

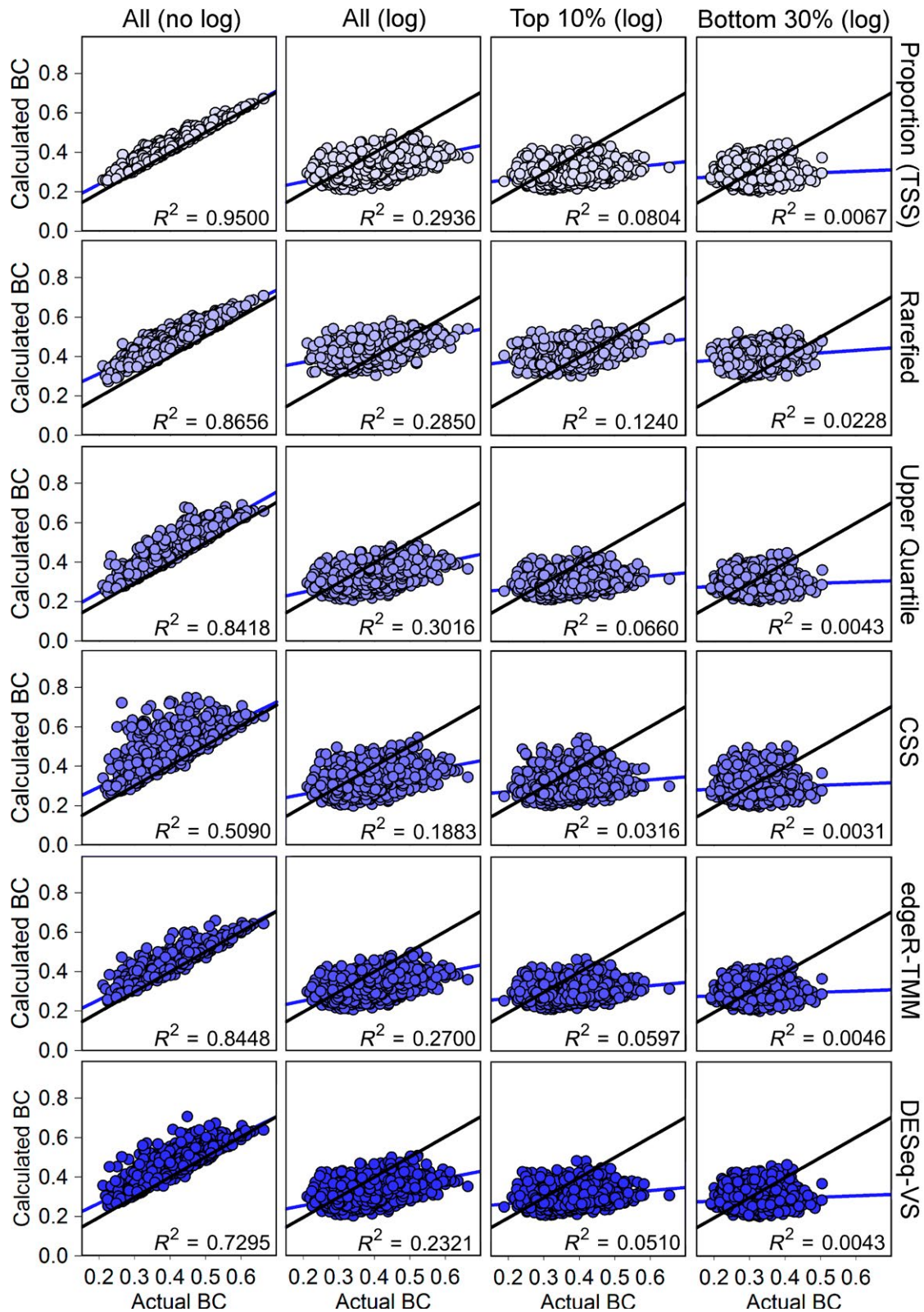


FIGURE 5 Correlations between the Bray–Curtis (BC) dissimilarities for the original communities (“Actual BC”) and the BC dissimilarities following normalization. Black lines show a slope of 1 and intercept of 0. Data are from 200 iterations of the simulator (per column). All = all OTUs were allowed to vary between the two distributions on which the populations were based, Top 10% = only the OTUs in the 90th percentile and above (based on DNA yield for population 1’s distribution) varied between distributions, Bottom 30% = only the OTUs in the 30th percentile and below varied, log = the data were transformed with a $\log_2(x + 1)$ transformation

particularly pronounced when only the bottom 30% of OTUs varied in the initial distributions. In contrast, when only the top 10% of OTUs varied in the initial distributions, log transforming the data often obscured clusters that were apparent in the original data. Additionally, log-transformed ordination plots generally explained less of the variance in the data and frequently clustered most individuals tightly, often with a few distant outliers.

4 | DISCUSSION

The results of both the mouse gut data and simulated data agreed strongly with our predictions, suggesting that methods other than proportions and rarefying distort community-level comparisons. First, with the exception of proportions and rarefying, none of the methods successfully standardized read depth across samples, and those remaining differences in read depths influenced the results, often affecting the BC dissimilarities. This is discouraging, as standardizing read depths are the initial impetus for normalizing the data (i.e., if all samples had equal read depths after sequencing, there would be no need to normalize).

In all analyses, transforming the data to proportions without log transforming returned the most accurate BC dissimilarities compared to the original communities. This is in agreement with previous studies (McMurdie & Holmes, 2014; Weiss et al., 2017) and suggests that, although proportions are not suitable for differential abundance testing (Bullard et al., 2009; Dillies et al., 2013; McMurdie & Holmes, 2014; Weiss et al., 2017), they are the most suitable method for community-level comparisons using dissimilarity and distance measures. Furthermore, proportions produced PCoAs that most closely matched the original data. Rarefied data also performed well, but tended to have more variation than data transformed to proportions. All other methods generally performed well prior to a log transformation, but they had more variation than proportions or rarefied data, suggesting they were still inferior.

Additionally, for all methods, applying a log transformation distorted the BC values, resulting in BC dissimilarities that poorly matched the original values. As a result, the subsequent analyses were strongly influenced by the log transformation. We expected the log transformation to decrease the importance of the most dominant members of the microbial community, while increasing the importance of differences in the rare members, and we observed this in both the PERMANOVAs and PCoAs. This was most clearly illustrated by the comparisons where either only the top 10% of OTUs (i.e., the most abundant OTUs) or the bottom 30% (i.e., the least abundant OTUs) differed between the initial distributions upon which the populations were based. When the initial distributions differed only in the most abundant OTUs, log transforming the data suppressed the differences between populations, resulting in a loss of power to both detect differences among populations and ordinate them into clusters. Conversely, when only the least abundant OTUs varied, the log transformation exaggerated those differences, and both the PERMANOVAs and PCoAs detected differences and

clusters that were not apparent in the original data. Furthermore, because microbial communities typically consist of a few common, and many rare, OTUs, even when all OTUs varied between the initial distributions, log transforming the data often ordinated the data into clusters and produced significant differences between the communities that were not evident in the original data. It should also be stressed that these patterns occurred across log-transformed normalization methods (including log transforming the original data), and the log transformation had a much greater impact on the results than did the choice of normalization method.

Although the loss of power when only common OTUs varied was clearly problematic, for most microbial communities, a log transformation should boost the statistical power, because most communities include many rare OTUs. Whether that boost in statistical power is desirable is, however, debatable. On one hand, because the log transformation detects differences that are not apparent in the original communities, it could be argued that the log transformation results in the detection of exceedingly minor differences that have little ecological relevance. This line of reasoning is especially relevant when you consider the small differences that were often statistically significant following a log transformation (Supporting Information). Indeed, in simulations where only the bottom 30% of OTUs varied in the initial distributions, the BC between the initial communities was only 0.005 on average, and the OTUs in the bottom 30% of initial distributions only varied from 0–5 reads, even when the mean dissimilarity was set to 0.8 (the highest setting we tested). Nevertheless, such slight differences were often statistically significant following the log transformations. On the other hand, because the initial distributions were different (albeit only slightly), it could be argued that the log transformation really is boosting statistical power and allowing the detection of previously obscured trends, rather than detecting inflated differences. Our purpose in this paper was not to give a definitive resolution to the discussion of whether it is beneficial to differentiate communities based on slight differences in rare OTUs, but rather to encourage researchers to think carefully about the ecological questions they are asking when comparing microbial communities.

Nevertheless, some general recommendations are warranted. In most cases, we think that researchers should strive to obtain the most accurate possible representation of the original communities. Thus, given that methods involving a log transformation distort communities and alter species evenness, we argue that community-level comparisons should generally use proportions (preferably) or rarefied data. There are, however, situations in which other normalization methods may be preferable. For example, if the communities in question contain several dominant members (i.e., have low evenness) that are similar across communities, researchers may want to use log-based methods, like CSS, so that differences in the rare members of the communities can be detected. The results should, however, be interpreted within that context, because any detected differences will reflect differences in the rare members of the community, rather than differences in the community as a whole. In other words, when using normalization methods that involve a

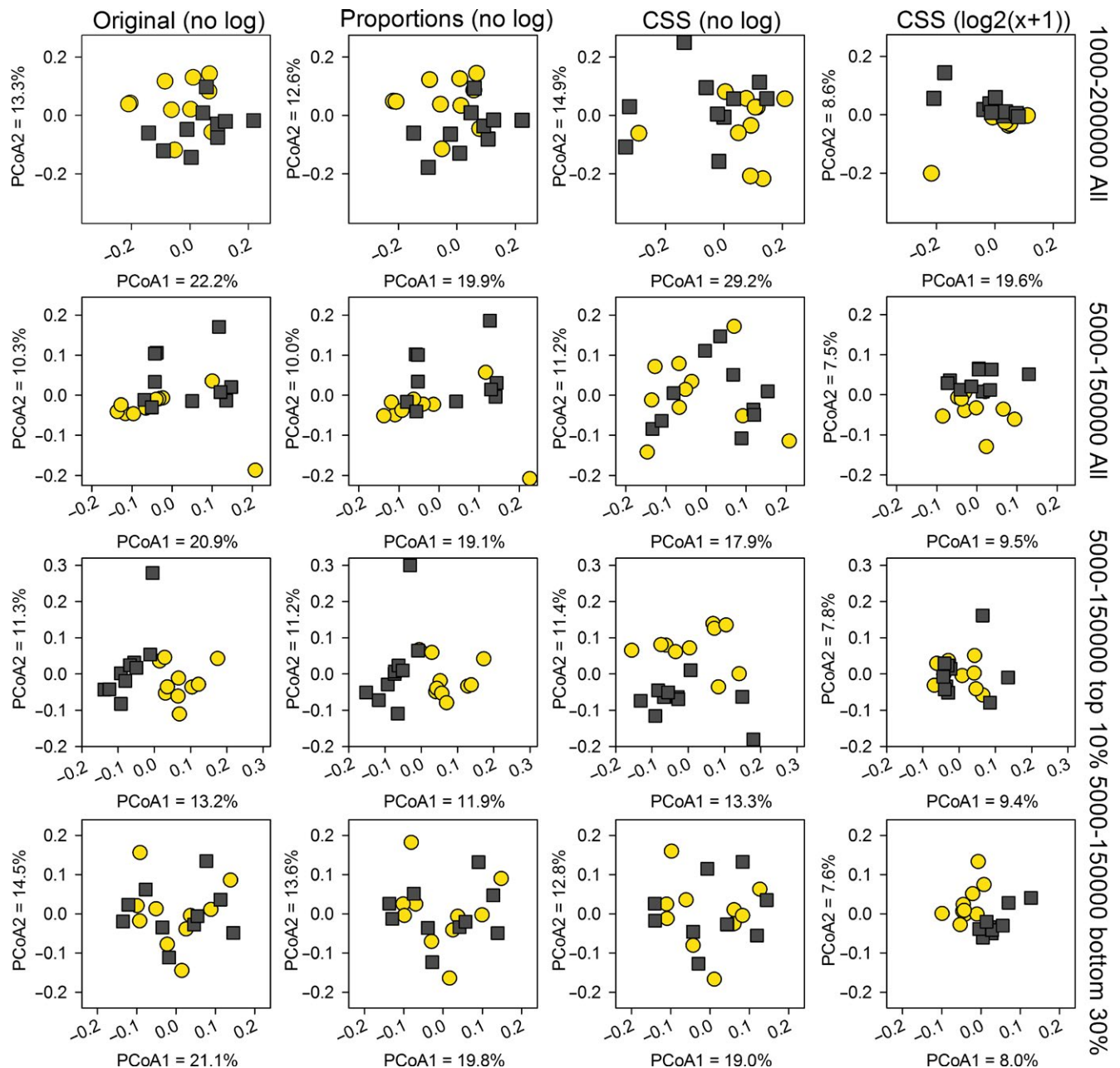


FIGURE 6 Example simulation results of PCoAs comparing population 1 (yellow circles) with population 2 (dark squares) using different normalization methods. Original = the real communities prior to sequencing. Proportions and rarefying generally produced results that were very similar to the original data. Following a log transformation, all methods often produced clusters that were not present in the original data (when all operational taxonomic units (OTUs) or only the bottom 30% varied between the initial distributions) or failed to produce clusters that were present in the original data (when only the top 10% of OTUs varied between the initial distributions). For log-transformed data, only CSS is presented here because of that method's popularity, but other methods involving a log transformation produced similar results (full results are available in Supporting Information). 1,000–20,000 and 5,000–15,000 = the range from which the numbers of reads per sample were randomly selected for each sample, All = all OTUs were allowed to vary between the two distributions on which the populations were based, Top 10% = only the OTUs in the 90th percentile and above (based on DNA yield for population 1's distribution) varied between distributions, Bottom 30% = only the OTUs in the 30th percentile and below varied. For rows 1 and 2, the mean dissimilarity was set to 0.2, for row 3 it was 0.3, and for row 4 it was 0.8

log transformation, it would be incorrect to say that the communities as a whole differ, and it would be more accurate to state that uncommon members of the community differ after reducing the importance of the common members. Conversely, if a significant

difference is not detected when using log-based methods, it would be misleading to say that the communities are not different, because log-based methods suppress differences in abundant OTUs and can mask differences between communities.

5 | CONCLUSIONS AND RECOMMENDATIONS

Both rarefied data and, especially, proportions outperformed all the other normalization methods for producing accurate BC dissimilarities and subsequent PCoAs and PERMANOVAs. They were the only methods that were capable of truly standardizing read depths, and they avoided the spurious correlations that were produced by the other methods. Therefore, although previous studies have raised serious concerns over their applicability for differential abundance testing, we do not think that they should be dismissed for community-level comparisons.

Further, although log transformations are a standard component of many normalization procedures, we showed that they can often distort comparisons of communities by suppressing large differences in common OTUs and amplifying slight differences in rare OTUs. In cases when populations of samples differ only in the most abundant OTUs, log transformations make the populations artificially similar and can mask differences. Conversely, when there are many rare OTUs, as is often the case in microbial communities, they can reveal differences that are not otherwise detectable. Whether that trait is a desirable boost in power or an undesirable false positive will depend on the specific ecological questions being asked. We are not, therefore, making blanket recommendations one way or the other, but simply want to encourage researchers and readers to carefully consider the ecology of their communities, the specific questions they are asking, and whether a given normalization method is suitable for addressing those questions.

CONFLICTS OF INTEREST

All the authors affirm that they have no conflicts of interest to declare.

AUTHORS' CONTRIBUTIONS

D.T.M. wrote the simulations and led the project design, analysis, and writing. R.H., R.A.A., L.S., D.S.B., and K.R.Z. supervised the research, assisted with project design and analysis, and edited the manuscript.

DATA ACCESSIBILITY

R scripts and the distribution used for our simulations are available via Dryad <https://doi.org/10.5061/dryad.tn8qs35>. The mouse microbiome dataset is available in the R package METAGENOMESeq <https://doi.org/10.18129/b9.bioc.metagenomeseq> (McKnight et al., 2018).

ORCID

Donald T. McKnight  <https://orcid.org/0000-0001-8543-098X>

Roger Huerlimann  <https://orcid.org/0000-0002-6020-334X>

REFERENCES

- Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2009). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1), 94. <https://doi.org/10.1186/1471-2105-11-94>
- Costea, P. I., Zeller, G., Sunagawa, S., & Bork, P. (2014). A fair comparison. *Nature Methods*, 11(4), 359. <https://doi.org/10.1038/nmeth.2898>
- Cottrell, M. T., & Kirchman, D. L. (2003). Contribution of major bacterial groups to bacterial biomass production (thymidine and leucine incorporation) in the Delaware estuary. *Limnology and Oceanography*, 48(1), 168–178. <https://doi.org/10.4319/lo.2003.48.1.0168>
- Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., & Jaffrézic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6), 671–683. <https://doi.org/10.1093/bib/bbs046>
- Fuhrman, J. A. (2009). Microbial community structure and its functional implications. *Nature*, 459(7244), 193–199. <https://doi.org/10.1038/nature08058>
- Ghazoul, J. (2006). Floral diversity and the facilitation of pollination. *Journal of Ecology*, 94, 295–304. <https://doi.org/10.1111/j.1365-2745.2006.01098.x>
- Hillebrand, H., Bennett, D. M., & Cadotte, M. W. (2008). Consequences of dominance: A review of evenness effects on local and regional ecosystem processes. *Ecology*, 89, 1510–1520. <https://doi.org/10.1890/07-1053.1>
- Hillebrand, H., & Cardinale, B. J. (2004). Consumer effects decline with prey diversity. *Ecology Letters*, 7, 192–201. <https://doi.org/10.1111/j.1461-0248.2004.00570.x>
- Hillebrand, H., Gruner, D. S., Borer, E. T., Bracken, M. E. S., Cleland, E. E., Elser, J. J., & Smith, J. E. (2007). Consumer versus resource control of producer diversity depends on ecosystem type and producer community structure. *Proceedings of the National Academy of Sciences*, 104(26), 10904–10909. <https://doi.org/10.1073/pnas.0701918104>
- Jackson, D. A. (1997). Compositional data in community ecology: The paradigm or peril of proportions. *Ecology*, 78(3), 929–940. [https://doi.org/10.1890/0012-9658\(1997\)078\[0929:CDICET\]2.0.CO;2](https://doi.org/10.1890/0012-9658(1997)078[0929:CDICET]2.0.CO;2)
- Lin, Y., Golovnina, K., Chen, Z.-X., Lee, H. N., Negron, Y. L. S., Sultana, H., & Harbison, S. T. (2016). Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*. *BMC Genomics*, 17(1), 28. <https://doi.org/10.1186/s12864-015-2353-z>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40, 4288–4297. <https://doi.org/10.1093/nar/gks042>
- McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (2018). Data from: Methods for normalizing microbiome data: An ecological perspective. *Dryad Digital Repository*, <https://doi.org/10.1111/2041-210X.13115>
- McMurdie, P. J., & Holmes, S. (2013). PHYLOSEQ: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, 8, e61217. <https://doi.org/10.1371/journal.pone.0061217>
- McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, 10(4), e1003531. <https://doi.org/10.1371/journal.pcbi.1003531>

- Oksanen, J. F., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGinn, D., ... Wagner, H. (2017). *vegan*: Community ecology package. Retrieved from <https://cran.r-project.org/package=vegan>
- Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Robust methods for differential abundance analysis in marker gene surveys. *Nature Methods*, 10(12), 1200–1202. <https://doi.org/10.1038/nmeth.2658>
- Pedros-Alí, C. (2006). Marine microbial diversity: Can it be determined? *Trends in Microbiology*, 14, 257–263. <https://doi.org/10.1016/j.tim.2006.04.007>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). *edgeR*: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Stirling, G., & Wilsey, B. (2001). Empirical relationships between species richness, evenness, and proportional diversity. *The American Naturalist*, 158, 286–299. <https://doi.org/10.1086/321317>
- Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., & Gordon, J. I. (2009). The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine*, 1(39857), 1–19. <https://doi.org/10.1126/scitranslmed.3000322>
- Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1), 27. <https://doi.org/10.1186/s40168-017-0237-y>
- Wilsey, B. J., Chalcraft, D. R., Bowles, C. M., & Willig, M. R. (2005). Relationships among indices suggest that richness is an incomplete surrogate for grassland biodiversity. *Ecology*, 86, 1178–1184. <https://doi.org/10.1890/04-0394>
- Wilsey, B. J., & Polley, H. W. (2002). Reductions in grassland species evenness increase dicot seedling invasion and spittle bug infestation. *Ecology Letters*, 5, 676–684. <https://doi.org/10.1046/j.1461-0248.2002.00372.x>
- Wittebolle, L., Marzorati, M., Clement, L., Balloi, A., Daffonchio, D., Heylen, K., & Boon, N. (2009). Initial community evenness favours functionality under selective stress. *Nature*, 458(7238), 623–626. <https://doi.org/10.1038/nature07840>
- Zhang, Y., Jiao, N., Cottrell, M. T., & Kirchman, D. L. (2006). Contribution of major bacterial groups to bacterial biomass production along a salinity gradient in the South China Sea. *Aquatic Microbial Ecology*, 43, 233–241. <https://doi.org/10.3354/ame043233>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. Methods for normalizing microbiome data: An ecological perspective. *Methods Ecol Evol.* 2019;10:389–400. <https://doi.org/10.1111/2041-210X.13115>