Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities

Lukas Beule[*] and Petr Karlovsky[*]

Molecular Phytopathology and Mycotoxin Research, Georg-August Universität Göttingen, Göttingen, Germany

[*] These authors contributed equally to this work.

ABSTRACT

Background. Analysis of species count data in ecology often requires normalization to an identical sample size. Rarefying (random subsampling without replacement), which is the current standard method for normalization, has been widely criticized for its poor reproducibility and potential distortion of the community structure. In the context of microbiome count data, researchers explicitly advised against the use of rarefying. Here we introduce a normalization method for species count data called scaling with ranked subsampling (SRS) and demonstrate its suitability for the analysis of microbial communities.

Methods. SRS consists of two steps. In the scaling step, the counts for all species or operational taxonomic units (OTUs) are divided by a scaling factor chosen in such a way that the sum of scaled counts equals the selected total number of counts C_{min}. The relative frequencies of all OTUs remain unchanged. In the subsequent ranked subsampling step, non-integer count values are converted into integers by an algorithm that minimizes subsampling error with regard to the population structure (relative frequencies of species or OTUs) while keeping the total number of counts equal C_{min}. SRS and rarefying were compared by normalizing a test library representing a soil bacterial community. Common parameters of biodiversity and population structure (Shannon index *H'*, species richness, species composition, and relative abundances of OTUs) were determined for libraries normalized to different size by rarefying as well as SRS with 10,000 replications each. An implementation of SRS in R is available for download (https://doi.org/10.20387/BONARES-2657-1NP3).

Results. SRS showed greater reproducibility and preserved OTU frequencies and alpha diversity better than rarefying. The variance in Shannon diversity increased with the reduction of the library size after rarefying but remained zero for SRS. Relative abundances of OTUs strongly varied among libraries generated by rarefying, whereas libraries normalized by SRS showed only negligible variation. Bray–Curtis index of dissimilarity among replicates of the same library normalized by rarefying revealed a large variation in species composition, which reached complete dissimilarity (not a single OTU shared) among some libraries rarefied to a small size. The dissimilarity among replicated libraries normalized by SRS remained negligibly low at each library size. The variance in dissimilarity increased with the decreasing library size after rarefying, whereas it remained either zero or negligibly low after SRS.

Submitted 16 March 2020
Accepted 1 July 2020
Published 3 August 2020

Corresponding author
Lukas Beule,
lukas.beule@agr.uni-goettingen.de

Academic editor
Craig Moyer

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.9593**

How to cite this article Beule L, Karlovsky P. 2020. Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities. *PeerJ* **8**:e9593 http://doi.org/10.7717/peerj.9593

Conclusions. Normalization of OTU or species counts by scaling with ranked subsampling preserves the original community structure by minimizing subsampling errors. We therefore propose SRS for the normalization of biological count data.

Subjects Biodiversity, Bioinformatics, Microbiology, Statistics, Population Biology
Keywords Species count, Microbiome, Amplicon sequencing, Count data normalization, Scaling with ranked subsampling (SRS), Rarefying, Alpha diversity, Beta diversity, Relative abundance

INTRODUCTION

Species counts are fundamental data in studies of ecology and biological diversity. A specific kind of count data, used in studies of the microbiome using next generation sequencing (NGS), are counts of nucleotide sequences that represent operational taxonomic units (OTUs). The so-called amplicon sequencing by NGS became the key technique for the exploration of microbial communities inhabiting diverse environments, such as deep-sea sediments (e.g., *Sogin et al., 2006*), soils (e.g., *Gilbert, Jansson & Knight, 2014*), and the human gut (e.g., *Yatsunenko et al., 2012*). Amplicon sequencing by NGS is also increasingly popular in studies of invertebrate diversity (*Hajibabaei et al., 2011*; *Carew et al., 2013*; *Morinière et al., 2016*; *Vivien, Lejzerowicz & Pawlowski, 2016*). Accumulation of these data motivated the development of bioinformatic tools and pipelines for their processing. Recently, it has been shown that the choice of bioinformatic tools can affect the results and, in some cases, even lead to different interpretation of the results (*Siegwald et al., 2019*). Therefore, the choice of analysis tools should be taken into account when comparing microbiome studies (*Allali et al., 2017*).

In studies of microbial communities by NGS, samples are represented by libraries, which consist of DNA fragments amplified by PCR and attached to adapters required for the sequencing. Multiplex sequencing, which is sequencing pooled libraries, in a single sequencing run, is widely used to lower sequencing costs. A disadvantage of multiplexing is that the number of sequences obtained per library (sample) can span orders of magnitude (*McMurdie & Holmes, 2014*). Comparative analysis requires identical sample size, therefore microbiome count data are commonly normalized to the same total count per library. Over half a century ago, *Sanders (1968)* proposed random subsampling without replacement, designated 'rarefying', to this end. Since then, rarefying has been used for the normalization of species count data in ecology as well as for NGS data in microbiology. For libraries with counts above a selected threshold, a subsample from each library is generated by randomly picking reads without replacement until the selected number of counts is reached. Although rarefying has become the standard tool in microbiome data analysis (*Weiss et al., 2017*), its disadvantages have been recognized (*McMurdie & Holmes, 2014*; *Weiss et al., 2017*; *Willis, 2019*). For example, *McMurdie & Holmes (2014)* demonstrated that rarefying is statistically inadmissible and should not be used. More recently, *Willis (2019)* pointed at the strong bias in alpha diversity estimates for unequal or rarefied microbiome count data. This is because rare OTUs may be overrepresented or underrepresented in libraries normalized to a small size by rarefying. The growing number of normalization methods indicates that the issue of normalization has not been conclusively solved yet. It is evident that any kind of normalization leads to loss of information that should be avoided if possible. Most diversity indices and statistical tests used in community analysis however do not account for the effect of library size; therefore they require normalization of libraries to the same size.

An alternative normalization to rarefying is scaling, which adjusts the size of all samples to the same value by multiplying the counts by a constant factor. Simple scaling preserves the relative frequencies of OTUs but it also keeps the number of OTUs unchanged, preserving the disparity between large and small libraries: a larger number of sampled individuals or sequence reads likely contains a larger number of species or OTUs and thus possess higher alpha diversity (e.g., species richness). Simple scaling does not compensate for the effect of the sample size or library size on species richness. In the analysis of microbial communities, differences in library size mainly originate from unequal pooling of PCR products prior to sequencing. In order to pool PCR products from individual samples in equimolar amounts (e.g., *Kozich et al., 2013*), DNA concentrations are commonly determined by UV spectroscopy, fluorescence spectroscopy, real-time PCR or digital PCR (*Nakayama et al., 2016*; *Robin et al., 2016*). Although some of these methods offer high accuracy (*Robin et al., 2016*), identical library size across samples cannot be achieved. Therefore, normalization of read counts will remain inevitable for comparative analyses requiring equal library size. Here we introduce a normalization method for biological count data called scaling with ranked subsampling (SRS).

METHODS

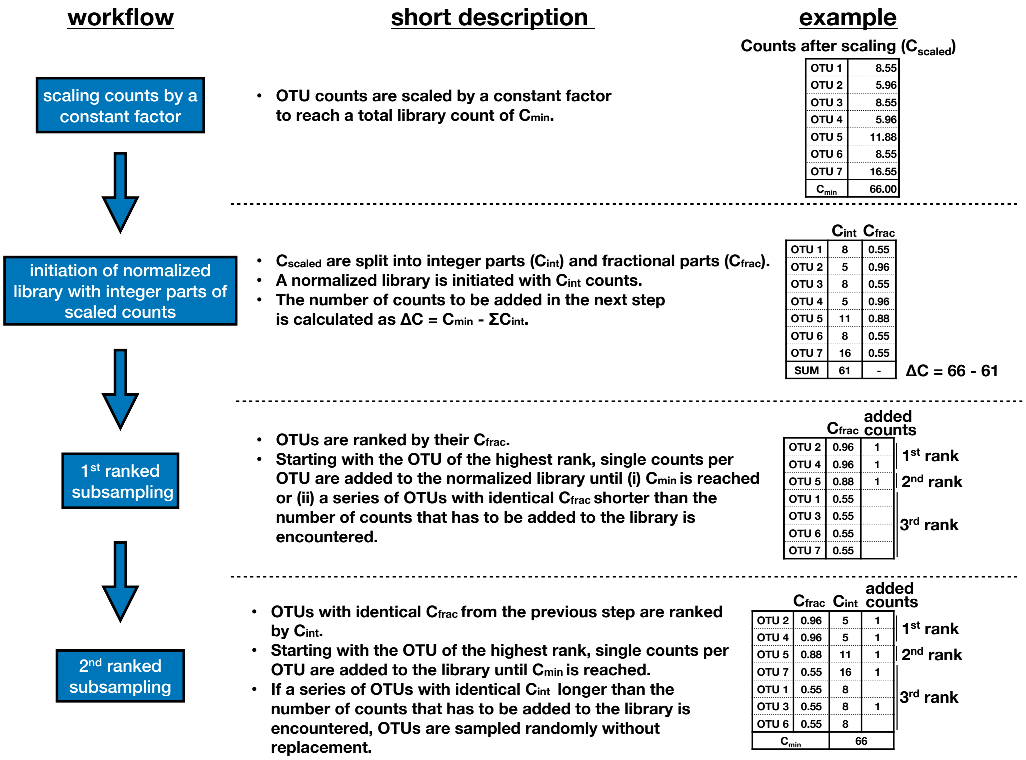
Normalization methods

Rarefying

Rarefying was conducted using the 'rrarefy'-function in the 'vegan' R-package v2.5-6 (*Oksanen et al., 2019*). The function randomly subsamples OTU counts within each library without replacement until the selected number of counts C_{min} is achieved. Rarefying was performed in the R environment v3.6.1 (*R Core Team, 2017*).

Scaling with ranked subsampling (SRS)

The normalization by SRS reduces the number of counts in each sample in such a way that (i) the total count equals C_{min}, (ii) each removed OTU is less or equally abundant than any preserved OTU, and (iii) the relative frequencies of OTUs remaining in the sample after normalization are as close as possible to the frequencies in the original sample. The algorithm consists of two steps (Fig. 1). In the first step, the counts for all OTUs are divided by a scaling factor chosen in such a way that the sum of the scaled counts (C_{scaled} with integer or non-integer values) equals C_{min}. The relative frequencies of all OTUs remain unchanged. In the second step, the non-integer count values are converted into integers by an algorithm that we dub ranked subsampling (Fig. 1). The scaled count C_{scaled} for each OTU is split into the integer-part C_{int} by truncating the digits after the decimal separator ($C_{\mathrm{int}} = \mathrm{floor}(C_{\mathrm{scaled}})$) and the fractional part C_{frac} ($C_{\mathrm{frac}} = C_{\mathrm{scaled}}$ - C_{int}). Since $\Sigma C_{\mathrm{int}} \leq C_{\mathrm{min}}$, additional $\Delta C = C_{\mathrm{min}}$ - ΣC_{int} counts have to be added to the library to reach the


workflow
short description
example
scaling counts by a constant factor
initiation of normalized library with integer parts of scaled counts
1st ranked subsampling
2nd ranked subsampling

- OTU counts are scaled by a constant factor to reach a total library count of C_{min}.

Counts after scaling (C_{scaled})

OTU 1	8.55
OTU 2	5.96
OTU 3	8.55
OTU 4	5.96
OTU 5	11.88
OTU 6	8.55
OTU 7	16.55
C_{min}	66.00

- C_{scaled} are split into integer parts (C_{int}) and fractional parts (C_{frac}).
- A normalized library is initiated with C_{int} counts.
- The number of counts to be added in the next step is calculated as $\Delta C = C_{min} - \Sigma C_{int}$.

	C_{int}	C_{frac}
OTU 1	8	0.55
OTU 2	5	0.96
OTU 3	8	0.55
OTU 4	5	0.96
OTU 5	11	0.88
OTU 6	8	0.55
OTU 7	16	0.55
SUM	61	-

$\Delta C = 66 - 61$

- OTUs are ranked by their C_{frac}.
- Starting with the OTU of the highest rank, single counts per OTU are added to the normalized library until (i) C_{min} is reached or (ii) a series of OTUs with identical C_{frac} shorter than the number of counts that has to be added to the library is encountered.

	C_{frac}	added counts	
OTU 2	0.96	1	1st rank
OTU 4	0.96	1	
OTU 5	0.88	1	2nd rank
OTU 1	0.55		3rd rank
OTU 3	0.55		
OTU 6	0.55		
OTU 7	0.55		

- OTUs with identical C_{frac} from the previous step are ranked by C_{int}.
- Starting with the OTU of the highest rank, single counts per OTU are added to the library until C_{min} is reached.
- If a series of OTUs with identical C_{int} longer than the number of counts that has to be added to the library is encountered, OTUs are sampled randomly without replacement.

	C_{frac}	C_{int}	added counts	
OTU 2	0.96	5	1	1st rank
OTU 4	0.96	5	1	
OTU 5	0.88	11	1	2nd rank
OTU 7	0.55	16	1	3rd rank
OTU 1	0.55	8		
OTU 3	0.55	8	1	
OTU 6	0.55	8		
C_{min}		66		

Figure 1 Workflow of scaling with ranked subsampling (SRS). SRS consists of two steps. In the first step, the counts for all OTUs (operational taxonomic untis) are divided by a scaling factor chosen in such a way that the sum of the scaled counts (C_{scaled} with integer or non-integer values) equals C_{min}. In the second step, the non-integer count values are converted into integers by an algorithm that we dub ranked subsampling. The scaled count C_{scaled} for each OTU is split into the integer part C_{int} by truncating the digits after the decimal separator ($C_{int} = \text{floor}(C_{scaled})$) and the fractional part C_{frac} ($C_{frac} = C_{scaled} - C_{int}$). Since $\Sigma C_{int} \leq C_{min}$, additional $\Delta C = C_{min} - \Sigma C_{int}$ counts have to be added to the library to reach the total count of C_{min}. This is achieved as follows. OTUs are ranked in the descending order of their C_{frac} values. Beginning with the OTU of the highest rank, single count per OTU is added to the normalized library until the total number of added counts reaches ΔC and the sum of all counts in the normalized library equals C_{min}. When the lowest C_{frag} involved in picking ΔC counts is shared by several OTUs, the OTUs used for adding a single count to the library are selected in the order of their C_{int} values. This selection minimizes the effect of normalization on the relative frequencies of OTUs. OTUs with identical C_{frag} as well as C_{int} are sampled randomly without replacement.

Full-size DOI: 10.7717/peerj.9593/fig-1

total count of C_{min}. This is achieved as follows. OTUs are ranked in the descending order of their C_{frac} values, which lie in the open interval (0, 1). Beginning with the OTU of the highest rank, single count per OTU is added to the normalized library until the total number of added counts reaches ΔC and the sum of all counts in the normalized library equals C_{min}. For example, if $\Delta C = 5$ and the seven top C_{frac} values are 0.96, 0.96, 0.88, 0.55, 0.55, 0.55, and 0.55, the following counts are added: a single count for each OTU with C_{frac} of 0.96; a single count for the OTU with C_{frac} of 0.88; and a single count each for two OTUs among those with C_{frac} of 0.55. When the lowest C_{frag} involved in picking ΔC counts is shared by several OTUs, the OTUs used for adding a single count to the library are selected in the order of their C_{int}values. This selection minimizes the effect of normalization on the relative frequencies of OTUs. OTUs with identical C_{frag} as well as C_{int} are sampled randomly without replacement. An R script that enables the reproduction of this study as well as the test library used in the present study are available at the BonaRes Repository: https://doi.org/10.20387/BONARES-T40J-7VAG.

An implementation of SRS in R is also available at the BonaRes Repository: https://doi.org/10.20387/BONARES-2657-1NP3.

Test library

The library used for the evaluation of our normalization method represents a soil bacterial microbial community sampled at an agricultural field in Germany. The library is part of a dataset consisting of 60 samples that were sequenced in a single multiplex sequencing run. Total soil DNA was extracted and amplified using the primer pair 341F (5′-CCT ACG GGN GGC WGC AG-3′)/785R (5′-GAC TAC HVG GGT ATC TAA KCC-3′) (*Herlemann et al., 2011*), and sequenced using the Illumina MiSeq Reagent Kit v3 (2× 300 bp) (Illumina, San Diego, CA, USA). Data processing in the QIIME 2 environment (v2019.10) (*Bolyen et al., 2019*) included denoising, merging, chimera filtering, and removing singletons using dada2 (*Callahan et al., 2016*), clustering amplicon sequence variants, and taxonomic assignment against the SILVA SSU database (release 132) (*Quast et al., 2013*) using VSEARCH (*Rognes et al., 2016*). The library consisted of a total of 162,888 counts distributed among 936 OTUs. The library further contained 1,110 OTUs with zero counts, which were included because these OTUs were found in other samples from the same dataset, and thus contained 2,046 OTUs in total. The Shannon diversity of the test library was 5.645.

Comparison of rarefying and SRS

The 162,888 counts of our test library were normalized to 39 different C_{min} values comprising all integer divisors (factors) 11, 12, 22, 24, 33, 44, 66, 88, 132, 264, 617, 1,234, 1,851, 2,468, 3,702, 4,936, 6,787, 7,404, 13,574, 14,808, 20,361, 27,148, 40,722, 54,296, and 81,444 as well as an exponential series 10, 20, 40, 80, 160, 320, 640, 1,280, 2,560, 5,120, 10,240, 20,480, 40,960, and 81,920. Both rarefying and SRS were used, each with 10,000 replications. Two alpha diversity measures were calculated for each replication: Shannon index *H*' and species richness. The Shannon index was calculated using the 'diversity'-function in the 'vegan' R-package v2.5-6 (*Oksanen et al., 2019*). Species richness was determined employing the 'specnumber'-function in the 'vegan' R-package v2.5-6. Both alpha diversity measures were calculated in the R environment v3.6.1 (*R Core Team, 2017*).

The two methods were further compared by artificially raising the total counts of the test library while keeping the relative frequencies of OTUs and alpha diversity constant. To this end, each OTU count was multiplied by 100 and the resulting library of 16,288,800 counts was normalized to 2.5×10^5, 5×10^5, 7.5×10^5, 1×10^6, 2.5×10^6, 5×10^6, 7.5×10^6, and 1×10^7 counts using both rarefying and SRS methods with 10,000 replications each. Two alpha diversity measures were calculated and compared to the known alpha diversity of the original library.

The effect of rarefying and SRS on species composition and their implications for beta diversity were evaluated by determining the Bray–Curtis index of dissimilarity for all pairs of normalized library replications. As in the investigation of the effect of normalization on the alpha diversity described above, the test library was normalized to 39 different C_{min} (see above) using both rarefying and SRS, each with 10,000 replications. The Bray–Curtis index of dissimilarity among all normalized library replications at each C_{min} was determined for both normalization methods. The Bray–Curtis index of dissimilarity was calculated using the 'vegdist'-function in the 'vegan' R-package version 2.5-6 (*Oksanen et al., 2019*).

In order to examine changes in the relative abundances of OTUs, the test library was normalized to 1×10^3, 1×10^4, and 1×10^5 counts using both rarefying and SRS with 10,000 replications each. The OTUs of the non-normalized library were ranked in a descending order of counts, every 50th OTU starting at the top rank was selected and its relative abundance in all replications of the normalized libraries was determined.

RESULTS

Normalization of the test library

SRS showed on average greater alpha diversity as compared to rarefying (Figs. 2A, 2B). The variance of the diversity measures was consistently lower after normalization by SRS as compared to rarefying across all tested C_{min} (Figs. 2C, 2D). No variation of the Shannon index and species richness was observed after SRS. For normalization by rarefying, the variance of the Shannon diversity increased as C_{min} decreased (Figs. 2C, 2D). Variance × the species richness after normalization by rarefying was bell-curve-shaped (Fig. 2D).

Normalization of a library with artificially raised counts

To confirm our results from the normalization of our test library, we multiplied all counts of our test library by 100 (162,888 × 100 = 16,288,800) -which does not affect alpha diversity- and then normalized the library to different C_{min} above its initial size of 162,888 counts. Ideally, alpha diversity after normalization would remain unchanged in all replications (zero variance). The Shannon diversity of libraries normalized by SRS differed only marginally from the Shannon diversity of the original library (Fig. S1A). Rarefying underestimated or overestimated the Shannon diversity in an extent growing with decreasing C_{min} (Fig. S1A); on the average, Shannon diversity was slightly underestimated. SRS returned the species richness of the original library at all selected C_{min}, whereas rarefying underestimated species richness by up to 9 species (Fig. S1B). Libraries normalized by SRS showed no variance for both diversity measures (Figs. S1C, S1D). After rarefying, the variance increased with decreasing C_{min} for both diversity measures (Figs. S1C, S1D).

Effect of normalization on species composition and implications for beta diversity

The species composition was evaluated by determining the beta diversity as the Bray–Curtis index of dissimilarity among replications of normalized libraries. Ideally, the index of dissimilarity among replications of the same library would be zero, corresponding to identical species composition. Across 10,000 replications of normalized libraries by rarefying, Bray–Curtis dissimilarity values above 0.5 were observed for the lowest 17 C_{min} (10, 11, 12, 20, 22, 24, 33, 40, 44, 66, 80, 88, 132, 160, 264, 320, and 617 counts) (Fig. 3A).

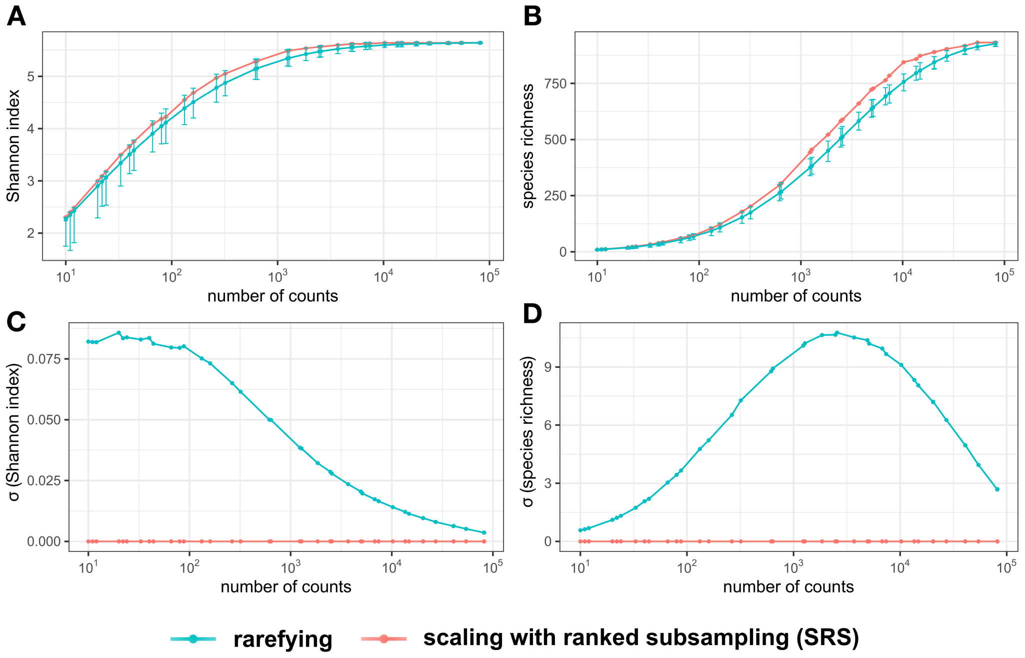


Figure 2 Alpha diversity measures (Shannon index *H'* (A) and species richness (B) of the test library and their standard deviation (*σ*) (C, D) normalized by rarefying or SRS. The sampled numbers of counts were 10, 11, 12, 20, 22, 24, 33, 40, 44, 66, 80, 88, 132, 160, 264, 320, 617, 640, 1,234, 1,280, 1,851, 2,468, 2,560, 3,702, 4,936, 5,120, 6,787, 7,404, 10,240, 13,574, 14,808, 20,361, 20,480, 27,148, 40,722, 40,960, 54,296, 81,444, and 81,920. Means of 10,000 replications (A, B) are represented by points. The minimum and maximum alpha diversity among all 10,000 replications are represented by error bars (A, B). SRS, scaling with ranked subsampling.

Full-size DOI: 10.7717/peerj.9593/fig-2

From these 17 C_{min}, the index of dissimilarity for some pairs of libraries normalized by rarefying to the first nine C_{min} values was one, which showed that some replications did not share any OTUs. After rarefying, the variance in species composition increased with decreasing C_{min}, whereas this was not observed for SRS (Fig. 3B). Differences in species composition among replications of libraries normalized by SRS were only observed when random subsampling of OTUs with the lowest rank of C_{frac} was necessary. For SRS, the maximum dissimilarity found across all replications of all C_{min} was 3.125×10^{-3} (found at C_{min} of 640) (Fig. 3A).

Evaluation of the relative abundance of OTUs

All OTUs above zero counts in the test library showed varying relative abundance among replications of libraries normalized by rarefying to all three selected C_{min} (1,000, 10,000, and 100,000) (Fig. 4). In contrast, the maximum standard deviation after normalization by SRS amounted to 0.001% relative abundance (OTU '1913' at a C_{min} of 1×10^5 counts) (Figs. 4E–4F). When normalized by rarefying to 1,000 counts, the relative abundance of our most abundant OTU (3.186% relative abundance in the non-normalized library) varied by factor of 4.6 (1.2 to 5.5% relative abundance) (Fig. 4A). Furthermore, our 51st most abundant OTU (OTU '348'; 0.404% relative abundance in the non-normalized library) was

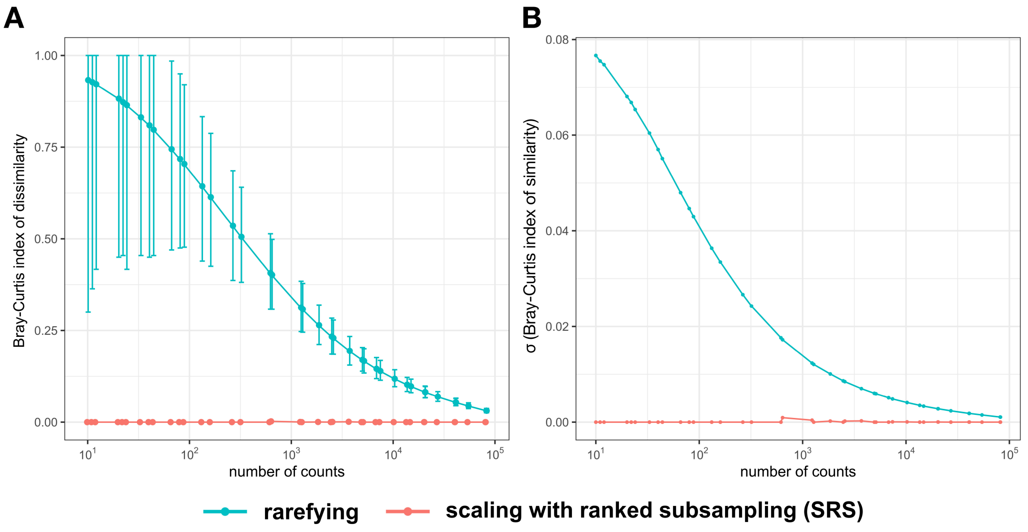


Figure 3 Bray–Curtis index of dissimilarity (A) among 10,000 replications of the normalized test library and its standard deviation (σ) (B) normalized by rarefying or SRS. The sampled numbers of counts were 10, 11, 12, 20, 22, 24, 33, 40, 44, 66, 80, 88, 132, 160, 264, 320, 617, 640, 1,234, 1,280, 1,851, 2,468, 2,560, 3,702, 4,936, 5,120, 6,787, 7,404, 10,240, 13,574, 14,808, 20,361, 20,480, 27,148, 40,722, 40,960, 54,296, 81,444, and 81,920. Means of 10,000 replications (A) are represented by points. The minimum and maximum dissimilarity among all 10,000 replications are represented by error bars (A). SRS, scaling with ranked subsampling.

Full-size DOI: 10.7717/peerj.9593/fig-3

removed from some normalized libraries, whereas it reached 1.2% relative abundance in other replications after rarefying to 1,000 counts (Fig. 4A). Overall, the variance in relative abundance increased with decreasing C_{min} when rarefying was used, whereas this was not observed for SRS (Fig. 4).

DISCUSSION

It is well established that primer choice and library preparation can cause biases in microbiome studies that use NGS technologies (*Van Dijk, Jaszczyszyn & Thermes, 2014*; *Schirmer et al., 2015*; *Tedersoo & Lindahl, 2016*). In addition, a number of studies reported that the choice of bioinformatic tools used to process the data can affect the results (*cf. Plummer et al., 2015*; *Allali et al., 2017*; *López-García et al., 2018*; *Siegwald et al., 2019*). Among these tools, the normalization of microbiome count data is much debated. Rarefying has become the standard procedure for normalization (*Weiss et al., 2017*), although it is statistically inadmissible (*McMurdie & Holmes, 2014*). In the present study, we introduced SRS as an alternative to rarefying.

Our results demonstrated that SRS has greater reproducibility and accuracy than rarefying when alpha diversity measures (Shannon index *H*' and species richness) were investigated (Fig.2, S1). This was particularly true when the library size differed by multiple orders of magnitude (Fig. 2, S1), which is not uncommon in microbiome studies (*McMurdie & Holmes, 2014*). Additionally, we observed a strong variation in the relative abundance of OTUs among library replicates normalized by rarefying (Fig. 4). Again, the variance

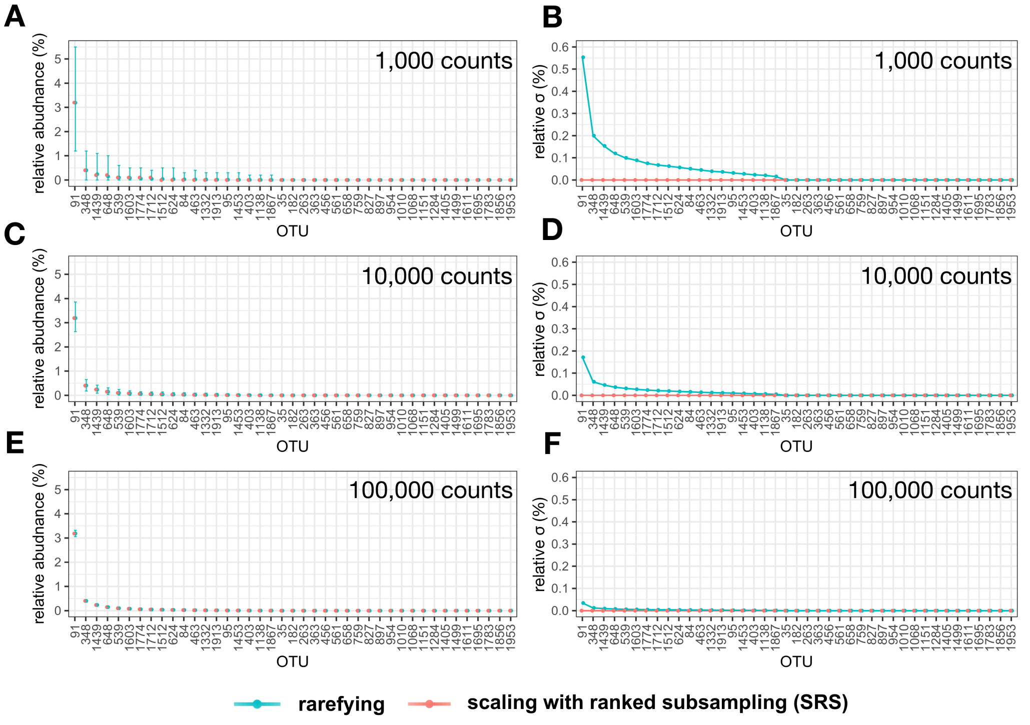


Figure 4 Relative abundance (%) of selected operational taxonomic units (OTUs) at varying library size (A, C, E) and their standard deviation (σ) (B, D, F) normalized by rarefying or SRS. The sampled number of counts were 1×10^3, 1×10^4, and 1×10^5. Means of 10,000 replications (A, C, E) are represented by points. The minimum and maximum relative abundance among all OTUs in all 10,000 replications are represented by error bars (A, C, E). SRS, scaling with ranked subsampling.

Full-size DOI: 10.7717/peerj.9593/fig-4

increased with the reduction of the library size (Fig. 4). Rarefying uses random subsampling without replacement which follows the hypergeometric distribution (*Simberloff, 1972*). Therefore, for an OTU occurring with the frequency f in the original library with N counts, the variance (var) of its abundance a after rarefying to C_{min} is:

$$\mathrm{var}(a) = C_{min} \times f \times (1 - f) \times \frac{(N - C_{min})}{(N - 1)}$$

For the most abundant OTU in Fig. 4 and C_{min} of 1,000, var(a) equals 30.65, corresponding to a relative standard deviation of 0.554% (*cf.* Fig. 4B). This variance results from a subsampling error incurred by rarefying. SRS largely eliminates the subsampling error.

The purpose of ranking OTUs in the last step of SRS in the order of their C_{frag} and if necessary by C_{int} is to minimize distortion of the species/OTU composition. When OTUs with identical C_{frag} values were picked randomly instead of according to the order of their C_{int}, the variance of alpha diversity in libraries normalized to the same C_{min} slightly increased but never reached values comparable to rarefying (Fig. S2).

Due to the law of large numbers, the variance of relative frequencies of OTUs, alpha diversity measures and other parameters after rarefying are expected to grow with decreasing C_{min}. Figure 2D shows, however, that the variance of species richness reaches a maximum at a medium library size, asymptotically approaching zero at both very large and very small libraries (in Fig. S1D this effect is not apparent because only libraries of relatively large size have been analyzed). The drop of variance of species richness for small libraries is caused by a systematic error due to normalization to low counts. With the diminishing library size, the number of different OTUs that can be obtained by random subsampling declines; the variance in species richness declines accordingly. Concomitantly, differences in species composition among replicates of libraries normalized by rarefying are expected to grow with decreasing library size. Comparison of libraries normalized by rarefying to the same size confirmed this expectation (Fig. 3). In contrast, the variation in species composition among libraries normalized by SRS was either zero or negligibly low (Fig. 3). The reproducibility of data normalization and the preservation of the original community structure (OTU frequencies) is crucial for the determination of beta diversity among samples.

Our results support the conclusion of *McMurdie & Holmes (2014)* that rarefying should not be used to normalize microbiome count data. The reason is that random subsampling is the source of variance, which is superimposed on the biological and technical variance. The use of random subsampling in SRS is limited to a fraction of counts that have to be added to the sum of counts scaled and rounded down to integers in order to reach the desired library size. A complex combination of circumstances has to occur for random subsampling to be used in SRS: several OTUs have to share both the integer part and the decimal fraction of their scaled frequencies; in a list of OTUs ranked by frequencies, these OTUs have to appear before the desired total number of counts is reached; and the number of counts that is needed to fill the normalized library is lower than the number of these OTUs. As long as at least one of these conditions is not fulfilled, SRS does not use random subsampling with replacement and replicates of the normalized library are identical. Zero variance of diversity measures for replicates of most libraries in Figs. 2 to 4 is the consequence. If random subsampling is used, the relative abundance of the affected OTUs will be vary by at most a single count. As a consequence, the relative abundance of a rare OTU will be affected more than the relative abundance of a dominant OTU. Therefore, the effect of random subsampling in SRS is expected to be negligible in studies with a library size above 1,000 counts, unless some OTUs are removed while other are kept. Principally, libraries with a high proportion of rare OTUs cannot be normalized to lower integer counts in such a way that all OTUs as well as their frequencies are preserved. In analogy to quantization error in signal processing, no mathematical procedure can circumvent the loss of information due to downscaling counts to integer values. On this background, we believe that SRS is currently the most adequate method for the normalization of species count data and OTU libraries representing microbial communities.

CONCLUSION

SRS method for the normalization of species count data minimizes the subsampling error. In contrast to rarefying, common parameters assessed in studies of biodiversity and population structure (alpha diversity, species composition, and relative abundance of

OTUs) calculated from data normalized by SRS were highly reproducible and the original community structure (OTU frequencies) was preserved. We therefore propose SRS for the normalization of biological count data.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was financed by the German Federal Ministry of Education and Research (BMBF) in the framework of the Bonares-SIGNAL project (funding codes: 031A562A and 031B0510A). This work was support by the Open Access Publication Funds of the Göttingen University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
German Federal Ministry of Education and Research (BMBF): 031A562A, 031B0510A.
Open Access Publication Funds of the Göttingen University.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Lukas Beule conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Petr Karlovsky conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

An R script that enables the reproduction of this study as well as the test library used in the present study are available at the BonaRes Repository: https://doi.org/10.20387/BONARES-T40J-7VAG.

An implementation of SRS in R is also available at the BonaRes Repository: https://doi.org/10.20387/BONARES-2657-1NP3.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.9593#supplemental-information.

REFERENCES

Allali I, Arnold JW, Roach J, Cadenas MB, Butz N, Hassan HM, Koci M, Ballou A, Mendoza M, Ali R, Azcarate-Peril MA. 2017. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiology* **17**:194–1 DOI 10.1186/s12866-017-1101-8.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodrguez AM, Chase J, Cope EK, Silva RD, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vzquez-Baeza Y, Vogtmann E, Hippel Mvon, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* **37**:852–857 DOI 10.1038/s41587-019-0209-9.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes S. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods* **13**:581–583 DOI 10.1038/nmeth.3869.

Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA. 2013. Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology* **10**:45 DOI 10.1186/1742-9994-10-45.

Gilbert JA, Jansson JK, Knight R. 2014. The Earth Microbiome project: successes and aspirations. *BMC Biology* **12**:69 DOI 10.1186/s12915-014-0069-1.

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ. 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using River Benthos. *PLOS ONE* **6**:e17497 DOI 10.1371/journal.pone.0017497.

Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. 2011. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME Journal* **5**:1571–1579 DOI 10.1038/ismej.2011.41.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing Platform. *Applied and Environmental Microbiology* **79**:5112–5120 DOI 10.1128/AEM.01043-13.

López-García A, Pineda-Quiroga C, Atxaerandio R, Pérez A, Hernández I, García-Rodríguez A, González-Recío O. 2018. Comparison of Mothur and QIIME for the analysis of rumen microbiota composition based on 16S rRNA amplicon sequences. *Frontiers in Microbiology* **9** DOI 10.3389/fmicb.2018.03010.

McMurdie PJ, Holmes S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. *PLOS Computational Biology* **10**:e1003531 DOI 10.1371/journal.pcbi.1003531.

Morinière J, De Araujo BC, Lam AW, Hausmann A, Balke M, Schmidt S. 2016. Species identification in malaise trap samples by DNA barcoding based on NGS technologies and a scoring matrix. *PLOS ONE* **11**:e0155497 DOI 10.1371/journal.pone.0155497.

Nakayama Y, Yamaguchi H, Einaga N, Esumi M. 2016. Pitfalls of DNA quantification using dna-binding fluorescent dyes and suggested solutions. *PLOS ONE* **11**:e0150528 DOI 10.1371/journal.pone.0150528.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara B, Simpson G, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2019. The vegan package. *Available at https://cran.r-project.org/web/packages/vegan/vegan.pdf* (accessed on 28 January 2020).

Plummer EL, Twin J, Bulach D, Garl SM, Tabrizi SN. 2015. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *Journal of Proteomics & Bioinformatics* **8**:283–291 DOI 10.4172/jpb.1000381.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**:D590–D596 DOI 10.1093/nar/gks1219.

R Core Team. 2017. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Robin JD, Ludlow AT, LaRanger R, Wright WE, Shay JW. 2016. Comparison of DNA quantification methods for next generation sequencing. *Scientific Reports* **6**:1–10 DOI 10.1038/srep24067.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**:e2584 DOI 10.7717/peerj.2584.

Sanders HL. 1968. Marine benthic diversity: a comparative study. *The American Naturalist* **102**:243–282 DOI 10.1086/282541.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* **43**:e37–e37 DOI 10.1093/nar/gku1341.

Siegwald L, Caboche S, Even G, Viscogliosi E, Audebert C, Chabé M. 2019. The impact of bioinformatics pipelines on microbiota studies: does the analytical microscope affect the biological interpretation? *Microorganisms* **7**:393 DOI 10.3390/microorganisms7100393.

Simberloff D. 1972. Properties of the rarefaction diversity measurement. *The American Naturalist* **106**:414–418 DOI 10.1086/282781.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored rare biosphere.

Proceedings of the National Academy of Sciences of the United States of America **103**:12115–12120 DOI 10.1073/pnas.0605127103.

Tedersoo L, Lindahl B. 2016. Fungal identification biases in microbiome projects. *Environmental Microbiology Reports* **8**:774–779 DOI 10.1111/1758-2229.12438.

Van Dijk EL, Jaszczyszyn Y, Thermes C. 2014. Library preparation methods for next-generation sequencing: tone down the bias. *Experimental Cell Research* **322**:12–20 DOI 10.1016/j.yexcr.2014.01.008.

Vivien R, Lejzerowicz F, Pawlowski J. 2016. Next-Generation sequencing of aquatic oligochaetes: comparison of experimental communities. *PLOS ONE* **11**:e0148644 DOI 10.1371/journal.pone.0148644.

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vzquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**:27 DOI 10.1186/s40168-017-0237-y.

Willis AD. 2019. Rarefaction, alpha diversity, and statistics. *Frontiers in Microbiology* **10**: DOI 10.3389/fmicb.2019.02407.

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M. 2012. Human gut microbiome viewed across age and geography. *Nature* **486**:222–227 DOI 10.1038/nature11053.