Chapter 16
Performance Evaluation of Normalization Approaches for Metagenomic Compositional Data on Differential Abundance Analysis

Ruofei Du, Lingling An, and Zhide Fang

16.1 Introduction

Classical microbiological research requires microbial culture, by which the studied microbes reproduce in culture medium (Handelsman 2004). However, since a community of microbes (i.e., microbiome) is usually not able to survive under the predetermined laboratory condition, our understanding of microbes at aggregate level had been much hindered (National Research Council 2007). The scenario started to change since mid-1980s when a different approach was innovated (Woese 1987), in which microbiome samples are obtained from the site in situ; DNA contents are extracted and sequenced; sequence alignment is subsequently performed, and then followed by computational or statistical analysis (Wooley et al. 2010). The related study is named metagenomics, especially boosted by the rapid advancement of DNA sequencing technologies in the past decade (Metzker 2010; Bragg and Tyson 2014).

Having the entire genomic DNA or particular DNA contents (e.g., 16S rDNA) sequenced, metagenomic datasets can be classified as whole-genome sequence

R. Du
Biostatistics Shared Resource, University of New Mexico Comprehensive Cancer Center, Albuquerque, NM, USA
e-mail: RDu@salud.unm.edu

L. An
Department of Agricultural and Biosystems Engineering, University of Arizona, Tucson, AZ, USA

Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ, USA

Z. Fang (✉)
Biostatistics Program, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA, USA
e-mail: zfang@lsuhsc.edu

Y. Zhao, D.-G. Chen (eds.), *New Frontiers of Biostatistics and Bioinformatics*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-319-99389-8_16

(WGS) data or marker-gene survey data. They are together termed as metagenomic sequence data, or metagenomic count data in this chapter. The obtained sequence reads can be aligned against a database for taxonomic analysis (e.g., RDP database (Cole et al. 2013)) or functional analysis (e.g., COGs (Tatusov et al. 2003), eggNOGs (Powell et al. 2014) databases). The number of reads aligned to a feature, either a taxonomic unit or a functional family, indicates the abundance level of the feature in a sample. It is often of primary interest to identify the features of which the abundance levels differ between conditions, for example, to find the microbial species more abundantly appeared in a diseased human gut than in a healthy gut (Shreiner et al. 2015). This comparative study is named differential abundance analysis. However, due to the fact that the total amount of DNA undergone sequencing, conventionally referred as to library size, may differ substantially as observed, normalization of library size is inescapable before the differential abundance analysis is performed. Otherwise, a differentially abundant feature may be claimed because of uneven library sizes instead of the difference in the abundance of study interest.

Various normalization methods have also been developed for RNA-Seq data analysis (Dillies et al. 2013). As both metagenomic sequence data and RNA-Seq data share a common structure: the count of reads aligned to a feature (e.g., a gene for RNA-Seq data), there have been suggestions proposed to treat metagenomic sequence data as another variant of RNA-Seq data and simply apply the existing normalization methods for RNA-Seq data to metagenomics data analysis (Fernandes et al. 2014; Anders et al. 2013). Towards differential abundance analysis, McMurdie and Holmes (2014) classified the existing normalization methods widely used for metagenomic count data into three groups: (1) Model/None, in which a parametric model is employed to normalize the data or no normalization is applied in some cases, includes the Upper Quartile (UQ) (Bullard et al. 2010), Relative Log Expression (RLE) (Anders and Huber 2010), Trimmed Mean of M-value (TMM) (Robinson and Oshlack 2010), and Cumulative Sum Scaling (CSS) (Paulson et al. 2013); (2) Rarefied (McMurdie and Holmes 2013), in which samples with library size being less than a specified value will be discarded and the remaining samples will be subsampled such that all library sizes equal to the specified value (detailed later); (3) Proportion, in which raw counts are divided by total library size, is named as Total Sum Scaling (TSS) in this chapter. The UQ normalization shares the same spirit with CSS method, so we do not evaluate UQ method. The basic conclusion McMurdie and Holmes drew from their study is that "both proportions and rarefied counts result in a high rate of false positives in tests for species that are differentially abundant across sample classes" and they suggest that it is fine to use the normalization methods from Model/None group, "In particular, an analysis that models counts with the Negative Binomial—as implemented in DESeq2 or in edgeR with RLE normalization—was able to accurately and specifically detect differential abundance over the full range of effect sizes, replicate numbers, and library sizes that we simulated" (McMurdie and Holmes 2014).

There is increasing evidence that many metagenomic count data may be regarded as samples from the microbial ecosystems, and the count of reads to a feature

indicates the relative abundance (i.e., compositional proportion) of the feature in the ecosystem (Tsilimigras and Fodor 2016; Gloor et al. 2016). Mandal et al. (2015) provided an excellent example explaining the difference between the comparison of abundance across specimens, and that across microbial ecosystems. We summarize that the former is about absolute abundance, while the latter is about relative abundance. Weiss et al. (2017) explicitly pointed out that the metagenomic data from 16S rDNA amplicon sequencing possess the compositional data characteristics, and studied six normalization methods combined with different test approaches for differential abundance analysis. The simulation studies conducted in their paper utilized Multinomial, Dirichlet-multinomial, and Gamma-Poisson distributions. However, as indicated in the same paper, both Multinomial and Dirichlet-multinomial distributions may not be appropriate for metagenomic compositional data as these distributions imply a negative correlation between any pair of the features, while Gamma-Poisson distribution does not impose the simplex (i.e., the relative abundances sum to 1). Adequate simulation criteria are strongly needed for drawing correct conclusions about the performance of normalization methods on metagenomic compositional data.

In this chapter, we adopt a metagenomic dataset to show the ineffectiveness of some normalization methods, list the details of conducting simulation based on the characteristics learned from the dataset, and demonstrate the impact of normalization methods on the differential abundance analysis. We advocate, in order to avoid ineffective normalization, case-by-case simulation should be conducted according to the dataset to be analyzed. We are drawing attention to the research community and calling for normalization methods specially designed for metagenomics compositional data.

16.2 Motivating Example

The NIH Human Microbiome Project (HMP) (https://hmpdacc.org/hmp/ (Peterson et al. 2009)) provides the 16S rDNA sequencing output and the processed datasets, collected from different sites of healthy human bodies. We downloaded the saliva and stool sample data (170 saliva samples vs. 191 stool samples) from http://www.hmpdacc.org/HMQCP/ (last visited on February 28, 2018). The sequencing reads were processed by the bioinformatics tool Quantitative Insights Into Microbial Ecology (QIIME, (Caporaso et al. 2010)). For each taxonomic unit, the coefficients of variation (CV: the ratio of the sample standard deviation over the sample mean) of the counts can be calculated for the saliva and the stool samples, respectively. As the CV is an indication of the level of standardized variation between the samples for a feature, it is expected that after appropriate normalization the CV values from all the features under the same condition will decrease in general since the variation due to unequal library sizes should have been reduced. A subsampled dataset is obtained using the steps: randomly selecting the same number (i.e., 361) of samples from the HMP saliva and stool dataset with replacement, and then removing the

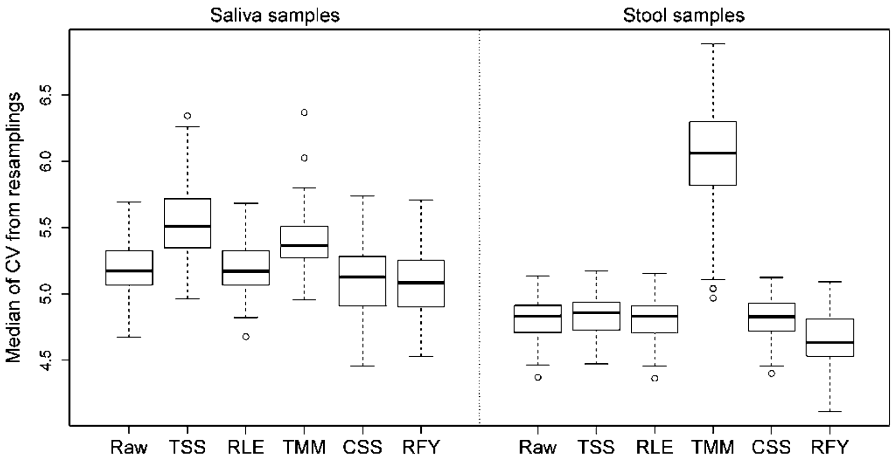

Fig. 16.1 Boxplots of the median values of Coefficients of Variation of the counts in the non-normalized subsampled datasets (Raw), and the normalized subsampled datasets by five different methods from the HMP saliva and stool dataset

duplicated ones. The resampling process repeated one hundred times. Figure 16.1 shows the boxplots of the median CV values of the non-normalized subsampled datasets (Raw), and the normalized subsampled datasets by five different methods. We can see instead of reducing the CV, the TMM normalization has noticeably increased CV values in both saliva and stool samples. This may imply that the TMM normalization is ineffective for the data intended for differential abundance analysis between saliva and stool microbiota. The TSS normalization results in higher CV values for the datasets subsampled from the saliva samples as well. However, it is worth noting that reduced CV itself does not sufficiently mean a good normalization because overreducing sample variation could lead to additional false positives. That is, we cannot conclude that RFY is superior than the other normalizations for this dataset either. This CV analysis on the HMP saliva and stool dataset shows a striking example, which motivated us to investigate how the existing normalization methods perform with metagenomic compositional datasets.

16.3 Data Notation and Methods

A metagenomic dataset can be organized as shown in Table 16.1. A column contains the sequence counts for all the features in a sample; a row lists the counts for a feature across all the samples. For example, y_{ij} denotes the count for feature i from sample j.

With these notations, the steps and the formula of the normalization methods studied in this chapter are briefly introduced as follows.

Table 16.1 Format of a metagenomic dataset of two conditions

	Condition 1			Condition 2		
	Sample 1	$\cdots$	Sample n_1	Sample n_1+1	$\cdots$	Sample n_1+n_2
Feature 1	y_{11}	$\cdots$	y_{1n_1}	y_{1,n_1+1}	$\cdots$	y_{1,n_1+n_2}
Feature 2	y_{21}	$\cdots$	y_{2n_1}	y_{2,n_1+1}	$\cdots$	y_{2,n_1+n_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Feature m	y_{m1}	$\cdots$	y_{mn_1}	y_{m,n_1+1}	$\cdots$	y_{m,n_1+n_2}

TSS (White et al. 2009): The total sum of the counts in a sample serves as the estimate of the library size of the sample. A TSS normalized count is calculated as

$$\tilde{y}_{ij}^{TSS} = \frac{y_{ij}}{\sum_i y_{ij}} N^{TSS},$$

where N^{TSS} is an appropriately chosen normalization constant.

RLE (Anders and Huber 2010): The geometric mean of the counts to a feature from all the samples is first calculated. The ratio of a raw count over the geometric mean to the same feature is then computed. The scale factor of a sample is obtained as the median of the ratios for the sample. A RLE normalized count can be calculated as

$$\tilde{y}_{ij}^{RLE} = y_{ij} / \text{median}_i \left\{ \frac{y_{ij}}{\left(\prod_j y_{ij}\right)^{\frac{1}{n_1+n_2}}} \right\}.$$

TMM (Robinson and Oshlack 2010): The ratio of two observed relative abundances for a feature in two samples is considered to be an estimate of the scale factor between the two samples. The log2 of the ratio is named M value; and the log2 of the geometric mean of the observed relative abundances is called A value. This name convention follows the M and A values given originally in the M-A plot (Yang et al. 2002). That is, for feature i from samples j, l,

$$M_{i(jl)} = \log_2 \frac{y_{ij}/\sum_i y_{ij}}{y_{il}/\sum_i y_{il}}; \quad A_{i(jl)} = \frac{1}{2}\log_2 \left(\frac{y_{ij}}{\sum_i y_{ij}} \frac{y_{il}}{\sum_i y_{il}} \right).$$

The features with specified upper or lower percent of M (default 30%) or A (default 5%) values are trimmed out. The weighted sum of the M values can be used to derive the scale factor,

$$\log_2 \left(SF_{jl}^{TMM} \right) = \frac{\sum_{i \in m_{jl}^{TMM}} \left(w_{i(jl)} M_{i(jl)} \right)}{\sum_{i \in m_{jl}^{TMM}} \left(w_{i(jl)} \right)},$$

where SF_{jl}^{TMM} denotes the scale factor of sample j relative to sample l by TMM method, and m_{jl}^{TMM} denotes the remaining features after the trimming step for the two samples. The weight $w_{i(jl)}$ is computed by,

$$w_{i(jl)} = \frac{\sum_i y_{ij} - y_{ij}}{y_{ij}\sum_i y_{ij}} + \frac{\sum_i y_{il} - y_{il}}{y_{il}\sum_i y_{il}}.$$

After appropriate steps, a TMM normalized count can also be expressed as the quotient of y_{ij} and some attainable value.

CSS (Paulson et al. 2013): For a sample, CSS is defined as the sum of counts that are less than or equal to a percentile, determined by the data. This cumulative sum excludes the raw counts from features that are preferentially amplified, and thus is considered to be relatively invariant across the samples. Using this sum as the scale factor, a CSS normalized count can be calculated as

$$\tilde{y}_{ij}^{CSS} = \frac{y_{ij}}{\sum_{i \in m^{CSS}} \left(y_{ij}\right)} N^{CSS},$$

where N^{CSS} is an appropriately chosen normalization constant, and m^{CSS} denotes the features included in the cumulative summation for the sample.

RFY (McMurdie and Holmes 2013): Rarefying normalization starts with selection of a library size, N^{RFY}. Then any sample, with library size less than N^{RFY}, is considered defective and discarded. For any remaining sample, the features are resampled using their counts as sampling weights. The resampled dataset, or the normalized samples, share the same library size. In this chapter, we use the same criterion as that in McMurdie and Holmes (2014) to set the 15th percentile of total sums of the counts of raw samples as the N^{RFY}. Note that, RFY does not provide an estimate of scale factor of a sample as other normalizations do. In this sense, TSS, RLE, TMM, and CSS are called scaling normalizations, but RFY is not.

16.4 Simulation Study

16.4.1 Parameters and Data Characteristics

Mandal et al. (2015) has made a remarkable comment for metagenomic compositional data analysis: "It is critical to understand what the observed data represent and what statistical parameters are being tested." As discussed in Introduction, in our opinion, the answer to the comment is: metagenomic compositional data should be deemed as samples from the microbial ecosystems, and the read counts to the features should be used as the indication of the relative abundances (i.e., compositional proportions) of the features in the ecosystems. For a statistical test, the relative abundance is the underlying parameter to be compared between

conditions. The relative abundance of feature i for condition k is denoted by $p_i^{(k)}$, subject to the simplex, i.e., $\sum_{i=1}^{m} p_i^{(k)} = 1$.

Through more than a decade of metagenomics research, it has been recognized that metagenomic data possess at least three outstanding characteristics: (1) a great proportion of the features have a sparse count, meaning that the data contain an inflated proportion of zero counts (Paulson et al. 2013; Sohn et al. 2015); (2) the data suffer from the under-sampling issue, that is, more features are found from sample with larger library size, in other words, zero counts could also be associated to library size (Srinivas et al. 2013); (3) the counts are usually overdispersed (McMurdie and Holmes 2014).

16.4.2 Data Simulation

Data simulation encompasses two consecutive steps: learning of real dataset on the characteristics outlined above, and statistical simulation using the parameters learned. To emphasize, both the learning of real dataset and statistical simulation are carried out for each condition separately.

Learning of real dataset. The expectation of y_{ij} is expressed as $\mu_{ij} = \mu_j p_i^{(k)}$, where μ_j is the expectation of the sum of the counts in sample j and is named sample scale here. An estimate of μ_{ij} can be obtained by,

$$\widehat{\mu}_{ij} = \widehat{\mu}_j \cdot \widehat{p}_i^{(k)} = \sum_i y_{ij} \cdot \frac{\sum_{j \in (k)} y_{ij}}{\sum_{i, j \in (k)} y_{ij}},$$

where $j \in (k)$ represents the samples from condition k only. Note that, as an estimate of count, $\widehat{\mu}_{ij}$ is rounded to the nearest integer.

The observed counts, with the same estimated expectation, of all the samples under the same condition, are put together to fit a Negative Binomial (NB) distribution. There is a fitted size parameter of NB distribution from each of the grouped raw counts. This size parameter indicates the level of overdispersion of the counts, which is detailed in Appendix. We will use the average of the fitted size values for the simulation.

After the NB fitting, for the group of observed counts that share the same estimated expectation, the probability of zero can be calculated using the fitted NB distribution. If the observed proportion of zeros is greater than this probability, their difference is recorded as the estimated probability of inflated zero counts for that expectation.

The samples (or columns in Table 16.1) under the same condition are sorted according to the values of $\widehat{\mu}_j$ (i.e., $\sum_i y_{ij}$) from the least to the greatest. Then, for a feature (or a row in Table 16.1), the cumulative sums of the counts from sample 1 to another sample are calculated, i.e., $\sum_{j=1}^{J} y_{ij}$, $J = 1, \ldots n_c$, where n_c is the sample size under that condition. Thus, for a feature, we use the maximum of the $\widehat{\mu}_j$'s, over

the samples (or columns) with the cumulative sums ≤ 3, to estimate the boundary library size of the under-sampling.

Simulation steps. Simulation is carried out for each of the conditions separately as well. First, the $\widehat{\mu}_j$'s from the real dataset are used to build an empirical distribution from which random numbers can be generated and serve as the sample scales (μ_j^{sim}'s) for the simulation. Second, the expectation of count is obtained following $\mu_{ij}^{sim} = \mu_j^{sim} \cdot \widehat{p}_i^{(k)}$. The simulated count ($y_{ij}^{sim}$) is randomly selected from either a zero point, or a random number from the NB distribution with the learned parameter values. Third, in simulated sample *j*, if the estimated boundary size of under-sampling for a feature is greater than $\sum_i y_{ij}^{sim}$, the corresponding count is replaced by zero. R codes for learning of a real dataset and subsequently data simulation are available at a Github webpage https://github.com/rdu2017/Normalization-Evaluation.

16.4.3 Normalization Performance

The purpose of normalization is to adjust all the samples to the same scale for differential abundance analysis. Although it is conventional to say that normalization is for library size, it is essentially the sample scale that needs to be normalized. After normalization, the counts for a feature in different samples under the same condition are assumed to have the same expectation. The expectations are compared between conditions to draw the conclusion for the analysis. Thereupon, the sample scale, the sum of expectations of counts in the sample, needs to be normalized among all the samples. In turn, the relative abundance is compared.

Using the HMP saliva and stool sample data as template, we generated 100 simulated datasets. The four methods (TSS, RLE, TMM, and CSS) were applied for estimation of the sample scales in the normalization. Since the RFY approach does not perform normalization through estimating sample scale, it is not included here. The Pearson correlation coefficient between the estimated sample scales and the true values is calculated to show how well a normalization works. The estimate is better when the coefficient is closer to one. Figure 16.2 displays the boxplots of the coefficients from the 100 simulated datasets. Among these four methods, TMM appears uncompetitive. Both TSS and CSS perform better than RLE, while the median of TSS (0.625) is slightly higher than that of CSS (0.61) but with two times larger standard deviation (0.08 vs. 0.04).

16.4.4 Impact of Normalization on Differential Abundance Analysis

To be able to set true/false differentially abundant features explicitly, we take only the simulated data from one condition, i.e., the stool metagenome. A simulated

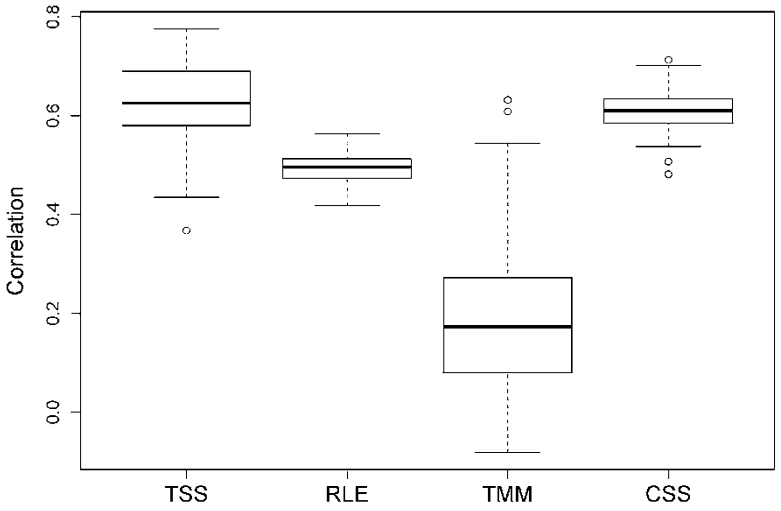

Fig. 16.2 Boxplots of Pearson correlation coefficients of sample scales estimated after different normalization methods and the true values, using the 100 simulated datasets

dataset, containing 191 samples, is randomly partitioned into two smaller datasets with 96 and 95 samples in each. Meanwhile, we intend to keep the compositional characteristics of the data. The quartiles of $p_i^{(k)}$'s are calculated. In the dataset that contains 96 samples, the features (i.e., rows) from the third and fourth quartiles are randomly swapped with the features from the first and second quartiles. By so doing, the two partitioned datasets share 50% true and 50% false differentially abundant features with the compositional structure still maintained. A two-sided T test is first performed to compare the normalized counts for each feature, and followed by the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) for false discovery rate controlling at 0.05 among all the tests. Figure 16.3a shows the boxplots of the observed false discovery rates (FDR) in the analysis output after a normalization procedure. It is noticeable that TMM normalization has much higher FDR, with 46% tests showing FDR greater than 0.05. RFY normalization performs the second worst regarding FDR controlling, with 12% tests having FDR greater than 0.05. As indicated in Fig. 16.3b, the true positive rates (TPR) associated with TMM and non-normalization are lower than TPR associated with the other normalizations. It is clear that an ineffective normalization (TMM here) will discourage the differential analysis in error rate controlling or statistical power, or both.

Our focus is to examine how a normalization impacts subsequent differential analysis. However, it should be pointed out that differential abundance analysis itself is influenced by both normalization and the statistical approach used for analysis. Figure 16.4 shows FDR and TPR on the same shuffled datasets above but analyzed using NB regression approach. It seems NB approach has better TPR rate but worse FDR controlling compared to T test. Nonetheless, TMM still shows ineffectiveness in the NB approach.

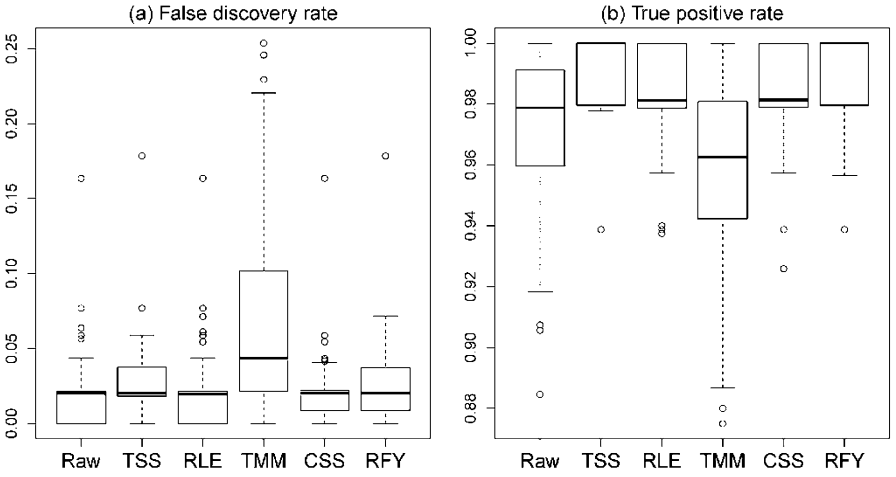

Fig. 16.3 Impact of normalization on differential abundance analysis in both FDR and TPR, by two-sided T test with 100 datasets shuffled from stool metagenome dataset. (**a**) False discovery rate. (**b**) True positive rate

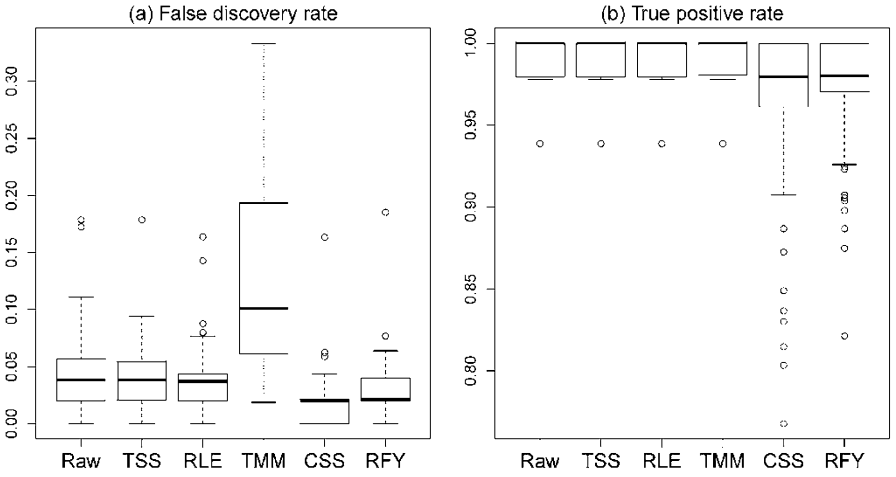

Fig. 16.4 Impact of normalization on differential abundance analysis in both FDR and TPR, by Negative Binomial regression with 100 datasets shuffled from stool metagenome dataset. (**a**) False discovery rate. (**b**) True positive rate

16.5 Discussion

16.5.1 TMM and RLE with Metagenomic Compositional Dataset

For gene expression studies, there is a widely used assumption that the majority of genes do not express differentially between conditions. Many of RNA-Seq normalization methods were developed based on this assumption, including TMM and RLE. The "non-differential" in the assumption is implemented as non-differential absolute abundance after normalization. Subsequent differential analysis is also to compare the normalized counts between conditions, instead of comparing the relative abundances as it is for compositional data. In Appendix, we use hypothetical datasets to explain why TMM and RLE normalizations may not work well with metagenomic compositional dataset. We would then like to suggest using RNA-Seq normalization with caution for metagenomic compositional data analysis.

16.5.2 Simulation Benchmark

Metagenomic studies have been frustrated by lack of good simulation benchmarks (Johnson et al. 2014). Meanwhile, contrary conclusions have been seen from the simulation studies conducted with different criteria (McMurdie and Holmes 2014; Weiss et al. 2017; Costea et al. 2014; Paulson et al. 2014). In our vision, the practice needs improvement from at least two aspects. First, the idea that a simulation study should be designed to apply for overall situations may not be realistic. Instead, a case-by-case simulation practice should be encouraged, based on the real dataset to analyze. Second, in terms of metagenomic compositional data, all the important data characteristics should be included when designing a simulation. Using a convenient statistical distribution is not a good strategy because it may not be capable to reflect the complex in a real dataset.

We suggest that a simulation be carried out for each condition independently for metagenomic compositional data. The distribution of library size, the relative abundance, the overdispersion parameter, the probability of zero count from a zero mass state, and the boundary library size in terms of under-sampling are learned from a real metagenomic dataset. Hopefully, the simulation approach we provide in this chapter can serve as a good basis for building up simulation benchmarks in the research community of metagenomic data analysis.

16.5.3 Novel Normalization Methods Are Needed

As observed, TMM method should be avoided for analysis of the HMP saliva and stool dataset. In Appendix, we also provide a figure showing that the RLE normalization does not work well for the mouse stool metagenomic dataset, which

has been used as the benchmark dataset in the chapter where CSS was introduced (Paulson et al. 2013). From our experience, no matter with real or simulated data, in most situations the CSS does not identify the data-driven percentile, up to which the raw counts will be summed, and then the default value 50th percentile is used. It is questionable to us whether there commonly exists a claimed percentile so that the raw counts are distributed differently lower or greater than it (see Supplementary Figure 1 in Paulson et al. 2013). In addition, there is no specific consideration of the compositional characteristics in the development of CSS. Conceptually, TSS may be fine for compositional data normalization as it uses a count divided by the total sum of the counts of a sample, as an estimate of the relative abundance. However, as many previous studies have shown, TSS is unreliable against the overdispersed counts, under-sampling issue, and aberrant counts in many situations. In a word, novel normalizations, specifically designed for metagenomic compositional data, are highly in demand. Developing novel normalization methods is our future research topics.

Acknowledgements The authors are grateful to two anonymous reviewers for their careful reading of the manuscript and their comments and suggestions. ZF's research is supported by grant U54 GM104940 from the National Institute of General Medical Sciences of the National Institutes of Health, which funds the Louisiana Clinical and Translational Science Center of Pennington Biomedical Research Center. LA's research is partially supported by National Science Foundation [DMS-1222592] and United States Department of Agriculture [Hatch project, ARZT-1360830-H22-138]. RD's research was supported in part by the UNM Comprehensive Cancer Center, a recipient of NCI Cancer Support Grant 2 P30 CA118100-11 (PI: Cheryl L. Willman, MD).

A.1 Appendix

A.1.1 Supplementary Data Distribution

Negative Binomial Distribution. A NB distribution is defined as,

$$P\left(X = x\right) = \frac{\Gamma\left(x + r\right)}{\Gamma\left(\mathrm{r}\right)\mathrm{x!}} p^r (1 - p)^x,$$

where r and p are two parameters, and r is called size parameter. The mean of the NB distribution is,

$$\mu = \frac{r\left(1 - p\right)}{p},$$

and the variance is,

$$V = \frac{r\left(1 - p\right)}{p^2} = \mu + \frac{1}{r}\mu^2.$$

Thus, r indicates the level of overdispersion in the counts.

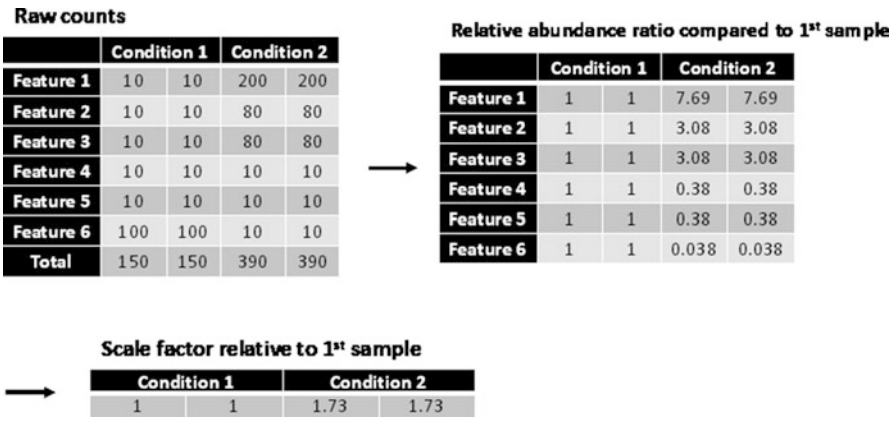

Fig. 16.5 TMM normalization on a hypothetical dataset

A.1.2 Supplementary Illustration of TMM and RLE with Compositional Dataset

For gene expression studies, there is a widely used assumption that the majority of genes do not express differentially between conditions. Many of RNA-Seq normalization methods were developed based on this assumption, including TMM and RLE. The "non-differential" in the assumption is implemented as non-differential absolute abundance after normalization. Subsequent differential analysis is also to compare the normalized counts between conditions, instead of comparing the relative abundances as it is for compositional data.

Focusing on the essence of a normalization procedure, the hypothetical datasets are made of the expectations of counts. For TMM approach, the logarithm function and the weighted sum are not applied since those are designed for reducing the effect of count variation. In Fig. 16.5, the relative abundance ratio, compared to the first sample, is first calculated from the raw counts, i.e., $\frac{y_{ij}/\sum_i y_{ij}}{y_{i1}/\sum_i y_{i1}}$. The trimmed mean of the ratios for each sample, after trimming the largest and smallest values, is used as the scale factor. The true scale factor is 2.6 (390/150), but the output from TMM is 1.73. Figure 16.5 shows a very likely situation for metagenomic compositional data, in which the relative abundances vary largely between conditions. TMM may not work well for such data since it merely relies on the assumption that after normalization most of features should share the same absolute abundance.

For RLE normalization, the geometric mean of the counts to each feature from all the samples is first calculated, see Fig. 16.6. Next, the ratio of a raw count over the mean count for the same feature is computed. The scale factor for a sample is obtained as the median of the ratios for the sample. For this hypothetic dataset, RLE approach does not suggest any normalization adjustment since all the scale factors equal to 1; however, the true library sizes are very different (e.g., 210 vs. 310).

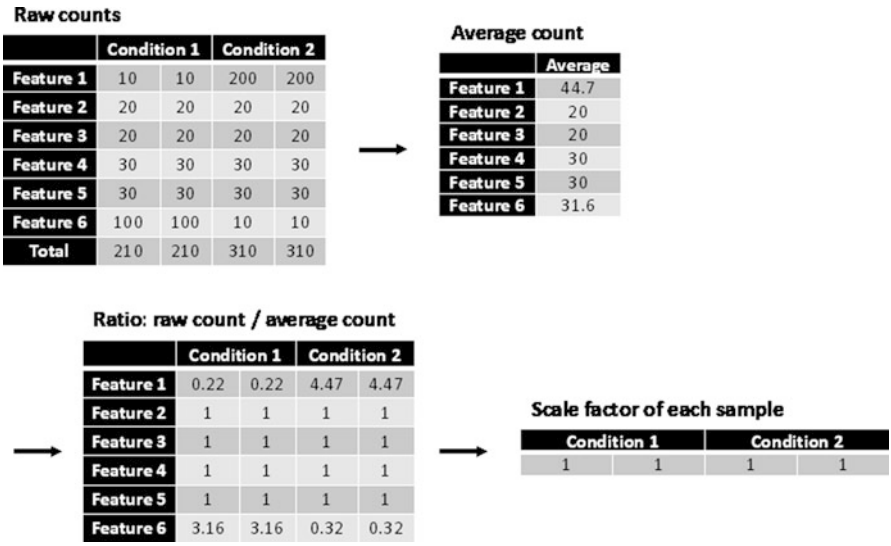

Fig. 16.6 RLE normalization steps on a hypothetical dataset

In Fig. 16.6, it is clear that the scale factor is determined by the absolute count of Feature 2, 3, 4, or 5, instead of the relative abundance of one of those features. A subsequently comparative analysis would reach the conclusion that there is no differential abundance for Feature 2, 3, 4, or 5 between the conditions. However, the relative abundances of the features have altered, for Feature 4 it is 14% and 10% under the two conditions, respectively.

A.1.3 Supplementary Example

Mouse stool metagenomic data. Fresh or frozen adult human fecal microbial communities were transplanted into guts of germ-free C57BL/6J mice. Here, germ-free environment is referred to as mice gut that does not previously expose to microbes. Following the transplanting, 12 recipient mice were fed with a standard low-fat, plant polysaccharide-rich diet for 4 weeks; after that, six mice were switched to take high-fat/high-sugar Western diet for another 6 weeks. Amplification and pyrosequencing of V2 region of 16S rRNA genes were performed periodically to record the changes of microbial community structure of fecal samples of the mice (Turnbaugh et al. 2009). There are 85 samples under condition one (associated to low-fat diet fed mice), and 54 samples under condition two (associated to Western diet fed mice). The bioinformatic tool RDP (Wang et al. 2007) was used to generate the count data, which is featured at species level.Together, there are 52 genera

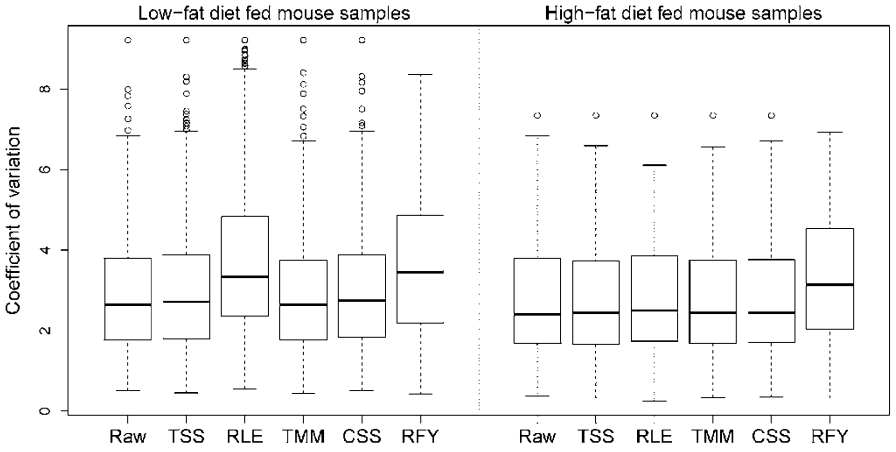

Fig. 16.7 Boxplots of coefficients of variation of counts in the raw data, and the normalized data for the mouse stool metagenomic data

shown under both conditions, and the data is considered to represent low complex metagenomic data. Figure 16.7 demonstrates that RLE and RFY should not be recommended for normalization of the metagenomic data.

References

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

Anders, S., et al. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols, 8*(9), 1765–1786.

Benjamini, Y., & 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)*, 289–300.

Bragg, L., & Tyson, G. W. (2014). Metagenomics using next-generation sequencing. *Environmental Microbiology: Methods and Protocols, 1096*, 183–201.

Bullard, J. H., et al. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics, 11*(1), 94.

Caporaso, J. G., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*(5), 335–336.

Cole, J. R., et al. (2013). Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research, 42*(D1), D633–D642.

Costea, P. I., et al. (2014). A fair comparison. *Nature Methods, 11*(4), 359.

Dillies, M.-A., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics, 14*(6), 671–683.

Fernandes, A. D., et al. (2014). Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome, 2*(1), 15.

Gloor, G. B., et al. (2016). It's all relative: Analyzing microbiome data as compositions. *Annals of Epidemiology, 26*(5), 322–329.

Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews, 68*(4), 669–685.

Johnson, S., et al. (2014). A better sequence-read simulator program for metagenomics. *BMC Bioinformatics, 15*(9), S14.

Mandal, S., et al. (2015). Analysis of composition of microbiomes: A novel method for studying microbial composition. *Microbial Ecology in Health and Disease, 26*(1), 27663.

McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS One, 8*(4), e61217.

McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Computational Biology, 10*(4), e1003531.

Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews Genetics, 11*(1), 31–46.

National Research Council. (2007). *The new science of metagenomics: Revealing the secrets of our microbial planet*. Washington, DC: National Academies Press.

Paulson, J. N., et al. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods, 10*(12), 1200–1202.

Paulson, J. N., Bravo, H. C., & Pop, M. (2014). Reply to: "A fair comparison". *Nature methods, 11*(4), 359–360.

Peterson, J., et al. (2009). The NIH human microbiome project. *Genome Research, 19*(12), 2317–2323.

Powell, S., et al. (2014). eggNOG v4. 0: Nested orthology inference across 3686 organisms. *Nucleic Acids Research, 42*(D1), D231–D239.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology, 11*(3), R25.

Shreiner, A. B., Kao, J. Y., & Young, V. B. (2015). The gut microbiome in health and in disease. *Current Opinion in Gastroenterology, 31*(1), 69.

Sohn, M. B., Du, R., & An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics, 31*(14), 2269–2275.

Srinivas, G., et al. (2013). Genome-wide mapping of gene–microbiota interactions in susceptibility to autoimmune skin blistering. *Nature Communications, 4*, 2462.

Tatusov, R. L., et al. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics, 4*(1), 1.

Tsilimigras, M. C., & Fodor, A. A. (2016). Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Annals of Epidemiology, 26*(5), 330–335.

Turnbaugh, P. J., et al. (2009). The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine, 1*(6), 6ra14.

Wang, Q., et al. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology, 73*(16), 5261–5267.

Weiss, S., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome, 5*(1), 27.

White, J. R., Nagarajan, N., & Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Computational Biology, 5*(4), e1000352.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews, 51*(2), 221.

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Computational Biology, 6*(2), e1000667.

Yang, Y. H., et al. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research, 30*(4), e15.