Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data

Sophie J Weiss, Zhenjiang Xu, Amnon Amir, Shyamal Peddada, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vazquez-Baeza, Amanda Birmingham, Rob Knight

Background: Data from 16S amplicon sequencing present challenges to ecological and statistical interpretation. In particular, library sizes often vary over several ranges of magnitude, and the data contains many zeroes. Also, since researchers sample a small fraction of the ecosystem, the observed sequences are relative abundances and therefore the data is compositional. Here we evaluate methods developed in the literature to address these three challenges in the context of normalization and ordination analysis, which is commonly used to visualize overall differences in bacterial composition between sample groups, and differential abundance analysis, which tests for significant differences in the abundances of microbes between sample groups. **Results.** *Effects of normalization on ordination:* Most normalization methods successfully cluster samples according to biological origin when many microbes differ between the groups. For datasets in which clusters are subtle and/or sequence depth varies greatly between samples, or for metrics in which rare microbes play an important role, rarefying outperforms other techniques. For abundance-based metrics, rarefying as well as alternatives like DESeq and metagenomeSeq's cumulative sum scaling (CSS), seem to correctly cluster samples according to biological origin. With these normalization alternatives, clustering by sequence depth as a confounding variable must be checked for, especially for low library sizes. *Effects of differential abundance testing model choice:* We build on previous work to evaluate each statistical method using rarefied as well as unrarefied data. When the mean library sizes in the differential abundance groups differ by more than 2-3x, or the library sizes differ in distribution, our simulation studies reveal that each statistical method improved in its false positive rate when samples were rarefied. However, when the difference in library size mean is less than 2-3x, and the library sizes are similarly distributed, rarefying results in a loss of power for all methods. In this case, DESeq2 has the highest power to compare groups, especially for less than 20 samples per group. MetagenomeSeq's fitZIG is a faster alternative to DESeq2, although it does worse for smaller sample sizes (<50 samples per group) and tends to have a higher false positives rate. For larger sample sizes (>50 samples), rarefying paired with a non-parametric test, such as the Mann-Whitney test, can also yield equally high sensitivity. Based on these

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1157v1 | CC-BY 4.0 Open Access | rec: 3 Jun 2015, publ: 3 Jun 2015

results, we recommend a stepwise procedure in which sample groups are first tested for significant differences in library size. If there is a significant difference, we recommend rarefying with a non-parametric test. Otherwise, DESeq2 and/or fitZIG offer increased sensitivity, especially for rare OTUs and small sample numbers. **Conclusions.** These findings help guide which technique to use depending on the data characteristics of a given study.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1157v1 | CC-BY 4.0 Open Access | rec: 3 Jun 2015, publ: 3 Jun 2015

Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data

Sophie J. Weiss[1], Zhenjiang Zech Xu[2], Amnon Amir[2], Shyamal Peddada[3], Kyle Bittinger[4], Antonio Gonzalez[2], Catherine Lozupone[5], Jesse R. Zaneveld[6], Yoshiki Vázquez-Baeza[2], Amanda Birmingham[7], Rob Knight[2,8a]

[1]Department of Chemical and Biological Engineering, University of Colorado at Boulder, Boulder, CO 80309
[2]Departments of Pediatrics, University of California San Diego, La Jolla, CA 92093
[3]Biostatistics and Computational Biology Branch, NIEHS, NIH
[4]Department of Microbiology, University of Pennsylvania, Philadelphia, PA 18014
[5]Department of Medicine, University of Colorado, Denver 80204
[6] Department of Microbiology, Oregon State University, 226 Nash Hall, Corvallis, OR 97331
[7]Center for Computational Biology and Bioinformatics, Dept. of Medicine, University of California San Diego, La Jolla, CA 92093
[8]Department of Computer Science & Engineering, University of California San Diego, La Jolla, CA 92093

[a]To whom correspondence should be addressed

Corresponding author
Rob Knight, University of California San Diego, 9500 Gilman Drive, MC 0763 La Jolla, CA 92093
robknight@ucsd.edu t:858-246-1184 f:858-246-1981

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1157v1 | CC-BY 4.0 Open Access | rec: 3 Jun 2015, publ: 3 Jun 2015

ABSTRACT

Background: Data from 16S amplicon sequencing present challenges to ecological and statistical interpretation. In particular, library sizes often vary over several ranges of magnitude, and the data contains many zeroes. Also, since researchers sample a small fraction of the ecosystem, the observed sequences are relative abundances and therefore the data is compositional. Here we evaluate methods developed in the literature to address these three challenges in the context of normalization and ordination analysis, which is commonly used to visualize overall differences in bacterial composition between sample groups, and differential abundance analysis, which tests for significant differences in the abundances of microbes between sample groups.

Results. *Effects of normalization on ordination:* Most normalization methods successfully cluster samples according to biological origin when many microbes differ between the groups. For datasets in which clusters are subtle and/or sequence depth varies greatly between samples, or for metrics in which rare microbes play an important role, rarefying outperforms other techniques. For abundance-based metrics, rarefying as well as alternatives like DESeq and metagenomeSeq's cumulative sum scaling (CSS), seem to correctly cluster samples according to biological origin. With these normalization alternatives, clustering by sequence depth as a confounding variable must be checked for, especially for low library sizes. *Effects of differential abundance testing model choice:* We build on previous work to evaluate each statistical method using rarefied as well as unrarefied data. When the mean library sizes in the differential abundance groups differ by more than 2-3x, or the library sizes differ in distribution, our simulation studies reveal that each statistical method improved in its false positive rate when samples were rarefied. However, when the difference in library size mean is less than 2-3x, and the library sizes are similarly distributed, rarefying results in a loss of power for all methods. In this case, DESeq2 has the highest power to compare groups, especially for less than 20 samples per group. MetagenomeSeq's fitZIG is a faster alternative to DESeq2, although it does worse for smaller sample sizes (<50 samples per group) and tends to have a higher false positives rate. For larger sample sizes (>50 samples), rarefying paired with a non-parametric test, such as the Mann-Whitney test, can also yield equally high sensitivity. Based on these results, we recommend a stepwise procedure in which sample groups are first tested for significant differences in library size. If there is a significant difference, we recommend rarefying with a non-parametric test. Otherwise, DESeq2 and/or fitZIG offer increased sensitivity, especially for rare OTUs and small sample numbers.

Conclusions. These findings help guide which technique to use, depending on the data characteristics of a given study.

INTRODUCTION

Although data produced by high-throughput sequencing has proven extremely useful for understanding microbial communities, the interpretation of these data is complicated by several statistical challenges. To ease data interpretation, data are often normalized to account for the sampling process and differences in sequencing efforts. Ordination analysis, such as principal coordinates analysis (PCoA) (Gower 1966), is subsequently applied to these normalized data to visualize broad trends of how similar or different bacteria are in certain sample types, such as healthy *vs*. sick patients). Samples containing similar bacteria will group, or cluster, close together, while differences in bacterial composition will cause separation in PCoA space. Next,

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1157v1 | CC-BY 4.0 Open Access | rec: 3 Jun 2015, publ: 3 Jun 2015

researchers may wish to determine, through statistical testing, which specific bacteria are significantly differentially abundant between two sample type clusters.

For example, patients with *Clostridium difficile* infection cluster separately from healthy patients in PCoA plots, and these overall differences in community composition are driven by differences in microbial relative abundances (Kelly et al. 2014; Shankar et al. 2014; Weingarden et al. 2015). Restoration of each intestinal bacteria type to healthy levels leads to patient recovery, and causes samples from treated patients to overlap with healthy individuals in PCoA plots. Significant changes in certain bacterial species abundances has also been linked to inflammatory bowel diseases (Gevers et al. 2014), diarrhea (Pop et al. 2014), obesity (Ley et al. 2005; Ridaura et al. 2013; Turnbaugh et al. 2009), HIV (Lozupone et al. 2013a), diet (David et al. 2014), culture, age, and antibiotic use (Lozupone et al. 2013b), among many other factors. However, the veracity of these discoveries depends upon how well the chosen normalization and differential abundance testing techniques address the statistical challenges posed by the underlying community sequence data.

Following initial quality control steps to account for errors in the sequencing process, microbial community sequencing data is typically organized into large matrices where the columns represent samples, and rows contain observed counts of clustered sequences commonly known as Operational Taxonomic Units, or OTUs, that represent bacteria types. These tables are often referred to as OTU tables. Several features of OTU tables can cause erroneous results in downstream analyses if unaddressed. First, the microbial community in each biological sample may be represented by very different numbers of sequences, reflecting differential efficiency of the sequencing process rather than true biological variation. This problem is exacerbated by the observation that the full range of species is rarely saturated, such that more bacterial species are observed with more sequencing. (Similar trends by sequencing depth hold for discovery of genes in shotgun metagenomic samples (Qin et al. 2010; Rodriguez & Konstantinidis 2014)). Thus, samples with relatively few sequences can have inflated beta (β, or between sample) diversity, because authentically shared OTUs are erroneously scored as unique to samples with more sequences (Lozupone et al. 2011). Second, most OTU tables are sparse, meaning that they contain a high proportion of zero counts (Paulson et al. 2013). This sparsity means that the counts of rare OTUs are uncertain, since they are at the limit of sequencing detection ability when there are many sequences per sample (i.e. large library size), and are undetectable when there are few sequences per sample. Third, each sample is only a small percentage of its original environment, constraining the total number of rRNA sequences to a constant sum; in such "compositional" data, researchers do not know the absolute counts of each type of OTU but only their relative abundances in relation to each other (Aitchison 1982; Friedman & Alm 2012; Lovell D 2010). Uneven sampling depth, sparsity, and compositionality represent serious challenges for interpreting these data. No normalization method or differential abundance testing method simultaneously addresses all of these challenges. Thus, investigators must choose methods based on relevant features of the dataset under consideration.

Normalization

Normalization is critical to address variability in sampling depths and number of zeros. Microbial ecologists in the era of high-throughput sequencing have commonly normalized their OTU matrices by rarefying, or drawing without replacement from each sample such that all samples have the same number of total counts. Samples with total counts below the defined threshold are excluded, sometimes leading researchers to face difficult trade-offs between sampling depth and the number of samples evaluated. To ensure the proper total sum is chosen, rarefaction curves can be constructed (Gotelli & Colwell 2001). These curves plot the number of counts sampled (rarefaction depth) vs. the expected value of species diversity. Rarefaction curves provide guidance that allows users to avoid negatively impacting the species diversity found in samples by choosing too low a rarefaction depth. The origins of rarefying sample counts are mainly in sample species diversity measures, or alpha diversity (Brewer & Williamson 1994; Gotelli & Colwell 2001). However, more recently rarefying has been used in the context of β-diversity (Horner-Devine et al. 2004; Jernvall & Wright 1998). Rarefying samples for normalization is now the standard in microbial ecology, and is present in all major data analysis toolkits for this field (Caporaso et al. 2010; Jari Oksanen 2015; McMurdie & Holmes 2013; Schloss et al. 2009). While rarefying is not an ideal normalization method, as it reduces statistical power by removing some data, and was not designed to address compositionality, alternatives to rarefying have not been sufficiently developed until recently.

Normalization alternatives to rarefying all involve some type of transformation, the most common of which are scaling or log-ratio transformations. Effects of scaling methods depend on the scaling factor chosen; often, a particular quantile of the data is used for normalization, but choosing the correct quantile is difficult (Anders & Huber 2010; Bullard et al. 2010; Dillies et al. 2013; Paulson et al. 2013; Robinson & Oshlack 2010), and scaling can overestimate or underestimate the prevalence of zero fractions, depending on whether zeroes are left in or thrown out of the scaling (Agresti & Hitchcock ; Friedman & Alm 2012). This is because putting all samples of varying sampling depth on the same scale ignores the differences in sequencing depth, and therefore resolution of species, between the samples. For example, a rare species having zero counts in a small rRNA sample can have a small fractional abundance in a large rRNA sample (unless further mathematical modeling beyond simple proportions is applied to correct for this). Scaling can also distort OTU correlations across samples, again due to zeroes, differences in sequencing depth, and sum constraints (Aitchison 1982; Buccianti et al. 2006; Friedman & Alm 2012; Lovell D 2010; Pearson 1896).

While rarefying and some scaling techniques, such as total sum scaling (proportions), treat OTU sequence counts as absolute environmental abundances, the counts are compositional and only a fraction from the original environment, making only their relative ratios known (Friedman & Alm 2012; Lovell D 2010). In contrast, log ratio transformations correct for compositionality by exploiting this relative ratio information, and can also alleviate some noise in the data (Aitchison 1982; Buccianti et al. 2006; Friedman & Alm 2012; Lovell D 2010). However, because the log transformation cannot be applied to zeros (which are often well over half of microbial data counts (Paulson et al. 2013)), sparsity is extremely problematic for methods that rely on this transformation. One approach to this issue is to replace zeros with a small value, known as a pseudocount. Despite active research on selection of pseudocount values for scaling methods (Egozcue et al. 2003; Greenacre 2011), the choice of pseudocount values can dramatically change the results (Costea et al. 2014; Paulson et al. 2014).

Differential Abundance Testing

For OTU differential abundance testing between conditions (e.g. case vs. control), a common approach is to first rarify the count matrix to a fixed depth and then apply a non-parametric test (e.g. Mann-Whitney test for tests of two classes; Kruskal-Wallis test for tests of multiple groups). Non-parametric tests are often preferred because most OTU counts are not normally distributed (Wagner et al. 2011). However, this approach does not account for the fact that the OTU counts are compositional. Also, nonparametric tests such as the Kruskal-Wallis test do not fare well in terms of power when the data are sparse, but perform well when the data are not sparse (Paulson et al. 2013). Recently, promising parametric models that make stronger assumptions about the data have been developed in the subfields of transcriptomics ('RNA-Seq') and metagenomic sequencing. These may additionally be useful for microbial marker gene data (Anders & Huber 2010; Anders et al. 2013; Law et al. 2014; Love MI 2014; McMurdie & Holmes 2014; Paulson et al. 2013; Robinson et al. 2010; Robinson & Smyth 2008). Such models have greater detection power if their assumptions about the data are correct; however, studies of these models on RNA-Seq data have shown that they can yield poor results (Rapaport et al. 2013) if relevant assumptions are not valid.

These parametric models are composed of a generalized linear model (GLM) that assumes a distribution (Cameron & Trivedi), and there is debate about which distribution to use (Auer & Doerge 2010; Cheung 2002; Connolly et al. 2009; Holmes et al. 2012; McMurdie & Holmes 2014; Paulson et al. 2013; Rapaport et al. 2013; Soneson & Delorenzi 2013; White et al. 2009; Yu et al. 2013). In the genomics field, the negative binomial (NB) GLM has replaced the Poisson GLM to allow for estimating overdispersion (Anders & Huber 2010; Anders et al. 2013; Robinson et al. 2010). This model type was also one of the first in the RNA-Seq field, and developed for use with a low number of replicates. NB models accommodate low replication by assuming that OTUs of similar mean expression strength have similar variance in their sample count distributions, estimating model parameters using this assumption, and then leveraging the GLM to perform exact statistical tests. Also, while allowing for some overdispersion, the NB often yields a poor fit in the case of a large number of zeroes, which is very typical in microbiome data (Cheung 2002; Paulson et al. 2013). Zero-inflated GLMs, the most promising of which is the zero-inflated Gaussian (ZIG), attempts to overcome this limitation (Paulson et al. 2013). The ZIG tries to address compositionality, sparsity and unequal sampling depth by separately modeling 'structural' zero counts generated by e.g. under-sequencing and zeros generated by the biological distribution of taxa. Log transformation of the non-zero counts yields the Gaussian. However, this mixture model distribution is designed for continuous data rather than discrete microbiome data. Hence, it is expected to do best in study designs that have large sample sizes and high sequencing depths, and thus best approximate continuous distributions.

Here, we evaluate some of the most widely used or promising techniques for analyzing sequencing data in the context of microbial ecology, with a focus on normalization and OTU differential abundance testing. In addition to these widely used or promising methods, we also test the naïve approaches of no normalization, and proportions (i.e. total sum scaling) for comparison purposes. Such comparisons are important, because while potential issues with many methodologies are known, the balance of sensitivity and specificity for these methods in situations commonly facing microbial ecologists is currently largely unknown. Recent work in this area (McMurdie & Holmes 2014), provides insights into the performance of parametric normalization and differential abundance testing approaches for microbial ecology studies.

However, the work is primarily focused on estimating proportions from discrete data. Here we update and expand these recent findings using both real and simulated datasets exemplifying the additional combined challenges of uneven library sizes, sparsity, and compositionality.

MATERIALS AND METHODS

Normalization

The basic test of how well broad differences in microbial sample composition are detected, as assessed by clustering analysis, was conducted as in ‘Simulation A’ from McMurdie and Holmes (McMurdie & Holmes 2014). Briefly, the ‘Ocean’ and ‘Feces’ microbiomes (the microbial data from ocean and human feces samples, respectively) from the ‘Global Patterns’ dataset (Caporaso et al. 2011b) were used as templates, modeled with a multinomial, and taken to represent distinct classes of microbial community because they have few OTUs in common. These two classes were mixed in many defined proportions (the ‘effect size’) in independent simulations in order to generate simulated samples of varying clustering difficulty. Samples were generated in sets of 40, as in McMurdie and Holmes (McMurdie & Holmes 2014). We also tested smaller and larger sample sizes but saw little difference in downstream results. Additional sets of 40 samples were simulated for varying library sizes (1000, 2000, 5000, and 10000 sequences per sample). These simulated samples were then used to assess normalization methods by the proportion of samples correctly classified into the two clusters by the partitioning around medioids (PAM) algorithm (Kaufman L. 1990; Reynolds A 2006).

McMurdie and Holmes (McMurdie & Holmes 2014) evaluated clustering accuracy with five normalization methods (none, proportion, rarefying with replacement as in the multinomial model (Colwell et al. 2012), DESeqVS (Anders & Huber 2010), and UQ-logFC (in the edgeR package) (Robinson et al. 2010)) and six beta diversity metrics (Euclidean, Bray-Curtis (Bray & Curtis 1957), PoissonDist (Witten 2011), top-MSD (Robinson et al. 2010), unweighed UniFrac (Lozupone & Knight 2005), and weighted UniFrac (Lozupone et al. 2007)). We modified the normalization methods to those in Table S1 (none, proportion, rarefying without replacement as in the hypergeometric model (Colwell et al. 2012), CSS (Paulson et al. 2013), logUQ (Bullard et al. 2010), DESeqVS (Anders & Huber 2010), and edgeR-TMM (Robinson & Oshlack 2010)) and the beta diversity metrics to those in Fig2 and Fig. S1 (binary Jaccard, Bray-Curtis (Bray & Curtis 1957), Euclidean, unweighed UniFrac (Lozupone & Knight 2005), and weighted UniFrac (Lozupone et al. 2007)), thus including more recent normalization methods (Bullard et al. 2010; Paulson et al. 2013), and only those beta diversity metrics that are most common in the literature. We amended the rarefying method to the hypergeometric model (Colwell et al. 2012), which is much more common in microbiome studies (Caporaso et al. 2010; Schloss et al. 2009). Negatives in the DESeq normalized values (Anders & Huber 2010) were set to zero as in McMurdie and Holmes (McMurdie & Holmes 2014), and a pseudocount of one was added to the count tables (McMurdie & Holmes 2014). McMurdie and Holmes (McMurdie & Holmes 2014) penalized the rarefying technique for dropping the lowest fifteenth percentile of sample library sizes in their simulations by counting the dropped samples as ‘incorrectly clustered’. Because the 15th percentile was used to set rarefaction depth, this capped clustering accuracy at 85%. We instead quantified cluster accuracy among samples that were clustered following normalization to exclude this rarefying penalty (Fig. S1). Conversely, it has since been confirmed that low-depth samples contain a higher proportion of contaminants (rRNA not from the intended sample) (Kennedy et al. 2014; Salter et al. 2014). Because the higher depth samples that rarefying keeps may be higher quality and therefore give rarefying an unfair advantage, Fig. 2 compares clustering accuracy for all the techniques based on the same set of samples remaining in the rarefied dataset.

On the real datasets, non-parametric multivariate ANOVA (PERMANOVA) (Anderson 2001) was calculated by fitting a Type I sequential sums of squares model (y ~ Library_Size + Biological_Effect). Thus, we control for library size differences before assessing the effects on the studied biological effect. All data was retrieved from QIITA (https://qiita.microbio.me).

Differential Abundance Testing

The simulation test for how well truly differentially abundant OTUs are recognized by various parametric and non-parametric tests was conducted as in 'Simulation B' in McMurdie and Holmes (McMurdie & Holmes 2014), with a few changes. The basic data generation model remained the same, but the creation of 'true positive' OTUs was either made symmetrical through duplication or moved to a different step, to avoid introducing compositionality artifacts (see below) depending on the simulation. The 'Global Patterns' (Caporaso et al. 2011b) dataset was again used, because it was one of the first studies to apply high-thoughput sequencing to a broad range of environments, which includes 9 environment types from 'Ocean', to 'Soil'; all simulations were evaluated for all environments. Additionally, we verified the results on the 'Lean' and 'Obese' microbiomes from a different study (Piombino et al. 2014). As in McMurdie and Holmes, significant changes were controlled for multiple comparisons using the Benjamini & Hochberg (Benjamini & Hochberg 1995) False Discovery Rate (FDR) threshold of 0.05.

A simple overview of the two methods used for simulating differential abundance is presented in Fig. S5a. In McMurdie and Holmes' (McMurdie & Holmes 2014) 'Original' simulation (second row), the distribution of counts from one environment (e.g. 'Ocean') was modeled off of a multinomial template (first row) for two similar groups ('Ocean_1' and 'Ocean_2'), ensuring a baseline of all 'true negative' OTUs. Following the artificial inflation of specific OTUs in the 'Ocean_1' samples to create 'true positives', fold-change estimates for every other OTU are affected. Thus, 'true negatives' are possible 'true positives.' This is because the counts in an OTU table are compositional, or relative abundances constrained to a sum. To control for this we inflate OTUs by pairs of differentially abundant OTUs in both the 'Ocean_1' and 'Ocean_2' samples (third row), creating a new 'Balanced' simulation.

We also tested the effect of differentially abundant organisms dominating one type of community by drawing from a multinomial distribution where solely that organism's template value is increased. This 'Compositional' approach is explained in Fig. S5b, and the results are shown in Fig. S7. In Fig S7, the environmental abundances of 25% of the OTUs in one group are increased.

Besides the above procedural changes to the McMurdie and Holmes (McMurdie & Holmes 2014) simulation, we also modified the rarefying technique from sampling with replacement (multinomial) to sampling without replacement (hypergeometric - as in the previous Normalization simulations) (Colwell et al. 2012). The testing technique was modified from a two-sided Welch t-test to non-parametric Mann-Whitney test, which is widely used and more appropriate because the OTU distributions in microbiome data usually deviate from normality. The techniques used (Table S2) differ only by the addition of another RNA-Seq method, Voom (Law et al. 2014). Finally, we corrected the FPR definition (McMurdie & Holmes 2014) from FP/(TP + FP) to FP/(TN + FP), where FP = number of false positive OTUs, TP = number of true positive OTUs, and TN = number of true negative OTUs. This new simulation code can be found in the supplemental R files (Differential_abundance.R, and Differential_abundance_with_compositionality.R).

Power Curve Calculations

Similar to Table S1 in McMurdie and Holmes [27], we considered a very simplistic set-up to evaluate the effect of rarefying on power when comparing two groups, labeled A and B. As in McMurdie and Holmes [27], we considered the extreme case of a microbial population consisting of only 2 species (or 2 OTUs), with OTU1 + OTU2 = library size. For power calculations, we assumed that the amount of OTU1 in group B is 85% of the amount of OTU1 in group A. Thus, it is enough to quantify the proportion of OTU1 in group A and library sizes of groups A and B to specify the whole system.

We considered varied patterns of proportions of OTU1 in group A ranging from very rare to common (0.5% to 50%). The library size of group A was fixed at either 500, 1000 or 10,000 sequences per sample. Meanwhile, the library size of group B was always taken to be at least as large as that of group A and was either 10,000 or 100,000 sequences per sample. Various rarefied percentages of the group B library size were considered. The percent-rarefied calculation for the first set of power curves is exemplified below using a library size of 500 for library A, and an unrarefied library size of 10,000 for B:

Library size for A	Library size for B
500	10,0000 (unrarefied case)
500	5,000 (50% rarefied)
500	1,000 (90% rarefied)
500	500 (95% rarefied)

For each scenario of proportion of OTU1 and library sizes, power was computed using Fisher's exact test. Power calculations were done using the statistical software SAS. Power calculation results are provided in Fig. 5.

Software Package Versions

R version 3.1.0 (Team 2014) was used with Bioconductor (Gentleman et al. 2004) packages phyloseq version 1.10.0, DESeq version 1.16.0, DESeq2 version 1.4.5, edgeR version 3.6.8, metagenomeSeq version 1.7.31, and Limma version 3.20.9. Also, we used python-based QIIME version 1.9.0, with Emperor (Vazquez-Baeza et al. 2013).

RESULTS AND DISCUSSION

Normalization

When there is a strong biological signal, and normalization is done properly, PCoA can yield clear clustering and insight into microbial community differences (Fig. 1a). However, low-depth samples can lead to poor cluster resolution (Fig. 1b), both by reducing information on community structure, and by being more readily influenced by contamination (Kennedy et al. 2014; Salter et al. 2014). Furthermore, if no data normalization is applied, or the normalization method fails to properly correct for differences in sequencing efficacy, the original library size of the samples can confound biological differences (Fig. 1c). This is because samples of lower sequencing depth fail to detect rare taxa. Highly sequenced samples will thus appear more similar to each other than to shallow sequenced samples because they are scored as sharing the same rare taxa.

To assess all the normalization methods (Table S1), we conducted simulations in the context of results that are highly critical of the rarefying technique (McMurdie & Holmes 2014). Briefly, only necessary modifications (Methods) were made to the code of McMurdie and Holmes (McMurdie & Holmes 2014), making our approach easily comparable. If rarefying is not penalized for the fifteenth percentile lowest depth samples that are thrown out, it can do better than other techniques (Fig. S1). This practice of removing low depth samples from the analysis is supported by the recent discovery that small biomass samples are of poorer quality and may contain contaminating sequences (Kennedy et al. 2014; Salter et al. 2014). Furthermore, alternatives to rarefying also recommend discarding low-depth samples, especially if they cluster separately from the rest of the data (Love MI 2014; Paulson et al. 2013). If all other techniques are run only on the same samples as rarefying, rarefying still does well (Fig. 2). These results demonstrate that previous microbiome ordinations using rarefying as a normalization method likely drew correct conclusions, even if some low depth samples were removed. However, these results also suggest that CSS (Paulson et al. 2013) and DESeq's variance-stabilizing transformation (Anders & Huber 2010) are promising alternatives for normalization prior to PCoA analysis, especially for weighted distance metrics. For unweighted metrics that are based on species presence and absence, like binary Jaccard and unweighted UniFrac, DESeq's variance-stabilizing transformation performs poorly. This is because the negatives resulting from DESeq's log-like transformation are set to zero (as in McMurdie and Holmes (McMurdie & Holmes 2014)), which ignores rare species.

No good solution exists for the negatives output by the DESeq technique. DESeq was developed mainly for use with Euclidean metrics (Lozupone & Knight 2005; Lozupone et al. 2007), for which negatives are not a problem; however, this issue yields misleading results for ecologically useful non-Euclidean measures, like Bray-Curtis (Bray & Curtis 1957) dissimilarity. Also, the negatives pose a problem to UniFrac's (Lozupone & Knight 2005; Lozupone et al. 2007) branch length. The alternative to setting the negatives to zero, or adding the absolute value of the lowest negative value back to the normalized matrix, will not work with distance metrics that are not Euclidean because it amounts to multiplying the original matrix by a constant due to DESeq's log-like transformation. Also, the addition of a constant (or pseudocount; here, one) to the count matrix prior to CSS (Paulson et al. 2013), DESeq (Anders & Huber 2010), and logUQ (Bullard et al. 2010) transformation as a way to avoid log(0) is not ideal, and clustering results have been shown to be very sensitive to the choice of pseudocount, due to the nonlinear nature of the log transform (Costea et al. 2014; Paulson et al. 2014). This underscores the need for a better solution to the zero problem so that log-like approaches inspired by Aitchison can be used (Aitchison 1982), and is especially critical since microbial matrices are almost always much more than half sparse (Paulson et al. 2013).

While simulations are a useful initial check, real datasets are often much more complex. Therefore, all normalization methods were also examined on real data to check for result and methodological consistency. To perform an initial, detailed comparison of normalization methods, we selected the data set from Caporaso *et al.* (Caporaso et al. 2012). The data included a wide variety of samples, representing both environmental and host-associated sources. To provide an extreme example of differences in sequencing depth, we artificially decreased the library size by 90% for half the samples in the data set. The samples selected for library size reduction were chosen randomly, and the same artificially altered data was used in all normalization comparisons.

Using the data set from Caporaso *et al.* (Caporaso et al. 2012), we observed substantial biases/confounding of results due to sequencing depth. In ordination of unweighted UniFrac distance by PCoA, the soil samples were split into two groups along the first principal coordinate when no normalization was used (Fig. 3a). Soil samples appearing in the group to the left had more reads than those appearing in the group to the right. Similarly, the two stool samples in the data set were arranged close to soil samples with similar library size. When the data was rarefied prior to ordination, soil and stool samples were arranged along the first two coordinates according to sample type rather than library size (Fig. 3b). Other methods of normalization preserved the characteristic pattern seen in the non-normalized data, where soil and stool samples were separated into groups according to library size (Fig. 3c-f).

Normalization did not affect conclusions drawn from non-parametric multivariate ANOVA (PERMANOVA) (Anderson 2001), but we did observe differences in the effect size estimated for sample type, and library size (R^2). Without normalization, the estimated effect size of sample type for unweighted UniFrac distance was R^2=0.40. When the data was rarefied prior to computing distances, the estimated effect size increased to R^2=0.56. Other methods of normalization produced effect sizes similar to the non-normalized result. Although the true effect size is not known for this data set, the environment of origin is known to be a dominant effect in the determination of bacterial species observed (Lozupone & Knight 2007). Without normalization, there is a large effect (R^2=0.14) corresponding to original library size, which is a known artifact of the sequencing process. Rarefying helps to remove the effect of sequencing depth (R^2=0.045), whereas other normalization techniques do not remove this signal artifact, again resembling the non-normalized data.

As another example, we selected the inflammatory bowel disease (IBD) data set from Gevers *et al.* (Gevers et al. 2014). In contrast to the previous data set, all samples here were taken from a single environment type, namely human stool, and were extremely low depth, having an average of 375 sequences per sample. In an ordination of unweighted UniFrac distance with no normalization, there is again strong clustering by library size, with a group of samples with low sequencing depth appearing slightly separate from the other samples (Fig. S2a). Samples in the low-depth group are either dominated by a lack of species detected due to few sequences, thus artificially inflating the β–diversity, or constitute different bacterial species than the main group of stool samples, which should raise suspicion of potential problems from contamination or poor quality PCR products. Furthermore, the first principal coordinate in Fig S2a is more strongly correlated with library size (R^2=0.055, Fig S2b) and poorly correlated with disease state (R^2=0.022), with sampling depth explaining twice the variance of the studied biological effect. Subsampling the data to uniform library size increased the correlation with disease state (R^2=0.036), while other methods did not (R^2=0.022 for proportion, DESeq, and CSS). Because the average library size is so low for this study, the library size also affects weighted UniFrac, where there is again low effect size for this gastrointestinal disorder. Thus, extremely low depth samples still need to be discarded from rarefying alternatives, especially if they are suspected of yielding a poor representation of the true bacterial community due to experimental factors.

PCoA plots using ecologically common metrics for all of the normalization techniques on a few key real datasets representing a gradient (Lauber et al. 2009), distinct body sites (Costello et al. 2009), and time series (Caporaso et al. 2011a) are shown in Supplemental Figures S3-S4. Most measures do well in these cases where there is strong separation between the categories. Clustering according to sequence depth is less of a problem in these datasets since they have strong clustering patterns, however, some clustering according to depth persists. For example, in the ‘Moving Pictures of the Human Microbiome’ dataset (Caporaso et al. 2011a), there is some clustering by sequence depth within each of the four main clusters when normalization alternatives to rarefying are applied. It is noteworthy that CSS normalization results appear robust to the distance metric used, including even Euclidean distance (results not shown), which have been reported to perform poorly on highly sparse matrices (Legendre & Gallagher 2001).

Thus, both simulations and real data suggest that rarefying remains a strong technique for sample normalization prior to ordination and clustering, especially for presence/absence distance metrics that have historically been very useful (such as binary Jaccard and unweighted UniFrac (Lozupone & Knight 2005) distances), subtle effects, small library sizes, and large differences in library size. Of the other methods, and for weighted distance measures, we recommend metagenomeSeq’s CSS (Paulson et al. 2013) or DESeq’s variance stabilizing transformation (Anders & Huber 2010); however, the researcher must check for erroneous clustering according to sequence depth.

Differential Abundance Testing

Differential abundance analysis is useful for testing whether certain microbes have higher relative abundance in one condition vs. another (e.g. healthy vs. diseased patients). More complex statistical methods specifically for RNA-Seq data have been developed and include DESeq (Anders & Huber 2010), DESeq2 (Love MI 2014), edgeR (Robinson et al. 2010; Robinson & Smyth 2008), and Voom (Law et al. 2014) (Table S2). MetagenomeSeq (Paulson et al. 2013) however, was developed specifically for microbial datasets, which usually contain many more zeros than RNA-Seq data. These five methods incorporate more sensitive statistical tests than the standard non-parametric tests such as the Wilcoxon rank-sum test, and they make some distributional assumptions. Therefore, they hold great potential for better prediction of rare OTU behavior.

Previous work in this area concluded that the newer differential abundance testing models are worthwhile, and that the traditional practice of rarefying causes a high rate of false positives (McMurdie & Holmes 2014). However, the latter conclusion was due to an artifact within the simulation (see Methods, Fig. S5a-b). Instead, we found that rarefying does not cause a high rate of false positives, but may lead to false negatives due to the decreased power that results from throwing away some of the data (Fig. 4). The severity of the power decrease caused by rarifying depends upon how much data has been thrown away. (This problem has been known for a long time, leading to the general guideline to rarefy to the highest depth possible without losing too many samples (Carcer et al. 2011).) In order to determine where the greatest loss in power or information occurs when a dataset is rarefied, we constructed power curves from a simple two-species simulation (Fig. 5). The greatest loss in power occurs for rare to common OTUs (e.g. relative abundance ranging from 0.5% to 50%) depending on the library size. This has also been observed in gene expression studies (Robles et al. 2012). Also, consistent with other studies on subsampling (Carcer et al. 2011; Robles et al. 2012), subsampling to library sizes close to the original does not have much effect on the results (50% is treated as "close to the original" in this simplified example, but real microbiome studies are much more complex and thus the real threshold is likely lower, and data-dependent). We also observed that the performance of rarefying degrades faster for smaller library sizes.

Since simulations do not necessarily mirror reality, we again investigated the performance of the techniques on real data. This was done for the techniques shown to be most promising in the simulations: DESeq2 (Love MI 2014), edgeR (Robinson et al. 2010; Robinson & Smyth 2008), metagenomeSeq (Paulson et al. 2013), and rarefying. Ranges of dataset sizes were analyzed for environments that likely contain differentially abundant OTUs, as evidenced by PCoA plots and significance tests (Fig. 6). Approximately 6 samples in each of the categories of human skin vs. soil from Caporaso *et al.* (Caporaso et al. 2012), 28 samples in each of the lean vs. obese categories from Piombino *et al.* (Piombino et al. 2014), and 500 samples in the tongue vs. left palm categories from Caporaso *et al.* (Caporaso et al. 2011a) were tested. Although we do not necessarily know which OTUs are true positives in these actual data, it is of interest to investigate how the most promising techniques compare to each other. While rarefying (at the 15th percentile as in McMurdie and Holmes (McMurdie & Holmes 2014)) finds fewer OTUs as significant, the OTUs it does find to be differentially expressed are remarkably stable. Agreeing with our modified simulation, it does not appear that rarefying causes a high type I error. For example, in Fig. 6 there is high agreement between rarefying and the other techniques. However, edgeR, which is known to be too lenient in its dispersion estimates (Love MI 2014; Paulson et al. 2013), predicts a large number of significantly differentially abundant OTUs relative to other methods, especially for studies with fewer samples (Fig. 6a), suggesting a high false positive rate in agreement with RNA-Seq studies (Love MI 2014; Rapaport et al. 2013; Soneson & Delorenzi 2013).

We also used simulated data to investigate the situation in which the average library size between the two groups was not approximately equal (Fig. 7). We found that of the newer methods, metagenomeSeq's figZIG (Paulson et al. 2013) has a high sensitivity and a low false positive rate (1-specificity) compared to the other techniques. However, the false positive rate is still high. Rarefying achieves the lowest false positive rate, but at a cost to sensitivity. Thus, the method employed by investigators may depend on the sensitivity of the analysis in question to false negatives *vs.* false positives. We often place higher importance in reducing false positives, but this will vary depending on experimental design. For example, study designs in which community analysis is used as a pre-screening, and significant changes will be confirmed in high-throughput follow-up experiments may allow greater tolerance of false positives. However, while both fitZIG or rarefying followed by Wilcoxon rank sum tests in isolation may be applicable for detecting differential abundance in particular situations, our results caution that fitZIG should not be used on rarified data, as this combination of methods caused extremely high false positive rates.

While the no-normalization or proportion approaches perform adequately where the average library size is approximately the same between the two groups (Fig. 4), they do not when one library is 10x larger than the other (Fig. 7). Therefore, we reiterate that neither the no-normalization nor the naive proportion approach should be used for most statistical analyses. To demonstrate this, we suggest the theoretical example of a data matrix with half the samples derived from diseased patients and half from healthy patients. If the samples from the healthy patients have a 10x larger library size, OTUs of all mean abundance levels will be found to be differentially abundant simply because they may have 10x the number of counts in the healthy patient samples. (Such systematic bias can happen if, for example, healthy vs. diseased patients are sequenced on separate sequencing runs or are being compared in a meta-analysis). The same warning applies for naive proportions, especially for rare OTUs that could be deemed differentially abundant simply due to differences in sequencing depth. This is seen even with some filtering to remove very rare OTUs (Fig. 7). We first observed a transition from the results of Fig. 4 to Fig. 7 at around 2-3x difference in library sizes (Fig S6). Further, we investigated uneven numbers of samples per class, with not much difference in results from Fig. 4.

While our previous simulations did not have compositionality, we next evaluated the performance of the techniques with a compositional OTU table (see Methods, Fig. S5b). In simulations where the abundances of 25% of the OTUs increased in one group, no method does well in terms of false positive rate (Fig. S7). Proportion normalization again performs poorly in the face of compositionality, which is present in all realistic datasets. For DESeq/DESeq2, poor performance may be due to the model's assumption that differentially abundant OTUs are not a large portion of the population (Dillies et al. 2013), or the model's overdispersion estimates (Paulson et al. 2013). Thus, compositionality is still a large unsolved problem in differential abundance testing (Lovell et al. 2015), and we would urge caution in data sets where compositionality may play a large role, e.g. when the alpha diversity of the samples is low (Friedman & Alm 2012).

CONCLUSIONS

We built on the pioneering work of McMurdie and Holmes (McMurdie & Holmes 2014), confirming that recently developed more complex techniques for normalization and differential abundance testing hold potential. More testing of the approaches on experimental data is necessary. Of methods for normalizing microbial data for ordination analysis, we found that DESeq normalization (Anders & Huber 2010; Love MI 2014), which was developed for RNA-Seq data and makes use of a log-like transformation, does not work well with ecologically useful metrics, except weighted UniFrac (Lozupone et al. 2007). In contrast, MetagenomeSeq's CSS normalization (Paulson et al. 2013) was developed for microbial data and does not result in troublesome negative output values. However, with techniques other than rarefying, library size can be a confounding factor with very low library sizes (under approximately 1000 sequences per sample), or if presence/absence metrics like unweighted UniFrac are used (Lozupone & Knight 2005). Extremely low-depth samples should be removed regardless of normalization technique, especially if it is suspected that they contain a higher proportion of contaminants (Kennedy et al. 2014; Salter et al. 2014). Also, when using alternatives to rarefying, the researcher must check that clustering by sequence depth does not obscure biologically meaningful results. Therefore, rarefying is still an extremely useful normalization technique, especially for presence/absence metrics. Rarefying can erase the artifact of sample library size better than other normalization techniques, and results in a higher PERMANOVA effect size (R^2) for the studied biological effect, especially for small (<1000 sequences per sample), and uneven library sizes between groups. For both normalization and differential abundance testing, we stress that no normalization and naive proportion approaches should not be used as they can generate artifactual clusters based on sequencing depth, and may result in mistaken OTU differential abundance significance or insignificance.

For differential abundance testing, we studied the methods using both simulations and real data. The most promising of current techniques are based on GLMs with either the negative binomial or zero-inflated Gaussian distributions. It appears that DESeq2 (Love MI 2014), metagenomeSeq's fitZIG (Paulson et al. 2013), and rarefying are all acceptable techniques for approximately even library sizes and numbers of samples per class. DESeq2 was designed for, and is a good option for, increased sensitivity on smaller datasets; however computation time becomes very slow for larger datasets, especially over 100 samples per category. MetagenomeSeq's fitZIG is a faster option for larger library sizes, although it may have a higher false positive rate. The fitZIG technique is designed for larger sample sizes, since more counts per OTU enables more accurate approximation of a continuous distribution. Rarefying, paired with traditional non-parametric tests to account for the non-normal distribution of microbial data, is useful for all dataset sizes, with sensitivity approaching parametric models in larger datasets. Rarefying yields fewer OTUs as significantly differentially abundant, but those OTUs are robust, in the sense that they are almost always identified as significant by at least one other differential abundance detection model. In the case of highly uneven library sizes per category (greater than 2-3x library size difference), we recommend rarefying, which provides higher specificity at a cost to sensitivity, or metagenomeSeq's fitZIG, giving higher sensitivity at a cost to specificity, over the DESeq2 technique. In situations with highly compositional data, no technique does well.

Prior to differential abundance analysis, we recommend checking for significant differences in library size means and distribution between categories (e.g. healthy vs. sick); and propose a Mann-Whitney test, although the subject could be investigated further. The Mann-Whitney test works on the library sizes simulated for this study, as well as that of McMurdie and Holmes (McMurdie & Holmes 2014). To check distributional differences, the library sizes of one sample category can be multiplied by a factor (e.g. 2) to make the means comparable prior to applying the Mann-Whitney test. If there is a significant difference in either mean or distribution, we recommend rarefying paired with a non-parametric test; if not, alternatives to rarefying may be used. For the parametric differential abundance techniques, it is recommended that rare OTUs be filtered out of the matrix prior to differential abundance testing. However, we advise OTU filtering after rarefying, and then applying non-parametric tests. Thanks to McMurdie and Holmes' previous work in this area (McMurdie & Holmes 2014), we recognize the potential of these newer techniques, and have incorporated DESeq2 (Love MI 2014) and metagenomeSeq (Paulson et al. 2013) normalization and differential abundance testing into QIIME version 1.9.0 (Caporaso et al. 2010), along with the traditional rarefying and non-parametric testing techniques.

Acknowlegements

S.J.W. was funded by the National Human Genome Research Institute Grant# 3 R01 HG004872-03S2, and the National Institute of Health Grant# 5 U01 HG004866-04. Research of S.D.P. was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences, NIH (Z01 ES101744-04). We also thank Joseph N. Paulson, Joey McMurdie, and Jonathan Friedman for helpful conversations, and Greg Caporaso and Jai Ram Rideout for coding advice.

Author Contributions

S.J.W, Z.Z.X., A.A., S.D.P., K.B., A.G., J.R.Z., and R.K. designed and conceived analyses. S.J.W. and J.N.P. wrote the QIIME scripts, and A.G.P. and Y.V.B. helped integrate the scripts into QIIME. S.J.W. wrote the initial manuscript, and all authors provided invaluable feedback and insights into analyses and the manuscript. All authors approved the final version of the manuscript.

Agresti A, and Hitchcock DB. *Bayesian inference for categorical data analysis : a survey*. Gainesville, Fla.?: University of Florida.

Aitchison J. 1982. The Statistical-Analysis of Compositional Data. *Journal of the Royal Statistical Society Series B-Methodological* 44:139-177.

Anders S, and Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* 11.

Anders S, McCarthy DJ, Chen YS, Okoniewski M, Smyth GK, Huber W, and Robinson MD. 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols* 8:1765-1786.

Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26:32-46.

Auer PL, and Doerge RW. 2010. Statistical design and analysis of RNA sequencing data. *Genetics* 185:405-416.

Benjamini Y, and Hochberg Y. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* 57:289-300.

Bray JR, and Curtis JT. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* 27:326-349.

Brewer A, and Williamson M. 1994. A New Relationship for Rarefaction. *Biodiversity and Conservation* 3:373-379.

Buccianti A, Mateu-Figueras G, and Pawlowsky-Glahn V. 2006. *Compositional data analysis in the geosciences : from theory to practice*. London: The Geological Society.

Bullard JH, Purdom E, Hansen KD, and Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94.

Cameron AC, and Trivedi PK. *Regression analysis of count data*.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, and Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335-336.

Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, and Knight R. 2011a. Moving pictures of the human microbiome. *Genome Biol* 12:R50.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, and Knight R. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6:1621-1624.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, and Knight R. 2011b. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* 108 Suppl 1:4516-4522.

Carcer DA, Denman SE, McSweeney C, and Morrison M. 2011. Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. *Appl Environ Microbiol* 77:8795-8798.

Cheung YB. 2002. Zero-inflated models for regression analysis of count data: a study of growth and development. *Stat Med* 21:1461-1469.

Colwell RK, Chao A, Gotelli NJ, Lin SY, Mao CX, Chazdon RL, and Longino JT. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* 5:3-21.

Connolly SR, Dornelas M, Bellwood DR, and Hughes TP. 2009. Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. *Ecology* 90:3138-3149.

Costea PI, Zeller G, Sunagawa S, and Bork P. 2014. A fair comparison. *Nat Methods* 11:359.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, and Knight R. 2009. Bacterial community variation in human body habitats across space and time. *Science* 326:1694-1697.

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, and Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505:559-563.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, and Jaffrezic F. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14:671-683.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, and Barcelo-Vidal C. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35:279-300.

Friedman J, and Alm EJ. 2012. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 8:e1002687.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge YC, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, and Zhang JH. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5.

Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, Gonzalez A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, and Xavier RJ. 2014. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15:382-392.

Gotelli NJ, and Colwell RK. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4:379-391.

Gower JC. 1966. Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika* 53:325-&.

Greenacre M. 2011. Measuring Subcompositional Incoherence. *Mathematical Geosciences* 43:681-693.

Holmes I, Harris K, and Quince C. 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* 7:e30126.

Horner-Devine MC, Lage M, Hughes JB, and Bohannan BJ. 2004. A taxa-area relationship for bacteria. *Nature* 432:750-753.

Jari Oksanen FGB, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens and Helene Wagner. 2015. vegan: Community Ecology Package. *R package version 22-1*.

Jernvall J, and Wright PC. 1998. Diversity components of impending primate extinctions. *Proc Natl Acad Sci U S A* 95:11279-11283.

Kaufman L. RP. 1990. *Finding Groups in Data: An introduction to Cluster Analysis*: JohnWiley & Sons.

Kelly CR, Ihunnah C, Fischer M, Khoruts A, Surawicz C, Afzali A, Aroniadis O, Barto A, Borody T, Giovanelli A, Gordon S, Gluck M, Hohmann EL, Kao D, Kao JY, McQuillen DP, Mellow M, Rank KM, Rao K, Ray A, Schwartz MA, Singh N, Stollman N, Suskind DL, Vindigni SM, Youngster I, and Brandt L. 2014. Fecal microbiota transplant for treatment of Clostridium difficile infection in immunocompromised patients. *Am J Gastroenterol* 109:1065-1071.

Kennedy K, Hall MW, Lynch MD, Moreno-Hagelsieb G, and Neufeld JD. 2014. Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Appl Environ Microbiol* 80:5717-5722.

Lauber CL, Hamady M, Knight R, and Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* 75:5111-5120.

Law CW, Chen YS, Shi W, and Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15.

Legendre P, and Gallagher ED. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129:271-280.

Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, and Gordon JI. 2005. Obesity alters gut microbial ecology. *Proc Natl Acad Sci U S A* 102:11070-11075.

Love MI HWaAS. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15.

Lovell D ea. 2010. Caution! compositions! can constraints on omics data lead analyses astray? *CSIRO*:1-44.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, and Bahler J. 2015. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol* 11:e1004075.

Lozupone C, and Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71:8228-8235.

Lozupone C, Lladser ME, Knights D, Stombaugh J, and Knight R. 2011. UniFrac: an effective distance metric for microbial community comparison. *ISME J* 5:169-172.

Lozupone CA, Hamady M, Kelley ST, and Knight R. 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 73:1576-1585.

Lozupone CA, and Knight R. 2007. Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A* 104:11436-11440.

Lozupone CA, Li M, Campbell TB, Flores SC, Linderman D, Gebert MJ, Knight R, Fontenot AP, and Palmer BE. 2013a. Alterations in the Gut Microbiota Associated with HIV-1 Infection. *Cell Host Microbe* 14:329-339.

Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vazquez-Baeza Y, Jansson JK, Gordon JI, and Knight R. 2013b. Meta-analyses of studies of the human microbiota. *Genome Research* 23:1704-1714.

McMurdie PJ, and Holmes S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* 8.

McMurdie PJ, and Holmes S. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput Biol* 10.

Paulson JN, Bravo HC, and Pop M. 2014. Reply to: "a fair comparison". *Nat Methods* 11:359-360.

Paulson JN, Stine OC, Bravo HC, and Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 10:1200-1202.

Pearson K. 1896. Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proc Roy Soc* 60:489-498.

Piombino P, Genovese A, Esposito S, Moio L, Cutolo PP, Chambery A, Severino V, Moneta E, Smith DP, Owens SM, Gilbert JA, and Ercolini D. 2014. Saliva from obese individuals suppresses the release of aroma compounds from wine. *PLoS One* 9:e85611.

Pop M, Walker AW, Paulson J, Lindsay B, Antonio M, Hossain MA, Oundo J, Tamboura B, Mai V, Astrovskaya I, Corrada Bravo H, Rance R, Stares M, Levine MM, Panchalingam S, Kotloff K, Ikumapayi UN, Ebruke C, Adeyemi M, Ahmed D, Ahmed F, Alam MT, Amin R, Siddiqui S, Ochieng JB, Ouma E, Juma J, Mailu E, Omore R, Morris JG, Breiman RF, Saha D, Parkhill J, Nataro JP, and Stine OC. 2014. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biol* 15:R76.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, and Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59-65.

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, and Betel D. 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 14:R95.

Reynolds A RG, Iglesia B, Rayward-Smith V. 2006. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* 5:475-504.

Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL, Griffin NW, Lombard V, Henrissat B, Bain JR, Muehlbauer MJ, Ilkayeva O, Semenkovich CF, Funai K, Hayashi DK, Lyle BJ, Martini MC, Ursell LK, Clemente JC, Van Treuren W, Walters WA, Knight R, Newgard CB, Heath AC, and Gordon JI. 2013. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* 341:1241214.

Robinson MD, McCarthy DJ, and Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-140.

Robinson MD, and Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11:R25.

Robinson MD, and Smyth GK. 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9:321-332.

Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, and Taylor JM. 2012. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics* 13:484.

Rodriguez RL, and Konstantinidis KT. 2014. Estimating coverage in metagenomic data sets and why it matters. *ISME J* 8:2349-2351.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, and Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 12:87.

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, Van Horn DJ, and Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75:7537-7541.

Shankar V, Hamilton MJ, Khoruts A, Kilburn A, Unno T, Paliy O, and Sadowsky MJ. 2014. Species and genus level resolution analysis of gut microbiota in Clostridium difficile patients following fecal microbiota transplantation. *Microbiome* 2:13.

Soneson C, and Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14:91.

Team RC. 2014. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, and Gordon JI. 2009. A core gut microbiome in obese and lean twins. *Nature* 457:480-484.

Vazquez-Baeza Y, Pirrung M, Gonzalez A, and Knight R. 2013. EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience* 2:16.

Wagner BD, Robertson CE, and Harris JK. 2011. Application of two-part statistics for comparison of sequence variant counts. *PLoS One* 6:e20296.

Weingarden A, Gonzalez A, Vazquez-Baeza Y, Weiss S, Humphry G, Berg-Lyons D, Knights D, Unno T, Bobr A, Kang J, Khoruts A, Knight R, and Sadowsky MJ. 2015. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. *Microbiome* 3:10.

White JR, Nagarajan N, and Pop M. 2009. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 5:e1000352.

Witten DM. 2011. Classification and Clustering of Sequencing Data Using a Poisson Model. *Annals of Applied Statistics* 5:2493-2518.

Yu D, Huber W, and Vitek O. 2013. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics* 29:1275-1282.

FIGURE CAPTIONS:

Figure 1: Effect of sampling depth on ordination methods. (a) Data rarefied at 500 sequences per sample**.** (b, c) Data not normalized, with a random half of the samples subsampled to 500 sequences per sample and the other half to 50 sequences per sample. (b) is colored by subject_ID, (c) is colored by sequences per sample. Non-parametric ANOVA (PERMANOVA) effect sizes (R^2) roughly represent the percent variance that can be explained by the given variable. Asterisk (*) indicates significance at $p < 0.01$. The distance metric of unweighted UniFrac was used for all panels.

Figure 2: Comparison of common distance metrics and normalization methods across library sizes when low-coverage samples are excluded. Clustering accuracy is shown for all combinations of five common distance metrics (panels arranged from left to right) across four library depths (panels arranged from top to bottom; N_L, median library size), six sample normalization methods (series within each panel), and several effect sizes (x-axis within panels). In all cases, samples below the 15th percentile of library size were dropped from the analysis in order to isolate the effects of rarifying from the effects of dropping low-coverage samples. The x-axis ('effect size') within each panel represents the multinomial mixing proportions of the two sample classes '*Ocean*' and '*Feces*'. A higher effect

size represents an easier clustering task. The y-axis (‘accuracy’) shows the accuracy of each classifier, as assessed by the fraction of simulated samples correctly clustered.

Figure 3: Rarefying clusters more according to biological origin, and diminishes the effect of library size. Rarefying exhibits a higher effect size (R^2) for biological origin, and a lower effect size (R^2) of original library size. Unweighted UniFrac was used for clustering, and a random half of samples were subsampled to 10 times fewer sequences per sample. The 45-degree line splits low from high depth samples in all but the rarefying technique. For each letter (a-f), the left PCoA plot is colored according to the ‘Canine Feces’, etc. legend, and the right PCoA plot is colored according to the ‘High/Low Library Size’ legend.

Figure 4: Differential abundance detection performance.
The AUC ('Area Under the Curve') version of the ROC ('Receiver Operator Characteristic’) curve is the ratio of sensitivity to (1-specificity), or true positive rate vs. false positive rate. A higher AUC indicates better differential abundance detection performance. The 'effect size' represents the fold-change of the 'true positive' OTUs from one condition (e.g. case) to another (e.g. control). The right axis represents the median library size (N_L), while the shading on the graph lines represents the number of samples per class. ‘Model/None’ represents data analyzed with a parametric statistical model (e.g. DESeq), or no normalization. Blue lines in, e.g. the DESeq column represents the data was rarefied, then DESeq was applied. Since the fitZIG model depends upon original library size information, the model does poorly on rarefied data.

Figure 5: The effect of rarefying on power for different OTU relative abundances and library sizes.
The detection power for differentially abundant OTUs of varying levels of relative abundance (very rare to common). This is for two samples A and B. For power calculations, we assumed that OTU1 fraction of group B is 85% of the OTU1 fraction of group A. Library type A was fixed, while library size B was subsampled at different percentages, creating the power curves calculated with Fisher’s exact test.

Figure 6: Comparison of the most promising differential abundance detection techniques on real datasets.
Each table’s diagonal represents the number of OTUs found significant (Benjamini & Hochberg FDR < 0.05) by that technique. The off-diagonal entries represent the number of shared differentially abundant OTUs between two techniques. The bar charts represents the percentage of differentially abundant OTUs shared by at least one other technique.

Figure 7: Differential abundance detection performance where one sample group average library size is 10 times the size of the other. Labels are the same as in Fig. 4. A significant difference from the results of Fig. 4 was first observed at 2-3-fold difference in library sizes (see Fig. S6).

Figure S1: Simulated clustering accuracy if rarefying is not penalized for removing the lowest 15th percentile samples.
The right axis represents the median library size (N_L), while the x-axis ‘effect size’ is the multinomial mixing proportions of the two classes of samples, *‘Ocean’* and *‘Feces’*. See caption

for Fig. 2 for further details.

Figure S2: Low library size samples can diminish result quality, regardless of normalization technique. We show the inflammatory bowel disease (IBD) dataset of Gevers et al.(Gevers et al. 2014), which has an average library size 375 sequences per sample. **(a)** Extremely low depth samples cluster in lower right hand corner of PCoA plots with no normalization, or rarefying alternatives, unweighted UniFrac. (b) The original library size of samples is a dominant effect, even influencing weighted UniFrac, with low library sizes and subtle biological clustering for rarefying alternatives. This diminishes if low library size samples are removed from analysis.

Figure S3: All normalization techniques on key microbiome datasets, Bray Curtis distance. Rows of panels show (from top to bottom) data from 88soils (Lauber et al. 2009), Body Sites (Costello et al. 2009), Moving Pictures (Caporaso et al. 2011a). 88 soils is colored according to a color gradient from low to high pH. The Costello et al. body sites dataset is colored according to body site: feces (blue), oral cavity (purple), the rest of the colors are external auditory canal, hair, nostril, skin, and urine. Moving Pictures dataset: Left and Right palm (red/blue), tongue (green), feces (orange). It is important to note that all the samples in these datasets are approximately the same depth, and there are very strong driving gradients.

Figure S4: All normalization techniques on key microbiome datasets, unweighed UniFrac distance. See Figure S3 caption for details.

Figure S5: Simple example of the reasoning behind differential abundance simulations. (a) In actual OTU tables generated from sequencing data, the counts (left column) are already compositional and therefore only relative (left column). Application of the 'effect size' to the original 'Multinomial' template to create fold-change differences disturbs the distinction between true positive (TP) and true negative (TN) OTUs in the 'Original' simulation, but not the 'Balanced' simulation. (c) Creation of a 'Compositional' OTU table from the 'Multinomial' template, where the counts/relative abundances are intentionally blurred for the TN OTUs.

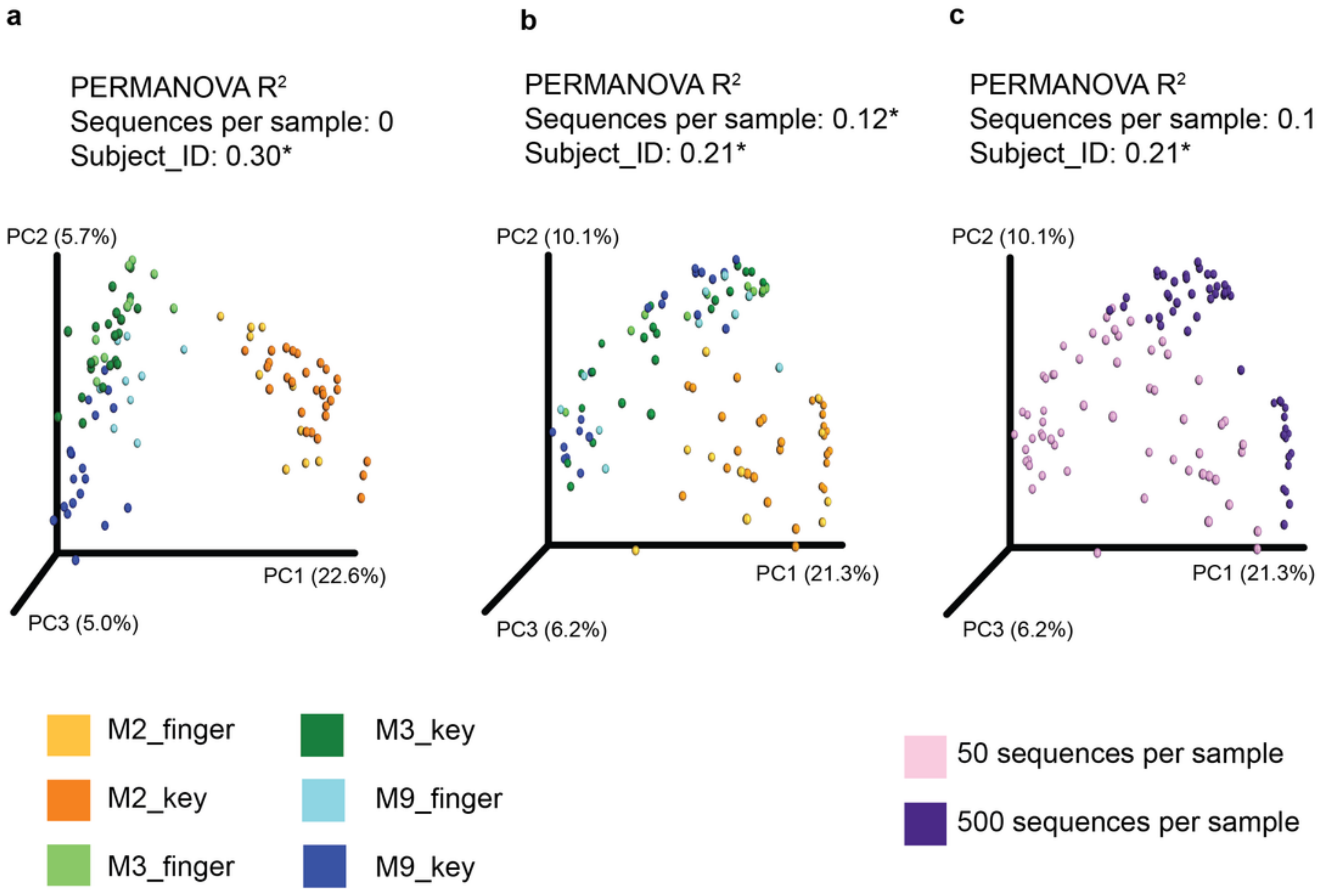
Figure S6: Differential abundance detection performance where one sample group average library size is 3 times the size of the other. Labels are the same as in Fig. 4.

Figure S7: Differential abundance detection performance when the dataset is compositional. 25% of OTUs are differentially abundant. Labels are the same as in Fig. 4.

1

Effect of sampling depth on ordination methods

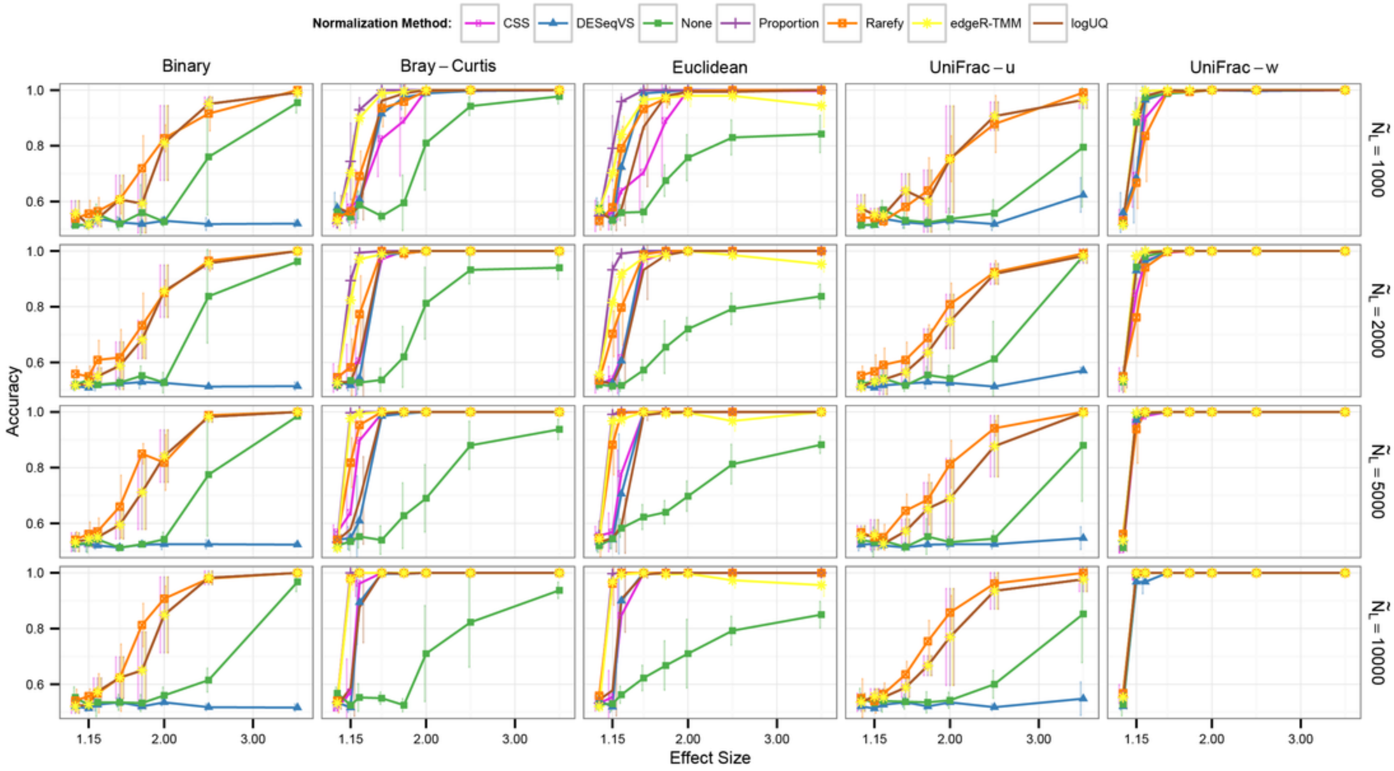
(a) Data rarefied at 500 sequences per sample**.** (b, c) Data not normalized, with a random half of the samples subsampled to 500 sequences per sample and the other half to 50 sequences per sample. (b) is colored by subject_ID, (c) is colored by sequences per sample. Non-parametric ANOVA (PERMANOVA) effect sizes (R^2) roughly represent the percent variance that can be explained by the given variable. Asterisk (*) indicates significance at p < 0.01. The distance metric of unweighted UniFrac was used for all panels.



2

Comparison of common distance metrics and normalization methods across library sizes when low-coverage samples are excluded.

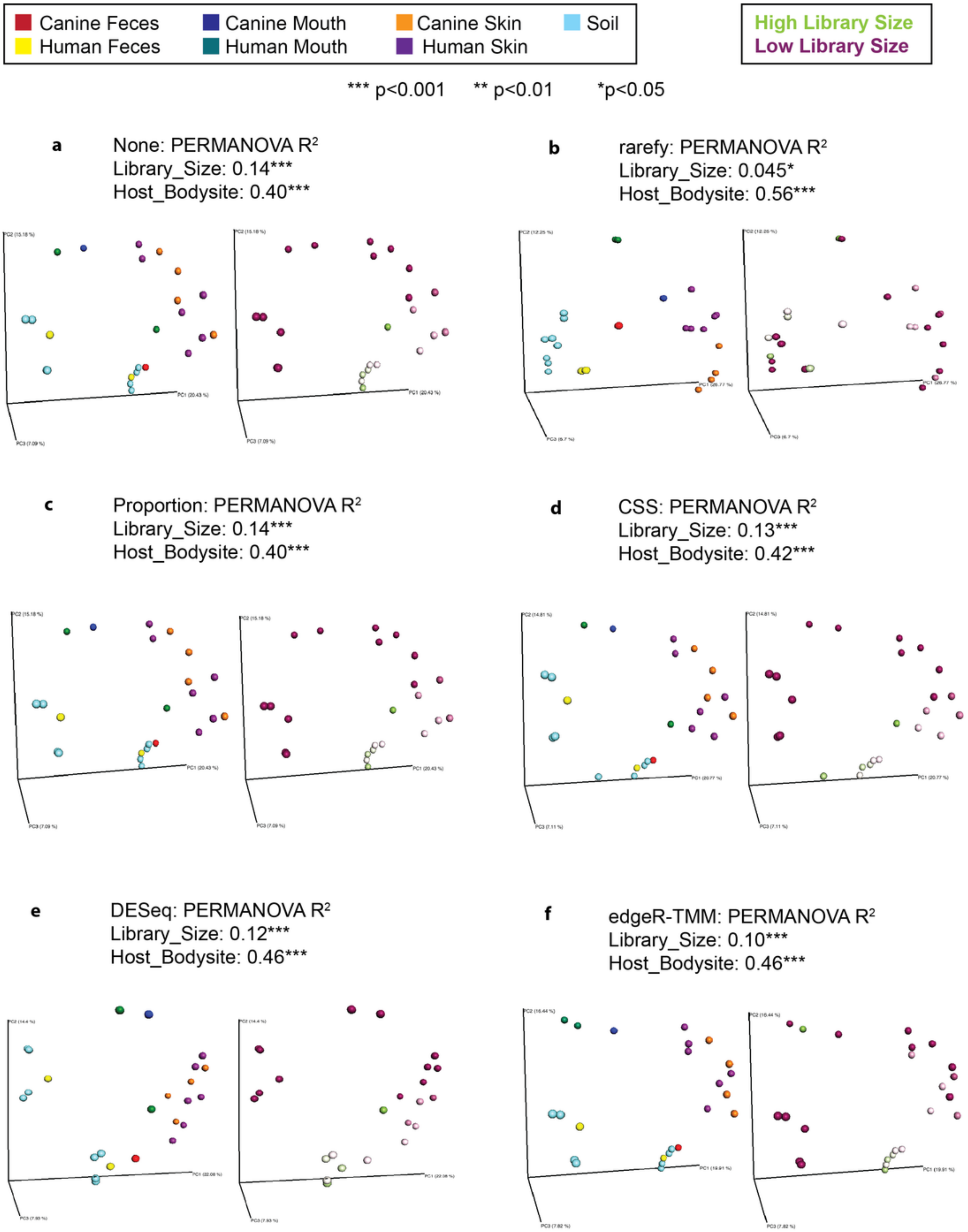
Clustering accuracy is shown for all combinations of five common distance metrics (panels arranged from left to right) across four library depths (panels arranged from top to bottom; N_L, median library size), six sample normalization methods (series within each panel), and several effect sizes (x-axis within panels). In all cases, samples below the 15th percentile of library size were dropped from the analysis in order to isolate the effects of rarifying from the effects of dropping low-coverage samples. The x-axis ('effect size') within each panel represents the multinomial mixing proportions of the two sample classes '*Ocean*' and '*Feces*'. A higher effect size represents an easier clustering task. The y-axis (‘accuracy’) shows the accuracy of each classifier, as assessed by the fraction of simulated samples correctly clustered.



3

Rarefying clusters more according to biological origin, and diminishes the effect of library size.

Rarefying exhibits a higher effect size (R^2) for biological origin, and a lower effect size (R^2) of original library size. Unweighted UniFrac was used for clustering, and a random half of samples were subsampled to 10 times fewer sequences per sample. The 45-degree line splits low from high depth samples in all but the rarefying technique. For each letter (a-f), the left PCoA plot is colored according to the 'Canine Feces', etc. legend, and the right PCoA plot is colored according to the 'High/Low Library Size' legend.

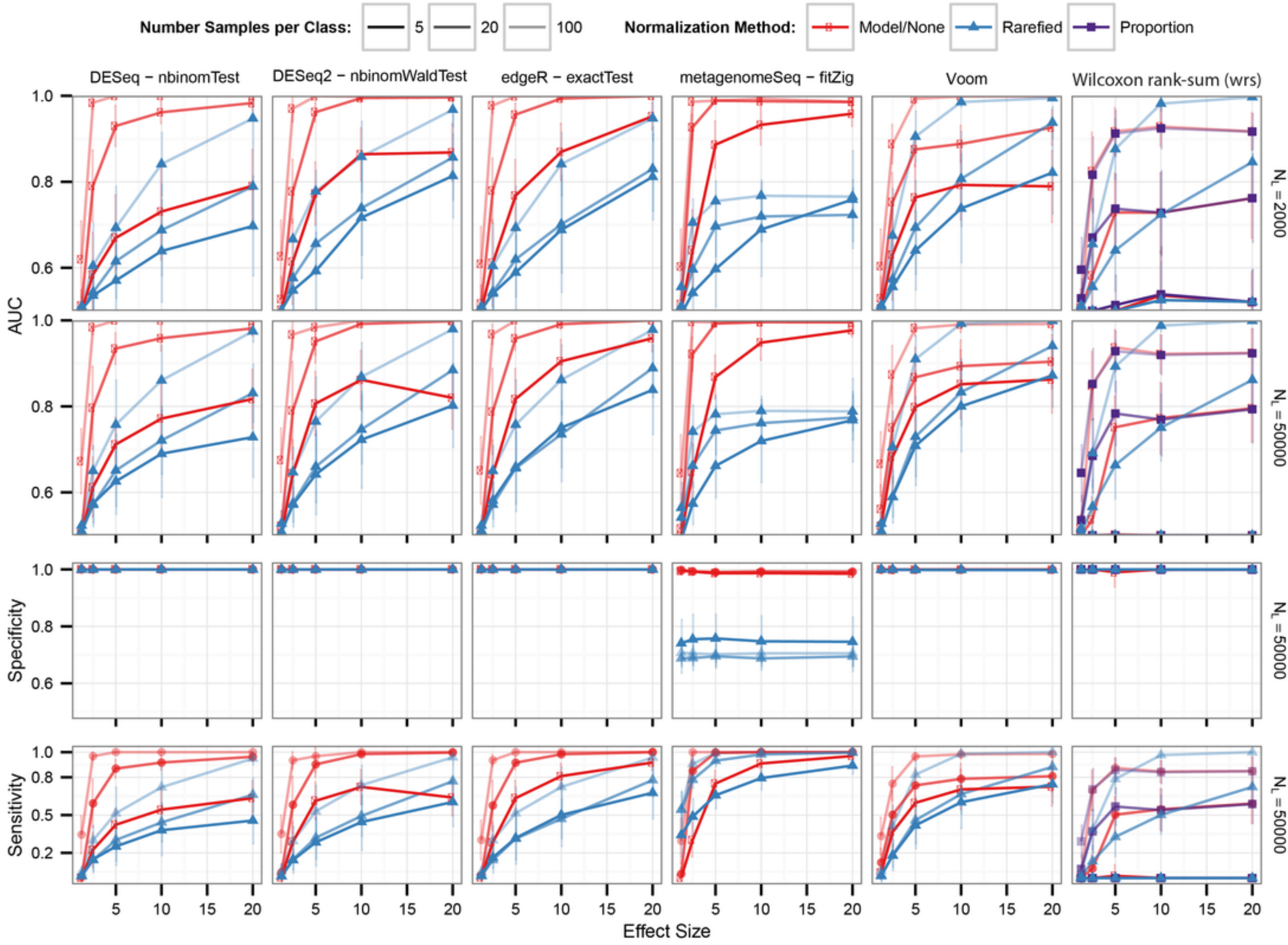

Canine Feces
Human Feces
Canine Mouth
Human Mouth
Canine Skin
Human Skin
Soil
High Library Size
Low Library Size
*** p<0.001
** p<0.01
*p<0.05
a
None: PERMANOVA R²
Library_Size: 0.14***
Host_Bodysite: 0.40***
PC2 (15.18 %)
PC1 (20.43 %)
PC3 (7.09 %)
b
rarefy: PERMANOVA R²
Library_Size: 0.045*
Host_Bodysite: 0.56***
PC2 (12.25 %)
PC1 (26.77 %)
PC3 (6.7 %)
c
Proportion: PERMANOVA R²
Library_Size: 0.14***
Host_Bodysite: 0.40***
PC2 (15.18 %)
PC1 (20.43 %)
PC3 (7.09 %)
d
CSS: PERMANOVA R²
Library_Size: 0.13***
Host_Bodysite: 0.42***
PC2 (14.81 %)
PC1 (20.77 %)
PC3 (7.11 %)
e
DESeq: PERMANOVA R²
Library_Size: 0.12***
Host_Bodysite: 0.46***
PC2 (14.4 %)
PC1 (22.08 %)
PC3 (7.93 %)
f
edgeR-TMM: PERMANOVA R²
Library_Size: 0.10***
Host_Bodysite: 0.46***
PC2 (16.44 %)
PC1 (19.91 %)
PC3 (7.82 %)

4

Differential abundance detection performance.

The AUC ('Area Under the Curve') version of the ROC ('Receiver Operator Characteristic') curve is the ratio of sensitivity to (1-specificity), or true positive rate vs. false positive rate. A higher AUC indicates better differential abundance detection performance. The 'effect size' represents the fold-change of the 'true positive' OTUs from one condition (e.g. case) to another (e.g. control). The right axis represents the median library size (N_L), while the shading on the graph lines represents the number of samples per class. ‘Model/None’ represents data analyzed with a parametric statistical model (e.g. DESeq), or no normalization. Blue lines in, e.g. the DESeq column represents the data was rarefied, then DESeq was applied. Since the fitZIG model depends upon original library size information, the model does poorly on rarefied data.

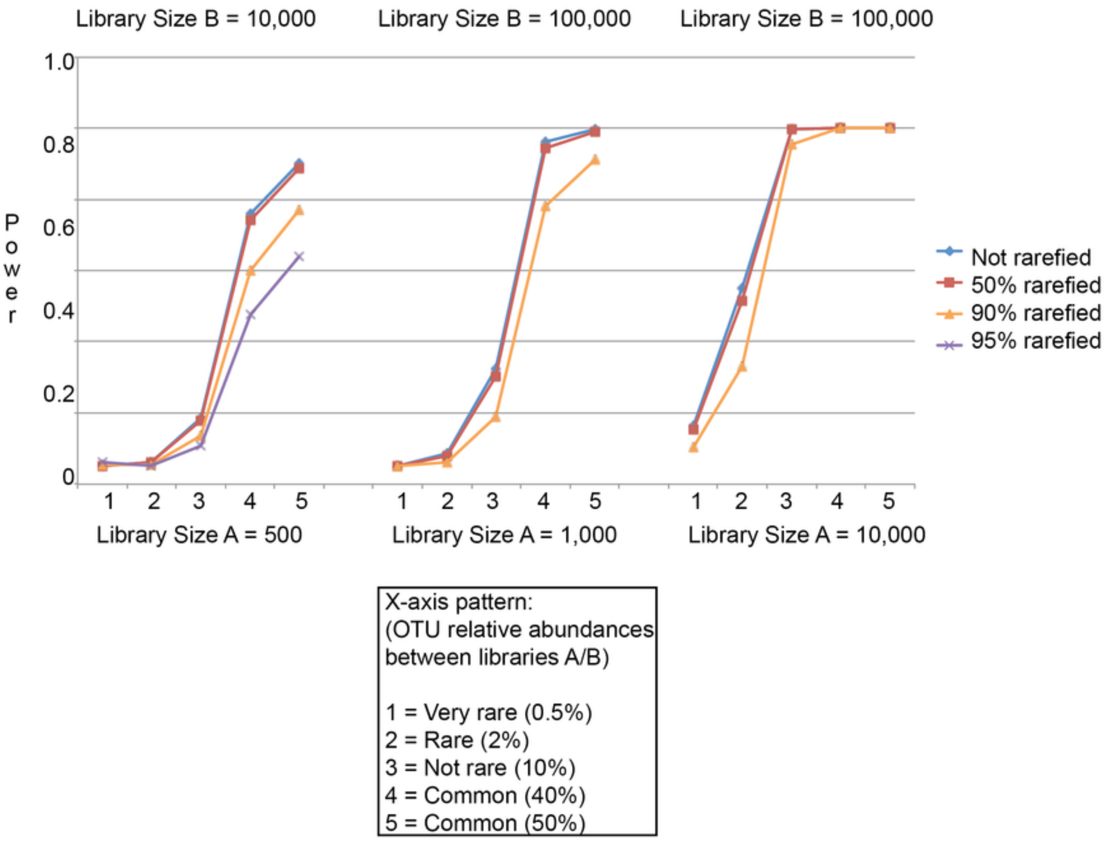
PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1157v1 | CC-BY 4.0 Open Access | rec: 3 Jun 2015, publ: 3 Jun 2015


Number Samples per Class:
5
20
100
Normalization Method:
Model/None
Rarefied
Proportion
DESeq − nbinomTest
DESeq2 − nbinomWaldTest
edgeR − exactTest
metagenomeSeq − fitZig
Voom
Wilcoxon rank-sum (wrs)
AUC
Specificity
Sensitivity
1.0
0.8
0.6
0.5
0.2
$\tilde{N}_L = 2000$
$\tilde{N}_L = 50000$
5
10
15
20
Effect Size

5

The effect of rarefying on power for different OTU relative abundances and library sizes.

The detection power for differentially abundant OTUs of varying levels of relative abundance (very rare to common). This is for two samples A and B. For power calculations, we assumed that OTU1 fraction of group B is 85% of the OTU1 fraction of group A. Library type A was fixed, while library size B was subsampled at different percentages, creating the power curves calculated with Fisher’s exact test.



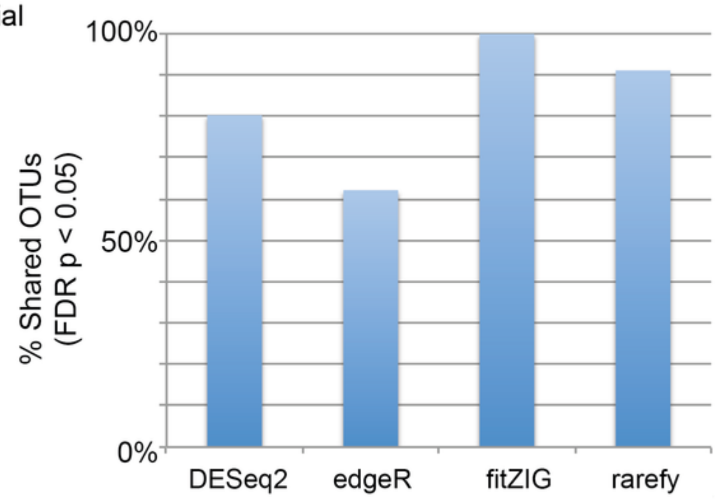
6

Comparison of the most promising differential abundance detection techniques on real datasets.

Each table's diagonal represents the number of OTUs found significant (Benjamini & Hochberg FDR < 0.05) by that technique. The off-diagonal entries represent the number of shared differentially abundant OTUs between two techniques. The bar charts represents the percentage of differentially abundant OTUs shared by at least one other technique.

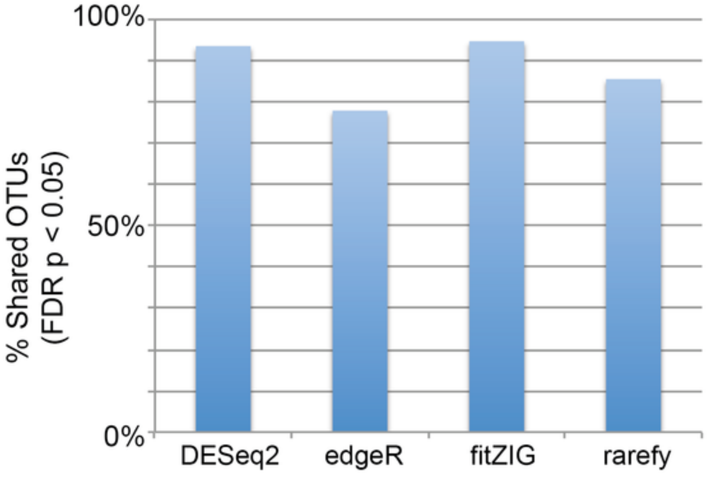
a Caporaso *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms *ISME* (2012).
~ 6 skin samples, 8 soil samples
mean sequences per sample: 1.3 million

	DESeq2	edgeR	fitZIG	rarefy
DESeq2	**1182**	1135	934	376
edgeR		**2357**	1344	706
fitZIG			**1445**	680
rarefy				**763**



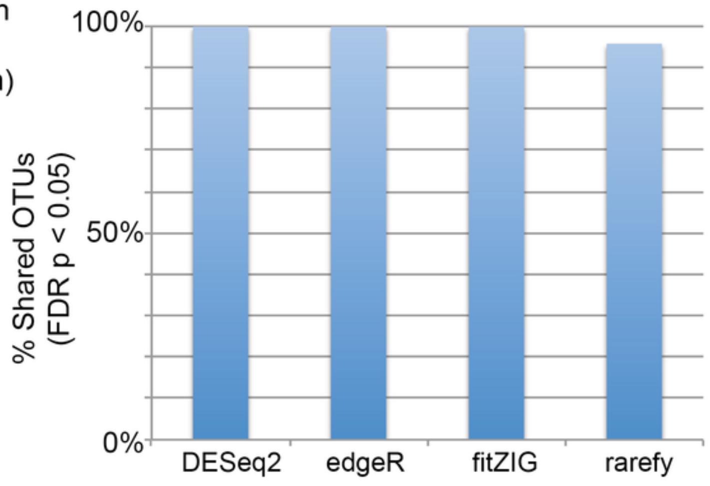
b Piombino *et al.* Saliva from Obese Individuals Suppresses the Release of Aroma Compounds from Wine. *PLoS One* (2014).
~28 samples per category (lean vs. obese)
mean sequences per sample: 75,580

	DESeq2	edgeR	fitZIG	rarefy
DESeq2	**232**	206	164	122
edgeR		**267**	159	113
fitZIG			**189**	128
rarefy				**145**



c Caporaso *et al.* Moving Pictures of the Human Microbiome. *Genome Biol.* (2011).
~500 samples per category (tongue vs. left palm)
mean sequences per sample: 25,600

	DESeq2	edgeR	fitZIG	rarefy
DESeq2	**1070**	1051	955	1038
edgeR		**1122**	995	1097
fitZIG			**1038**	1004
rarefy				**1185**

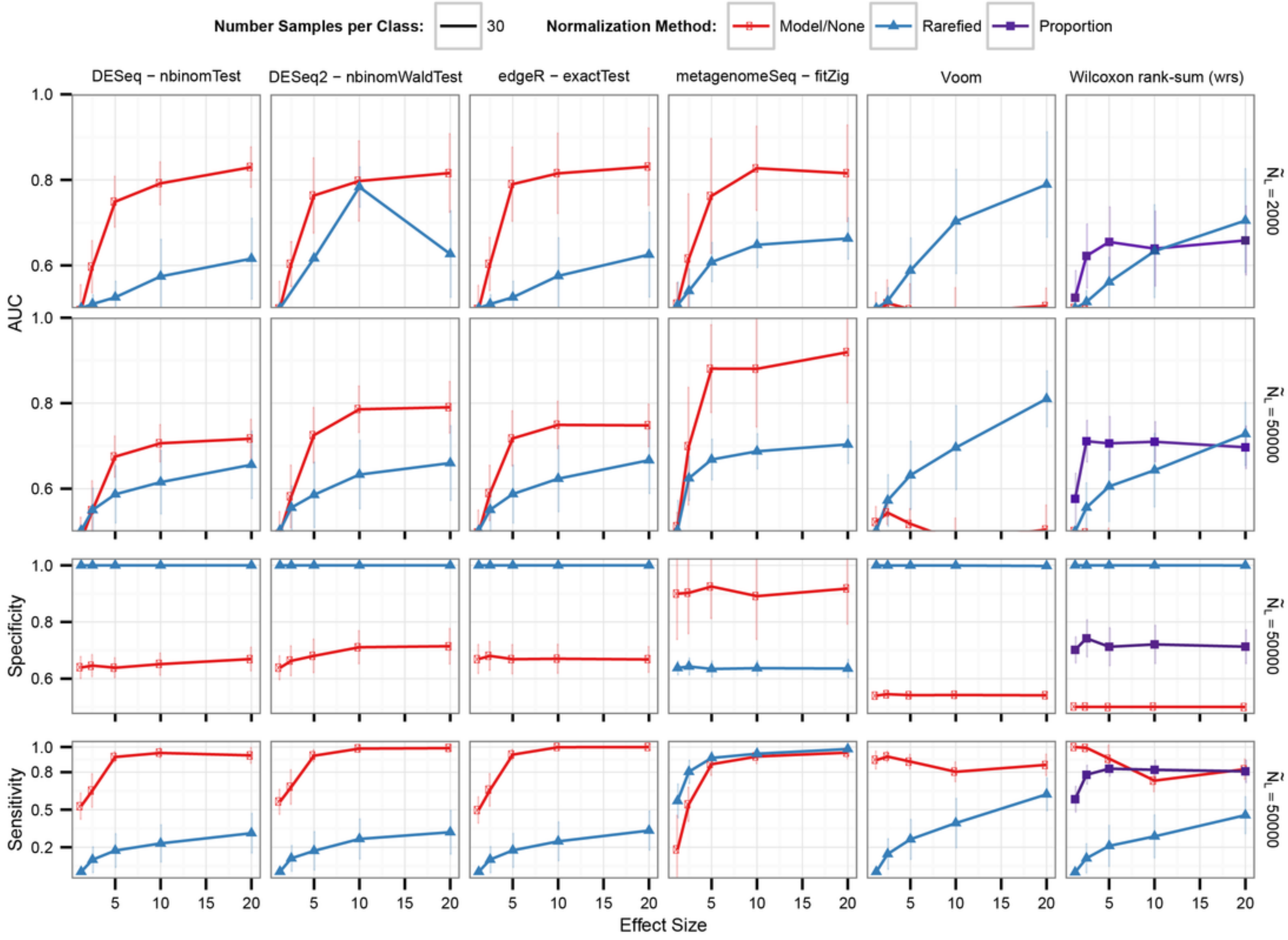


PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1157v1 | CC-BY 4.0 Open Access | rec: 3 Jun 2015, publ: 3 Jun 2015

7

Differential abundance detection performance where one sample group average library size is 10 times the size of the other.

Labels are the same as in Fig. 4. A significant difference from the results of Fig. 4 was first observed at 2-3-fold difference in library sizes (see Fig. S6).


Number Samples per Class: 30
Normalization Method: Model/None Rarefied Proportion
DESeq − nbinomTest
DESeq2 − nbinomWaldTest
edgeR − exactTest
metagenomeSeq − fitZig
Voom
Wilcoxon rank-sum (wrs)
AUC
Specificity
Sensitivity
Effect Size
$\tilde{N}_L = 2000$
$\tilde{N}_L = 50000$