A peer-reviewed version of this preprint was published in PeerJ on 2 April 2018.

View the peer-reviewed version (peerj.com/articles/4600), which is the preferred citable publication unless you specifically need to cite this preprint.

Chen L, Reeve J, Zhang L, Huang S, Wang X, Chen J. 2018. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. PeerJ 6:e4600 https://doi.org/10.7717/peerj.4600

GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data

Li Chen [1], **James Reeve** [2], **Lujun Zhang** [3], **Shengbin Huang** [2], **Jun Chen** Corresp. [2, 4]

[1] Department of Health Outcomes Research and Policy, Harrison School of Pharmacy, Auburn University, Auburn, Alabama, United States

[2] Bioinformatics and Computational Biology Program, University of Minnesota - Rochester, Rochester, Minnesota, United States

[3] College of Environmental and Resource Sciences, Zhejiang University, Hangzhou, Zhejiang, China

[4] Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota, United States

Corresponding Author: Jun Chen
Email address: chen.jun2@gmail.com

Normalization is the first critical step in microbiome sequencing data analysis used to account for variable library sizes. Current RNA-Seq based normalization methods that have been adapted for microbiome data fail to consider the unique characteristics of microbiome data, which contain a vast number of zeros due to the physical absence or under-sampling of the microbes. Normalization methods that specifically address the zero inflation remain largely undeveloped. Here we propose GMPR - a simple but effective normalization method - for zero-inflated sequencing data such as microbiome data. Simulation studies and real datasets analyses demonstrate that the proposed method is more robust than competing methods, leading to more powerful detection of differentially abundant taxa and higher reproducibility of the relative abundances of taxa.

GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data

Li Chen[1], James Reeve[2], Lujun Zhang[3], Shengbing Huang[2], and Jun Chen[2, 4 *]

[1]**Department of Health Outcomes Research and Policy, Harrison School of Pharmacy, Auburn University, AL 36849, USA**
[2]**Bioinformatics and Computational Biology Program, University of Minnesota, Rochester, MN 55905, USA**
[3]**College of Environmental and Resource Sciences, Zhejiang University, Zhejiang, 310058, China**
[4]**Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905,USA**
[*]**Corresponding addressed to: Chen.Jun2@mayo.edu**

ABSTRACT

Normalization is the first critical step in microbiome sequencing data analysis used to account for variable library sizes. Current RNA-Seq based normalization methods that have been adapted for microbiome data fail to consider the unique characteristics of microbiome data, which contain a vast number of zeros due to the physical absence or under-sampling of the microbes. Normalization methods that specifically address the zero inflation remain largely undeveloped. Here we propose GMPR - a simple but effective normalization method - for zero-inflated sequencing data such as microbiome data. Simulation studies and real datasets analyses demonstrate that the proposed method is more robust than competing methods, leading to more powerful detection of differentially abundant taxa and higher reproducibility of the relative abundances of taxa.

INTRODUCTION

Hight-throughput sequencing experiments such as RNA-seq and microbiome sequencing are now routinely employed to interrogate the biological systems at the genome scale (Wang et al., 2009). After processing of the raw sequence reads, the sequencing data usually presents as a count table of detected features. The complex processes involved in the sequencing causes the sequencing depth (library size) to vary across samples, sometimes ranging several orders of magnitude. Normalization, which aims to correct or reduce the bias introduced by variable library sizes, is an essential preprocessing step before any downstream statistical analyses for high-throughput sequencing experiments (Dillies et al., 2013; Li et al., 2015). An inappropriate normalization method may either reduce statistical power with the introduction of unwanted variation, or more severely, result in falsely discovered features. Normalization is especially critical when the library size is a confounding factor that correlates with the variable of interest. One popular approach for normalizing the sequencing data involves calculating a size factor for each sample as an estimate of the library size. The size factors can be used to divide the read counts to produce normalized data (in the form of relative abundances), or to be included as offsets in count-based regression models such as DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010) for differential feature analysis. One simple normalization method is TSS (Total Sum Scaling), which uses the total read count for each sample as the size factor. However, there are a couple undesirable properties for TSS. First, it is not robust to outlier counts. Outliers have frequently been observed in sequencing samples due to technical artifacts such as preferential amplification by PCR (Aird et al., 2011). Several outliers could bias the library size estimates significantly. Second, it creates compositional effects: non-differential features will appear to be differential due to the constant-sum constraint (Tsilimigras and Fodor, 2016). Compositional effects

are much stronger for data where there are overly abundant features and the total number of features is relatively small. An ideal normalization method should thus capture the invariant part of the count distribution and be robust to outliers and differential features. The latter property is important to reduce the false positives due to compositionality.

Many normalization methods have been developed for sequencing data generally, and for RNA-Seq data in particular. These methods usually rely on the assumption that the majority of features do not change with respect to a certain condition so that a robust statistic (i.e. median or quantile), which is not sensitive to a small set of differential features, could be used to quantify the library size. Two popular normalization methods for RNA-Seq data include TMM (Trimmed Mean of M values, implemented in edgeR) (Robinson and Oshlack, 2010) and the DESeq normalization (equivalent to Relative Log Expression normalization implemented in edgeR. For simplicity, we label it as "RLE".) (Anders and Huber, 2010).

Compared to RNA-Seq data, microbiome sequencing data are more over-dispersed and contain a vast number of zeros. Take the COMBO data for example (Wu et al., 2011), it contains 1,873 non-singleton OTUs (Operational Taxonomic Units, a proxy for bacterial species) from 98 samples and more than 90% are zeros. The observed zeros are a mixture of "structural zeros" (due to physical absence) and "sampling zeros" (due to under-sampling). One popular strategy to circumvent the zero inflation problem is to add a pseudo-count. This practice has a Bayesian explanation and implicitly assumes that all the zeros are due to under-sampling (McMurdie and Holmes, 2014). However, this assumption may not be appropriate due to the large extent of structural zeros. Moreover, the choice of the pseudo-count is very arbitrary and it has been shown that the clustering results can be highly dependent upon the choice (Costea et al., 2014). Recently, a new normalization method CSS (Cumulative Sum Scaling) has been developed for microbiome sequencing data (Paulson et al., 2013). In CSS, raw counts are divided by the cumulative sum of counts, up to a percentile determined using a data-driven approach. The percentile is aimed to capture the relatively invariant count distribution for a dataset. However, the determination of the percentiles could fail for microbiome datasets that have high count variability. Therefore, a more robust method to address the zero-inflated sequencing data is still needed.

Here we propose a novel inter-sample normalization method GMPR (Geometric Mean of Pairwise Ratios), developed specifically for zero-inflated sequencing data such as microbiome sequencing data. By comprehensive tests on simulated and real datasets, we show that GMPR outperforms the other competing methods for zero-inflated count data.

METHODS

Our method extends the idea of RLE normalization for RNA-seq data. Assume we have a count table of OTUs by 16S rDNA targeted microbiome sequencing. Denote the c_{ki} as the count of the kth OTU ($k = 1, \cdots, q$) in the ith ($i = 1, \cdots, n$) sample. The RLE method consists of two steps:

- Step 1: Calculate the geometric means for all OTUs

$$\mu_k^{GM} = (c_{1k} c_{2k} \cdots c_{nk})^{1/n},\ k = 1, \cdots, q$$

- Step 2: For a given sample,

$$s_i = median_k\{c_{ik}/\mu_k^{GM}\},\ i = 1, \cdots, n$$

Since geometric mean is not defined for features with 0s, features with 0s are usually excluded for size calculation. However, for zero-inflated data such as microbiome sequencing data, as the sample size increases, the probability of existence of features without any 0s becomes smaller. It is not uncommon that a large dataset does not share any common taxa. In such cases, RLE fails. As an alternative, a pseudo-count such as 1 or 0.5 has been suggested to add to the original counts to eliminate 0s. Since the majority of the counts may be 0s for microbiome data, adding even a small pseudo-count could have a dramatic effect on the geometric means of most OTUs. To circumvent the problem, GMPR reverses the order of the two steps of RLE. The first step is to calculate r_{ij}, which is the median count ratio of nonzero counts between sample i and j,

$$r_{ij} = \prod_{j=1}^{n} \underset{k \in \{1, \cdots, q\} | c_{ki} \cdot c_{kj} \neq 0}{Median} \left\{ \frac{c_{ki}}{c_{kj}} \right\},$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

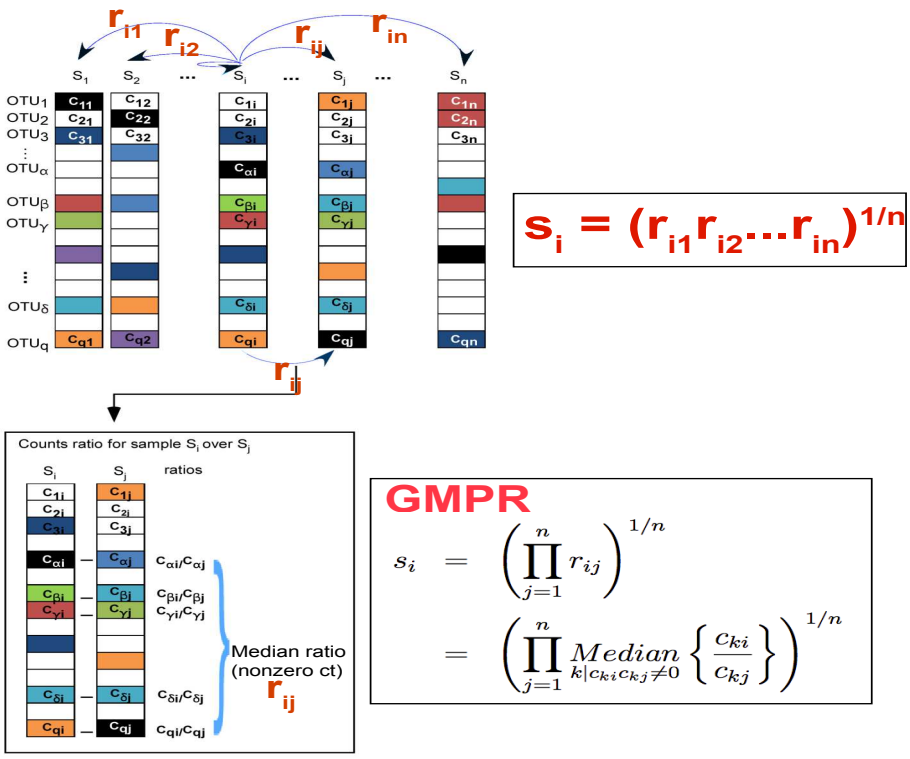

Figure 1. GMPR starts with pairwise comparisons (upper). Each pairwise comparison calculates the median abundance ratio of those common OTUs between the pair of samples (lower). The pairwise ratios are then synthesized into a final estimate.

The second step is to calculate the size factor s_i for a given sample i as

$$s_i = \left(\prod_{j=1}^{n} r_{ij} \right)^{1/n}.$$

Figure 1 illustrates the procedure of GMPR. The basic strategy of GMPR is that we conduct the pairwise comparison first and then combine the pairwise results to obtain the final estimate. Using this strategy, we do not need to calculate the geometric mean for each OTU as implemented in RLE. Although only a small number of OTUs (or none) are shared across all samples due to severe zero inflation, for every pair of samples, they usually share many OTUs. Thus, for pairwise comparison, we focus on these common OTUs that are observed in both samples to have a reliable inference of the abundance ratio between samples. We then synthesize the pairwise abundance ratios using a geometric mean to obtain the size factor. It should be noted that GMPR is a general method, which could be applied to any type of sequencing data in principle.

The R implementation of GMPR could be accessed by `https://github.com/jchen1981/GMPR`.

RESULTS

We compare GMPR to competing normalization methods including CSS, RLE, RLE+ (RLE with pseudo-count 1), TMM, TMM+ (TMM with pseudo-count 1) and TSS. The details of how to estimate the size factors using each normalization method are described in Box 1.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

Box 1. Normalization methods compared in the analysis.

- GMPR (Geometric Mean of Pairwise Ratios): The size factors for all samples are calculated by `GMPR` described in the Method section.
- CSS (Cumulative Sum Scaling): The size factors for all samples are calculated by applying `newMRexperiment`, `cumNorm` and `normFactors` in Bioconductor package metagenomeSeq. Normalized read counts are obtained by dividing the raw read counts by the size factors.
- RLE (Relative Log Expression): The size factors for all samples are calculated by the `calcNormFactors` with the parameter set as "RLE" in the edgeR Bioconductor package. The scaled size factors are obtained by multiplying the size factors with the total read count. Normalized read counts are obtained by dividing the raw read counts by the scaled size factors.
- RLE+ (Relative Log Expression plus pseudo-counts): The scaled size factors for all samples are calculated in the same way as RLE, except that each data entry is added with a pseudo-count 1. Normalized read counts are obtained by dividing the raw read counts by the scaled size factors.
- TMM (Trimmed Mean of M values): The size factors for all samples are calculated by the `calcNormFactors` function with the parameter set as "TMM" in the edgeR Bioconductor package. The scaled size factors are obtained by multiplying the size factors with the total read count. Normalized read counts are obtained by dividing the raw read counts by the scaled size factors.
- TMM+ (Trimmed Mean of M values plus pseudo-counts): The scaled size factors for all sample are calculated in the same way as TMM, except that each data entry is added with a pseudo-count 1. Normalized read counts are obtained by dividing the raw read counts by the scaled size factors.
- TSS (Total Sum Scaling): The size factors are taken to be the total read counts. Normalized read counts are obtained by dividing the raw read counts by the size factors.

We study the performance of GMPR using both simulated and real OTU datasets. In simulated datasets, we study its robustness to differential and outlier OTUs as well as the effect on the performance of differential abundance analysis of OTU data. In real datasets, since we do not know the ground truth, we focus on its ability to reduce the inter-sample variability as well as the ability to increase the reproducibility of the normalized taxa abundances.

Simulation: GMPR is robust to differential and outlier OTUs

We first use a perturbation-based simulation approach to evaluate the performance of normalization methods, focusing on their robustness to differentially abundant OTUs and sample-specific outlier OTUs. The idea is that we first simulate the counts from a common distribution so that the number of total counts is a proxy of the "true" library size. Next, we perturb the counts in different ways and apply different normalization methods on the perturbed counts and evaluate the performance based on the correlation between estimated size factor and "true" library size. Specifically, we generate zero-inflated count data based on a Dirichlet-multinomial model with known library sizes (Chen and Li, 2013). The mean and dispersion parameters of Dirichlet-multinomial distribution are estimated from the COMBO dataset (n=98) after filtering out rare OTUs with prevalence less than 10% (q=397) (Wu et al., 2011). The library sizes are also drawn from those of the COMBO data. To investigate the effect of sparsity (the number of zeros), OTU counts are simulated with different zero percentages ($\sim$ 60%, 70% and 80%) by adjusting the dispersion parameter. A varying percentage of OTUs (0%, 1%, 2%, 4%, 8%, 16% and 32%) are perturbed in each set of simulation, with varying strength of perturbation. The counts c_{ij} of perturbed OTUs are changed to $\sqrt{c_{ij}}$ or c_{ij}^2 for strong perturbation and $0.25c_{ij}$ or $4c_{ij}$ for moderate perturbation. Finally, size factors for all methods are estimated and the Spearman's correlation between the estimated size factors and "true" library sizes is calculated. The simulation is repeated 50 times and the average Spearman's correlation is reported.

We employ two perturbation approaches where we decrease/increase the abundances of a "fixed" or

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

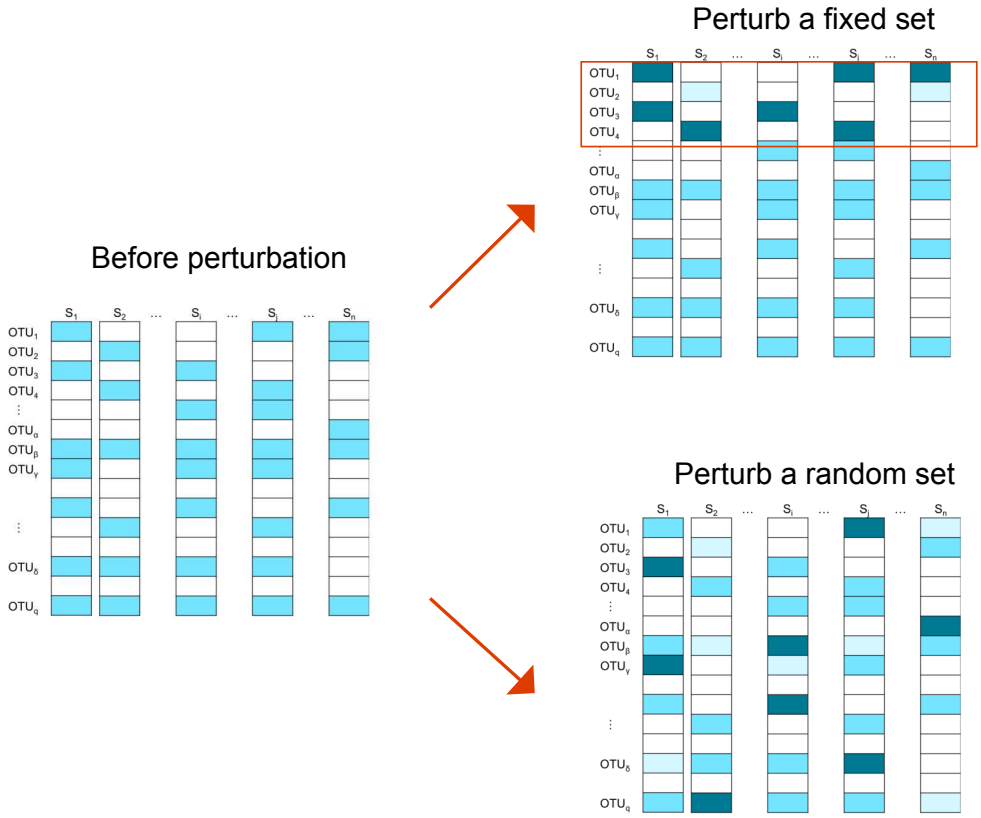

Figure 2. Illustration of the simulation strategy. In the "fixed" perturbation approach, the abundances of the same set of OTUs are decreased/increased for all samples, reflecting differentially abundant OTUs under certain conditions such as disease state. In the "random" perturbation approach, each sample has a random set of OTUs perturbed with a random direction, reflecting the sample-specific outliers. The darkness of the color indicates the OTU abundance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

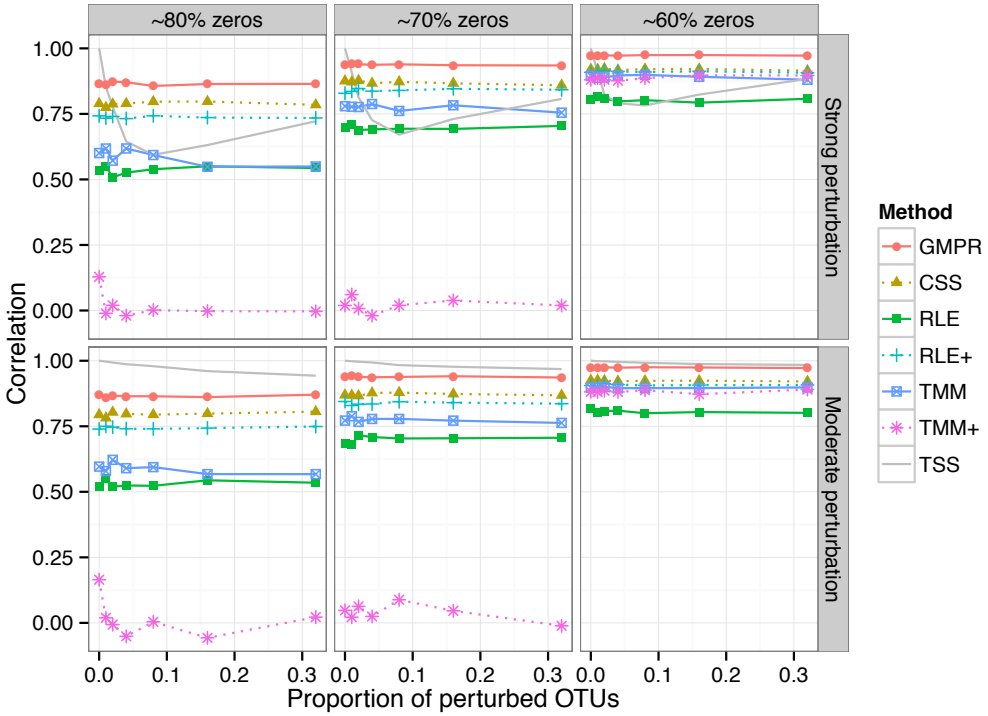

Figure 3. Spearman's correlation between the estimated size factors and the simulated "true" library sizes when a fixed set of OTUs are perturbed. The performance of different normalization methods are compared under different level of zero inflation, percentage of perturbed OTUs and strength of perturbation.

"random" set of OTUs. As shown in Figure 2, in the "fixed" perturbation approach, the same set of OTUs are decreased/increased in the same direction for all samples, reflecting differentially abundant OTUs under a certain condition such as disease state. In the "random" perturbation approach, each sample has a random set of OTUs perturbed with a random direction, mimicking the sample-specific outliers.

In the simulation of "fixed" perturbation (Figure 3), the performance of all methods trends to decrease with the increased zero percentage of counts and strength of perturbation. TSS has excellent performance under moderate perturbation but performs poorly under strong perturbation. GMPR, followed by CSS, consistently outperforms the other methods when the perturbation is strong. When the perturbation is moderate, GMPR is only secondary to TSS when the percentage of zeros is high (80%) and on par with TSS when the percentage of zeros is moderate (70%) or low (60%). For RNA-Seq based methods, TMM performs better than RLE in either strong or moderate perturbation. Though the performance of RLE+ improves by adding pseudo-counts to the OTU data, the size factor estimated by TMM+ merely correlates with true library size when the zero percentage is high (70% and 80%). In contrast, GMPR, together with CSS, performs stable in all cases and GMPR yields better size factor estimate than CSS.

In the "random" perturbation scenario (Figure 4), performance of all methods trends to decrease with the increased zero percentage and strength of perturbation as expected. The performance also decreases with the increased number of perturbed OTUs. Similar to the performance in "fixed" perturbation scenario, TSS has excellent performance under moderate perturbation but performs poorly under strong perturbation. When the perturbation is strong, GMPR, followed by CSS, still outperforms the other methods. RNA-Seq based methods including TMM, TMM+, RLE and RLE+ have similar trend as in "fixed" perturbation. However, different from "fixed" perturbation, the performance of TMM and RLE decreases significantly as the number of perturbed OTUs increases. In contrast, GMPR and CSS are more robust to sample-specific outlier OTUs in all cases and GMPR results in better size factor estimate than CSS.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

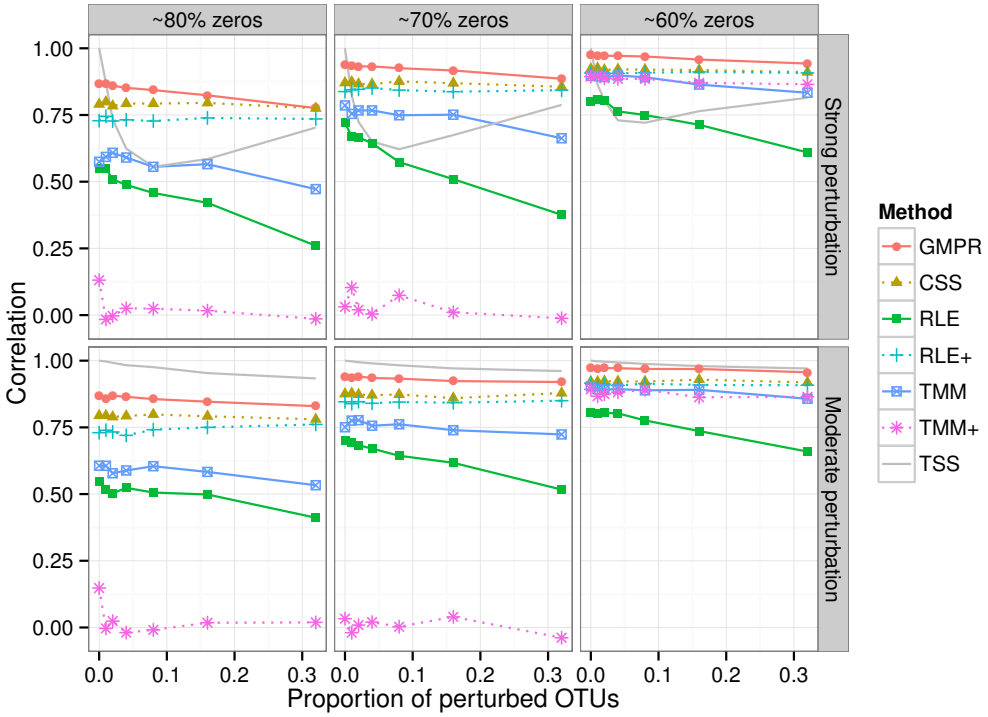

Figure 4. Spearman's correlation between the estimated size factors and the simulated "true" library sizes when a random set of OTUs are perturbed. The performance of different normalization methods are compared under different level of zero inflation, percentage of perturbed OTUs and strength of perturbation.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

Simulation: GMPR improves the performance of differential abundance analysis

In the previous section, we demonstrate that GMPR could better recover the "true" library size in presence of differentially abundant OTUs or sample-specific outlier OTUs. In this section, with a different perspective, we show that the robustness of GMPR method translates to a better false positive control and higher statistical power in the context of differential abundance analysis (DAA), where the aim is to detect differentially abundant OTUs between two sample groups. To achieve this end, we use DESeq2 and edgeR to perform DAA on the OTU table (McMurdie and Holmes, 2014) and we compare the performance of these two methods using their native normalization methods (RLE for DESeq2 and TMM for edgeR) to that using the GMPR method. We evaluate the performance based on the actual false discovery rate (FDR) control after the Benjamini-Hochberg FDR control procedure is applied (Benjamini and Hochberg, 1995) and ROC analysis, where the true positive rate is plotted against false positive rate at different P-value cutoffs.

We use Zero-inflated Negative Binomial distribution (ZINB) to simulate the microbiome data as more detailedly described in Chen et al. (2017). Let c_{ij} be the number of reads from taxon j in the i^{th} sample, the ZINB has the following probability distribution function

$$f_{zinb}(c_{ij}|p_{ij},\mu_{ij},\phi_{ij}) = p_{ij}\cdot I_0(c_{ij}) + (1-p_{ij})\cdot f_{nb}(c_{ij}|\mu_{ij},\phi_{ij}), \tag{1}$$

which is a mixture of a point mass at zero (I_0) and a negative binomial (f_{nb}) distribution of the form

$$f_{nb}(c_{ij}|\mu_{ij},\phi_{ij}) = \frac{\Gamma(c_{ij}+\frac{1}{\phi_{ij}})}{\Gamma(c_{ij}+1)\Gamma(\frac{1}{\phi_{ij}})}\cdot\left(\frac{\phi_{ij}\mu_{ij}}{1+\phi_{ij}\mu_{ij}}\right)^{c_{ij}}\cdot\left(\frac{1}{1+\phi_{ij}\mu_{ij}}\right)^{\frac{1}{\phi_{ij}}} \tag{2}$$

There are three parameters prevalence(p_{ij}), abundance(μ_{ij}) and dispersion(ϕ_{ij}), which fully captures the zero-inflated and dispersed count data. We generate the simulated datasets based on the estimated parameters from the COMBO dataset after filtering out rare OTUs (n=98, q=397). We simulate two sample groups of size 49 each and randomly select 5% of OTUs as differential OTUs by either multiplying or dividing a factor of 4 in one group. We then apply DESeq2 and edgeR on the simulated datasets with either their native normalization or GMPR normalization. We denote DESeq2-GMPR, DESeq2-RLE, edgeR-GMPR and edgeR-TMM as the four method-normalization combinations. For each approach, the P-values are calculated for each OTU and corrected for multiple testing using the BH procedure. The observed FDR is calculated as

$$\frac{FP}{max(1, FP+TP)},$$

where FP and TP are the number of false and true positives respectively.

As shown in Figure 5A, although all approaches have slightly elevated FDRs relative to the nominal levels, the observed FDRs of DAA methods using GMPR normalization are closer to the nominal levels than those of DAA methods with their native normalization. In terms of the power of different methods based ROC analysis (Figure 5B), DESeq2-GMPR achieves a higher AUC (Area Under the Curve) than DESeq2-RLE and edgeR-GMPR has a higher AUC than edgeR-TMM. Overall, GMPR has better FDR control and higher power invariant to the DAA method used.

Real data: GMPR reduces the inter-sample variability of normalized abundances

We next evaluate various normalization methods using 38 gut microbiome datasets from16S rDNA sequencing of the stool samples (Table 1). These real datasets are retrieved from qiita database (https://qiita.ucsd.edu/) with a sample size larger than 50. The 38 datasets come from different species of both invertebrates and vertebrates as well as a wide range of biological conditions. We choose stool samples because the stool microbiota is more studied than that from other body sites.

For the real data, it is not feasible to calculate the correlation between estimated size factors and "true" library sizes as done for simulations. As an alternative, we use the inter-sample variability as a performance measure since an appropriate normalization method will reduce the variability of the normalized OTU abundances (raw counts divided by the size factor) due to different library sizes. A similar measure has been used in the evaluation of normalization performance for microarray data (Fortin et al., 2014). We use the traditional variance as the metric to assess inter-sample variability. For each method, the variance of the normalized abundance of each OTU across all samples is calculated and the median of the variances of all OTUs or stratified OTUs (based on their prevalence) is reported for each

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

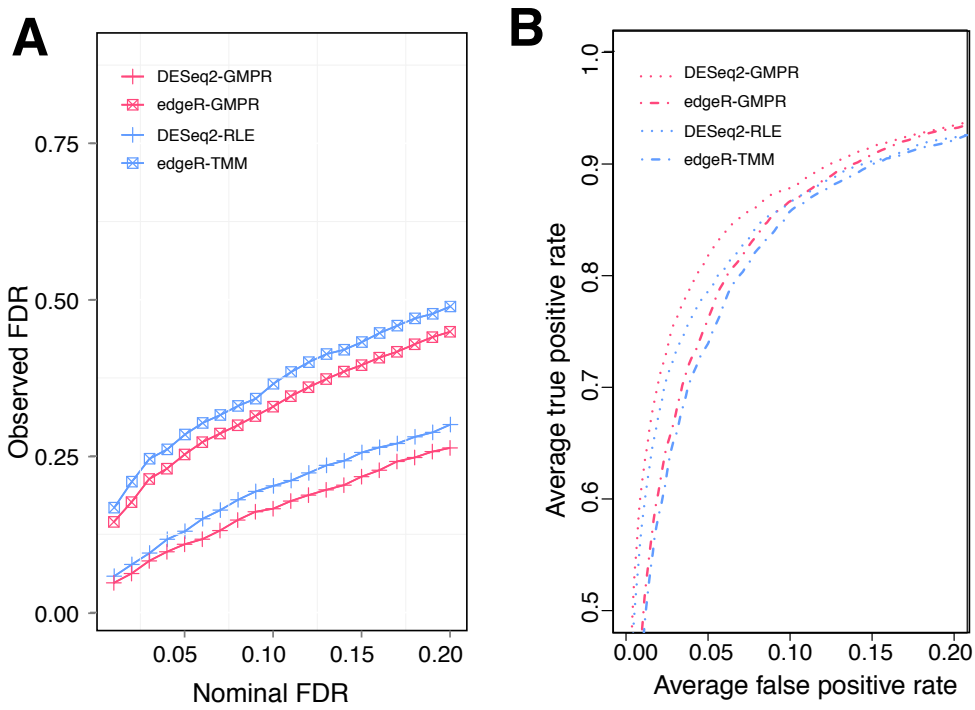

Figure 5. Comparison of the performance of different normalization methods in differential abundance analysis. A. Ability to control the FDR. The observed FDR is plotted against the nominal FDR level. B. ROC curves when 5% random OTUs are differentially abundant between two groups.

study. For each study, all methods are ranked based on these median variances. The distributions of their ranks across these 38 studies for each method are depicted in Figure 6. A higher ranking (lower values in the box plot) indicates a better performance in terms of minimizing inter-sample variability.

In Figure 6, we could see that GMPR achieves the best performance with top ranks in 22 out of 38 datasets, followed by CSS, which tops in 7 datasets (Table 2). This result is consistent with the simulation studies, where GMPR and CSS are overall more robust to perturbations than other methods. Moreover, GMPR consistently performs the best for reducing the variability of OTUs at different prevalence level. It is also noticeable that the inter-sample variability is the largest without normalization (RAW) and TSS does not perform well for a large number of studies. As expected, RLE only works for 8 out of 38 datasets due to a large percentage of zero read counts. By adding pseudo-counts, RLE+ improves the performance significantly compared to RLE. However, there is not much improvement of TMM+ compared to TMM. To see if the difference is significant, we performed paired Wilcoxon signed-rank tests between the ranks of the 38 datasets obtained by GMPR and by any other methods. GMPR achieves significantly better ranking than other methods (P-value$<$0.05 for all OTUs or stratified OTUs). Overall, GMPR achieves the best performance in terms of minimizing inter-sample variability.

Real data: GMPR improves the reproducibility of normalized abundances

When replicates are available, we could evaluate the performance of normalization based on its ability to reduce between-replicate variability. Normalization will increase the reproducibility of the normalized OTU abundances. In this section, we compare the performance of different normalization methods based on a reproducibility analysis of a dataset from the fecal stability study, which aims to compare the temporal stability of different stool collection methods (Sinha et al., 2016). In this study, 20 healthy volunteers provided the stool samples and these samples were subject to different treatment methods. The stool samples were then frozen immediately or after storage in ambient temperature for one or four days for the study of the stability of the microbiota. Each sample had two to three replicates for each condition and thus we could perform reproducibility analysis based on the replicate samples. We evaluate the reproducibility for the “no additive” treatment method, where the stool samples are left untreated.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

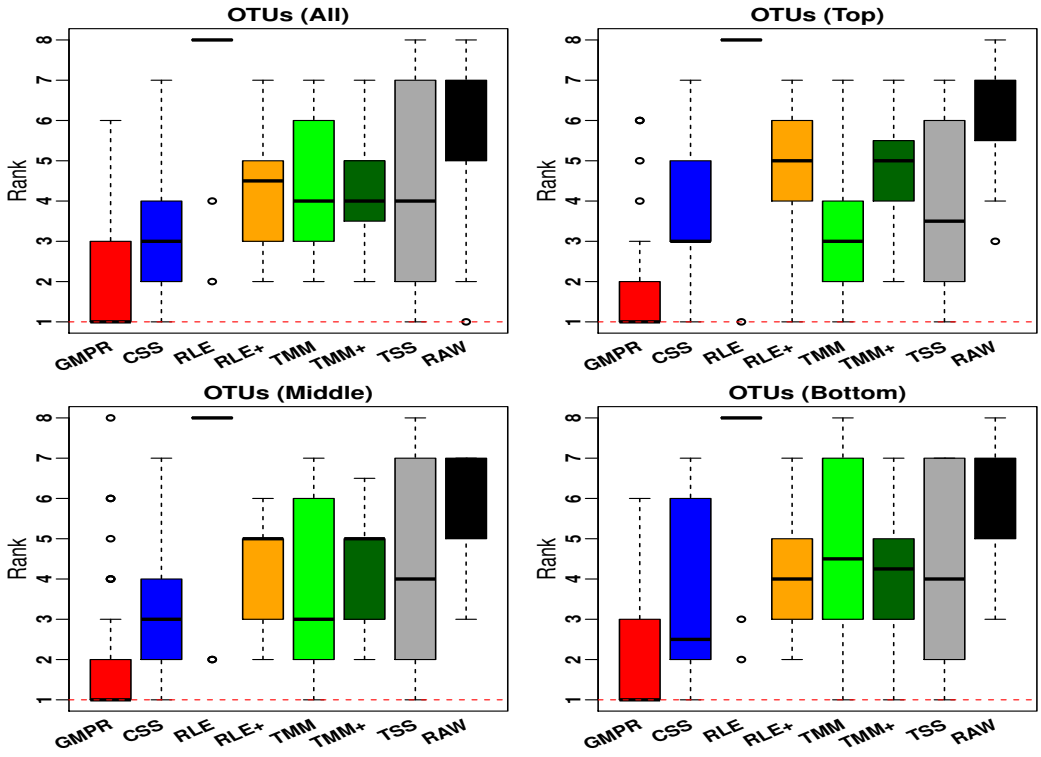

Figure 6. Comparison of normalization methods in reducing inter-sample variability of normalized OTU abundances based on 38 real stool microbiome datasets. Distribution of the ranks for the medians of the variances over the 38 datasets. The median is calculated over all OTUs or OTUs of different prevalence level (Top, middle and bottom)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

Table 1. 38 gut microbiome datasets (stool samples) from qiita ($n \geq 50$)

	study.object	study.ID	sample.size
1	infant gut fecal samples	101	63
2	infant fecal samples	10293	144
3	human and canine fecal samples	10394	1535
4	mice fecal sample	10469	391
5	human fecal samples	1561	52
6	human(HIV) fecal samples	1700	58
7	Cape Buffalo fecal samples	1736	642
8	Skin, oral and fecal samples	1841	3735
9	stool New-Onset Crohns Disease	1998	284
10	TwinsUK population fecal samples	2014	1081
11	Saliva, skin and fecal samples from ICU patients	2136	554
12	human fecal samples	455	92
13	human fecal samples	457	91
14	mice fecal microbiota	654	212
15	pregnant women fecal samples	867	1007
16	human infant gut	10297	85
17	monkey gut	10315	199
18	Grant gazelle gut	10323	768
19	human gut western Oklahoma	10342	58
20	human gastrointestinal gut	1070	118
21	human gut	1189	436
22	zebrafish gut	1192	50
23	Asian primates gut	1453	318
24	cow hindgut	1621	192
25	mice gut	1634	294
26	monkey gut	1696	172
27	bat gut	1734	96
28	colobine primates gut	2182	167
29	human gut and salivary	2202	820
30	bat gut	2338	192
31	human gut and mouthand skin	449	602
32	humann gut microbiome (mouse samples)	452	160
33	humann gut microbiome (mouse samples)	456	158
34	human gastrointestinal	492	77
35	human gut (obese and lean twins)	77	281
36	human gut	850	528
37	freshwater fish slime and gut	940	288
38	Iguanas gut	963	100

Under this condition, certain bacteria will grow in the ambient temperature and we thus expect a low agreement between replicates after four-day ambient temperature storage.

We conduct the reproducibility analysis on the core genera, which are present in more than 75% samples (a total of 26 genera are assessed). We first estimate the size factors based on the OTU-level data and the genus-level counts are divided by the size factors to produce normalized genus-level abundances. Intraclass correlation coefficients (ICC) is used to quantify the reproducibility for the genus-level normalized abundances. The ICC is defined as,

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2}$$

where σ_b^2 represents the biological variability, i.e., sample-to-sample variability and σ_ε^2 represents the replicate-to-replicate variability. We calculate the ICC for 26 core genera for "day 0" (immediately frozen) and "day 4" (frozen after four-day storage) respectively. The ICCs are estimated using the R package "ICC" based on the mixed effects model. An ICC closer to one indicates excellent reproducibility.

Figure 7 shows that the reproducibility of the genera in "day 0" has higher reproducibility than "day 4" regardless of the normalization method used since reproducibility decreases as certain bacteria grow as time elapses. While all the methods have resulted in comparable ICCs for "day 0", GMPR has achieved

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

Table 2. The frequency of 1st rank in the 38 real stool microbiome datasets.

	GMPR	CSS	RLE	RLE+	TMM	TMM+	TSS	RAW
OTU(All)	22	7	0	0	0	0	8	1
OTUs(Top)	23	3	1	1	3	0	7	0
OTUs(Middle)	20	8	0	0	1	0	9	0
OTUs(Bottom)	20	8	0	0	2	2	6	0

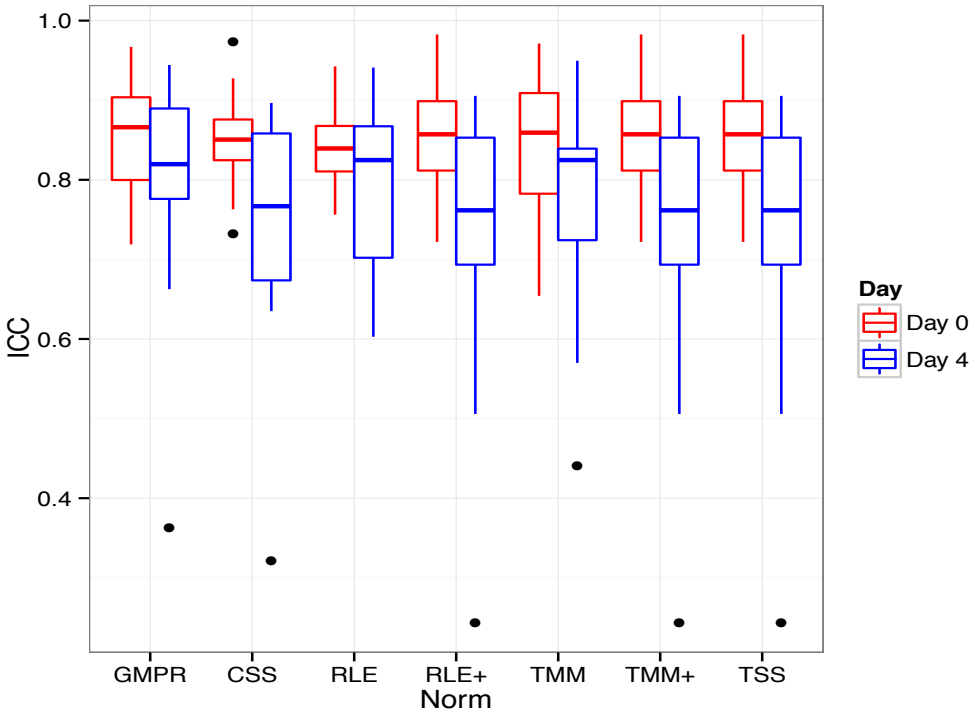

Figure 7. ICC as a measurement for reproducibility is calculated for 26 core genera normalized by different methods for “day 0” and “day 4” respectively.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

higher ICCs for "day 4" than the rest methods. Sinha et al. (2016) showed that most taxa were relatively stable over 4 days and only a small group of taxa (mostly *Gammaproteabacteria*) displayed a pronounced growth at ambient temperature. This suggests that most of the genera are temporally stable and their "day 4" ICCs should be close to the "day 0" ICCs. However, due to the compositional effect, if the data are not properly normalized, a few fast-growing bacteria will skew the relative abundances of other bacteria, leading to apparently lower ICCs for those stable genera. In contrast, the GMPR method is more robust to differential or outlier taxa as demonstrated by the simulation study, which explains higher ICCs for "day 4" samples.

CONCLUSION AND DISCUSSION

Normalization is a critical step in processing microbiome data, rendering multiple samples comparable by removing the bias caused by variable sequencing depths. Normalization paves the way for the downstream analysis, especially for differential abundance analysis of microbiome data, where proper normalization could reduce the false positive rates due to compositional effects. However, the characteristics of microbiome sequencing data, including over-dispersion and zero inflation, make the normalization a non-trivial task.

In this study, we propose the GMPR method for normalizing microbiome sequencing data by addressing the zero inflation. In one simulation study, we demonstrate GMPR's effectiveness by showing its better performance than other normalization methods in recovering the original library sizes when a subset of OTUs are differentially abundant or when random outlier OTUs exist. In another simulation study, GMPR yields better FDR control and higher power in detecting differentially abundant OTUs. In real data analysis, we show GMPR reduces the inter-sample variability and increases inter-replicate reproducibility of normalized taxa abundances. Overall, GMPR outperforms RNA-Seq normalization methods including TMM and RLE and modified TMM+ and RLE+. It also yields better performance than CSS, which is a normalization method specifically designed for microbiome data. As a general normalization method for zero-inflated sequencing data, GMPR could also be applied to other sequencing data with excessive zeros such as single-cell RNA-Seq data (Vallejos et al., 2017).

Although we demonstrate the use of GMPR method in the context of differentially abundant analysis and reproducibility analysis of taxa abundances, its use may not be limited to these applications. Other applications of GMPR normalization include distance-based statistical methods such as ordination, clustering and PERMANOVA (Caporaso et al., 2010; Chen et al., 2012), where the distance is calculated using the GMPR-normalized data. We note that this strategy only works with weighted distance measures, such as weighted UniFrac distance (Chen et al., 2012), where the taxa abundances are used in the calculation. For unweighted distance measures based on presence/absence information, rarefaction is still recommended to remove/reduce the effect of differing probabilities of being sampled as 0s due to uneven sequencing depths (Thorsen et al., 2016; Weiss et al., 2017).

GMPR is an inter-sample normalization method and has a computational complexity of $O(n^2q)$, where n and q are the number of samples and features respectively. While GMPR calculates the size factors for a typical microbiome dataset ($n < 1000$) in seconds, it does not scale linearly with the sample size. Large samples sizes are increasingly popular for epidemiological study and genetic association study of the microbiome (Robinson et al., 2016; Hall et al., 2017), where tens or hundreds of thousands of samples will be collected to detect weak association signals. For such large sample sizes, GMPR may take a much longer time. A potential strategy for efficient computation under ultra-large sample sizes is to divide the dataset into overlapping blocks, calculate GMPR size factors on these blocks and unify the size factors through the overlapping samples between blocks. To increase the computational efficiency of GMPR for ultra-large sample sizes will be the focus of our future research.

REFERENCES

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., et al. (2011). Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome Biol*, 12(2):R18.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stati. Soc. Series B*, 57(1):289–300.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). Qiime allows analysis of high-throughput community sequencing data. *Nat Methods*, 7(5):335–6.

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–13.

Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., et al. (2017). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, 10.1093/bioinformatics/btx650.

Chen, J. and Li, H. (2013). Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.*, 7(1).

Costea, P. I., Zeller, G., Sunagawa, S., and Bork, P. (2014). A fair comparison. *Nat Methods*, 11(4):359.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Brief Bioinform*, 14(6):671–83.

Fortin, J., Labbe, A., Lemire, M., Zanke, B., Hudson, T., and Fertig, E. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.*, 15(11):503.

Hall, A. B., Tolonen, A. C., and Xavier, R. J. (2017). Human genetic variation and the gut microbiome in disease. *Nat Rev Genet*, 18(11):690–699.

Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of illumina high-throughput rna-seq data. *BMC Bioinformatics*, 16:347.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):e1003531.

Paulson, J., Stine, O., Bravo, H., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat. methods*, 10(12):1200–1202.

Robinson, C. K., Brotman, R. M., and Ravel, J. (2016). Intricacies of assessing the human microbiome in epidemiologic studies. *Ann Epidemiol*, 26(5):311–21.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol.*, 11(3):R25.

Sinha, R., Chen, J., Amir, A., Vogtmann, E., Shi, J., Inman, K. S., Flores, R., Sampson, J., Knight, R., and Chia, N. (2016). Collecting fecal samples for microbiome analyses in epidemiology studies. *Cancer Epidemiol Biomarkers Prev*, 25(2):407–16.

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., et al. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16s rrna gene amplicon data analysis methods used in microbiome studies. *Microbiome*, 4(1):62.

Tsilimigras, M. C. and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol*, 26(5):330–5.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell rna sequencing data: challenges and opportunities. *Nat Methods*, 14(6):565–571.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27.

Wu, G., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y., Keilbaugh, S., et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3417v2 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017