1   **Title:** To rarefy or not to rarefy: Enhancing diversity analysis of microbial communities through

2   next-generation sequencing and rarefying repeatedly

3   **Authors:** Ellen S. Cameron[a], Philip J. Schmidt[b], Benjamin J.-M. Tremblay[a], Monica B.

4   Emelko[b], Kirsten M. Müller[a,*]

5   [a] Department of Biology, University of Waterloo, 200 University Ave. W, Waterloo, Ontario,

6   Canada, N2L 3G1

7   [b] Department of Civil and Environmental Engineering, University of Waterloo, 200 University

8   Ave. W, Waterloo, Ontario, Canada, N2L 3G1

9   [*] Corresponding author. Tel.: +1 519 888 4567x32224

10   E-mail address: kirsten.muller@uwaterloo.ca (K.M. Müller).

11

Abstract

Amplicon sequencing has revolutionized our ability to study DNA collected from environmental samples by providing a rapid and sensitive technique for microbial community analysis that eliminates the challenges associated with lab cultivation and taxonomic identification through microscopy. In water resources management, it can be especially useful to evaluate ecosystem shifts in response to natural and anthropogenic landscape disturbances to signal potential water quality concerns, such as the detection of toxic cyanobacteria or pathogenic bacteria. Amplicon sequencing data consist of discrete counts of sequence reads, the sum of which is the library size. Groups of samples typically have different library sizes that are not representative of biological variation; library size normalization is required to meaningfully compare diversity between them. Rarefaction is a widely used normalization technique that involves the random subsampling of sequences from the initial sample library to a selected normalized library size. Rarefying is often dismissed as statistically invalid because subsampling effectively discards a portion of the observed sequences. Nonetheless, it remains prevalent in practice. Notably, the superiority of rarefying relative to many other normalization approaches has been argued in diversity analysis. Here, repeated rarefying is proposed as a tool for diversity analyses to normalize library sizes. This enables (i) proportionate representation of all observed sequences and (ii) characterization of the random variation introduced to diversity analyses by rarefying to a smaller library size shared by all samples. While many deterministic data transformations are not tailored to produce equal library sizes, repeatedly rarefying reflects the probabilistic process by which amplicon sequencing data are obtained as a representation of the source microbial community. Specifically, it evaluates which data might have been obtained if a particular sample's library

34   size had been smaller and allows graphical representation of the effects of this library size

35   normalization process upon diversity analysis results.

36   **Keywords (Maximum 6 keywords)**

37   Library size normalization, amplicon sequencing, alpha diversity, beta diversity, Shannon index,
38   Bray-Curtis dissimilarity

39   **1.   Introduction**

40   Next-generation sequencing (NGS) has revolutionized the understanding of environmental

41   systems through the characterization of microbial communities and their function by examining

42   DNA collected from samples that contain mixed assemblages of organisms (Bartram et al., 2011;

43   Hugerth and Andersson, 2017; Shokralla et al., 2012). It is well known that fewer than 1% of

44   species in the environment can be isolated and cultured, limiting the ability to identify rare and

45   difficult-to-cultivate members of the community (Bodor et al., 2020; Cho and Giovannoni, 2004;

46   Ferguson et al., 1984). In addition to the limitations of culturing, microscopic evaluation of

47   environmental samples remains of limited utility because of challenges in high-resolution

48   taxonomic identification and the inability to infer function from morphology (Hugerth and

49   Andersson, 2017). Metagenomic evaluations employ NGS technology to analyze large quantities

50   of diverse environmental DNA (Thomas et al., 2012) and have largely eliminated challenges

51   associated with culturing and microscopic identification (McMurdie and Holmes, 2014).

52   Metagenomics encompasses a conglomerate of different sequencing experimental designs,

53   including amplicon sequencing (sequencing of amplified genes of interest) and shotgun

54   sequencing (sequencing of fragments of present genetic material). While shotgun sequencing

55   allows characterization of the entire community, including both taxonomic composition and

56   functional gene profiles, it is not widely accessible due to high sequencing costs and

3

57 computational requirements for analysis (Bartram et al., 2011; Clooney et al., 2016; Langille et

58 al., 2013). In contrast, the relatively low cost of amplicon sequencing has made it an increasingly

59 popular technique (Clooney et al., 2016; Langille et al., 2013). The amplification and sequencing

60 of specific genes (e.g., taxonomic marker genes) enables characterization of microbial

61 community composition (Hodkinson and Grice, 2015); as a result, it has been successfully

62 applied in many areas of environmental and water research. For example, amplicon sequencing

63 has been used to characterize and predict cyanobacteria blooms (Tromas et al., 2017), describe

64 microbial communities found in aquatic ecosystems (Zhang et al., 2020), and evaluate

65 groundwater vulnerability to pathogen intrusion (Chik et al., 2020). It has also been applied to

66 water quality and treatment performance monitoring in diverse settings (Vierheilig et al., 2015),

67 including drinking water distribution systems (Perrin et al., 2019; Shaw et al., 2015), drinking

68 water biofilters (Kirisits et al., 2019), anaerobic digesters (Lam et al., 2020), and cooling towers

69 (Paranjape et al., 2020).

70   Processing and analysis of amplicon sequencing data are statistically complicated for a

71 number of reasons (Weiss et al., 2017). In particular, library sizes (i.e., the total number of

72 sequencing reads within a sample) can vary widely among different samples, even within a

73 single sequencing run, and the disparity in library sizes between samples may not represent

74 actual differences in microbial communities (McMurdie and Holmes, 2014). Amplicon

75 sequencing libraries cannot be compared directly for this reason. For example, two replicate

76 samples with 5,000 and 20,000 sequence reads, respectively, are likely to have different read

77 counts for specific sequence variants simply due to the difference in library size. While

78 parametric tools such as generalized linear modelling (e.g., McMurdie and Holmes, 2014) can

79 provide a statistically sound framework for differential abundance analysis, drawing biologically

4

80   meaningful diversity analysis conclusions from amplicon sequencing data typically requires

81   normalization of library sizes to account for the additional variation in counts that is attributable

82   to differences in library sizes between samples (McKnight et al., 2019). For example, larger

83   samples may appear more diverse than smaller samples (Hughes and Hellmann, 2005). Notably,

84   a variety of normalization techniques that may affect the analysis and interpretation of results

85   have been suggested, including rarefaction (i.e., the process of rarefying libraries to a common

86   size).

87   Rarefaction is a normalization tool initially developed for ecological diversity analyses to

88   allow for sample comparison without associated bias from differences in sample size (Sanders,

89   1968). Rarefaction normalizes samples of differing sample size by subsampling each to a shared

90   threshold. Although initially developed for use in ecological studies, rarefaction is a commonly

91   used library size normalization technique for amplicon sequencing data. As a result, it is the

92   subject of considerable debate and statistical criticism (Gloor et al., 2017; McMurdie and

93   Holmes, 2014). Rarefying is typically conducted in a single iteration that only provides a

94   snapshot of the community that might have been observed at the smaller normalized library size.

95   This omits a random subset of observed sequences and potentially also samples with small

96   library sizes and introduces artificial variation to the data (McMurdie and Holmes, 2014).

97   Repeatedly rarefying, on the other hand, has the potential to address the statistical concerns

98   associated with omission of data and could provide a more statistically acceptable technique than

99   performing a single iteration of rarefying for diversity analyses. It characterizes what data might

100   have been obtained if a particular sample's library size had been smaller, revealing what can be

101   inferred about community diversity in the source from samples of equal library size. Rarefying

102   repeatedly has received only trivial consideration in the literature (e.g., McMurdie and Holmes,

103   2014; Navas-Molina et al., 2013).Diversity analysis approaches grounded in statistical inference

104   about source microbial diversity (that address the random probabilistic processes through which

105   NGS yields libraries of sequence reads) could conceptually be superior to rarefying (Willis,

106   2019), but they are not yet fully developed or readily available for routine diversity analysis to

107   support study of environmental microbial communities.

108   Here, we investigate the application of repeatedly rarefying as a library size normalization

109   technique specifically for diversity analyses. This paper graphically evaluates the impact of

110   subsampling with or without replacement and normalized library size selection on diversity

111   analyses such as the Shannon index and Bray-Curtis dissimilarity ordinations, specifically.

112   Rather than representing diversity as a single numerical value or point in an ordination plot

113   (often following transformation that may not be designed to compensate for differing library

114   sizes), rarefying repeatedly yields bands of values or patches of points that characterize how

115   diversity may vary among or between samples at a particular library size.

116   **2. Theory**

117   *2.1 Amplicon Sequencing and Diversity Analysis for Microbial Communities in Water – An*

118   *Overview*

119   Due to the inevitable interdisciplinarity of environmental water quality research and the

120   complexity and novelty of next generation sequencing relative to traditional microbiological

121   methods used in water quality analyses, further detail on amplicon sequencing is provided.

122   Amplification and sequencing of taxonomic marker genes has been used extensively to examine

123   phylogeny, evolution, and taxonomic classification of numerous groups across the three domains

124   of life (Quast et al., 2013; Weisburg et al., 1991; Woese et al., 1990). Taxonomic marker genes

125   include the 16S rRNA gene in mitochondria, chloroplasts, bacteria and archaea (Case et al.,

6

126   2007; Tsukuda et al., 2017; Weisburg et al., 1991; Yang et al., 2016), or the 18S rRNA gene

127   within the nucleus of eukaryotes (Field et al., 1988). Widely used reference databases have been

128   developed containing marker gene sequences across numerous phyla (Hugerth and Andersson,

129   2017).

130   The 16S rRNA gene consists of nine highly conserved regions separated by nine

131   hypervariable regions (V1-V9; Gray et al., 1984) and is approximately 1,540 base pairs in length

132   (Kim et al., 2011; Schloss and Handelsman, 2004). While sequencing of the full 16S rRNA gene

133   provides the highest taxonomic resolution (Johnson et al., 2019), many studies only utilize partial

134   sequences due to limitations in read length of NGS platforms (Kim et al., 2011). Next-generation

135   sequencing on Illumina platforms (Illumina Inc., San Diego, California) produces reads that are

136   up to 350 base pairs in length, requiring selection of an appropriate region of the 16S rRNA gene

137   to amplify and sequence for optimal taxonomic resolution (Bukin et al., 2019; Kim et al., 2011).

138   Sequencing the more conservative regions of the 16S rRNA gene may be limited to resolution of

139   higher levels of taxonomy, while more variable regions can provide higher resolution for the

140   classification of sequences to the genus and species levels in bacteria and archaea (Bukin et al.,

141   2019; Kim et al., 2011; Yang et al., 2016).

142   Different variable regions of the 16S rRNA gene may be biased towards different taxa

143   (Johnson et al., 2019) and be preferred for different ecosystems (Escapa et al., 2020). For

144   example, the V4 region has been shown to strongly differentiate taxa from the phyla

145   Cyanobacteria, Firmicutes, Fusobacteria, Plantomycetes, and Tenericutes but the V3 region best

146   differentiates taxa from the phyla Proteobacteria (e.g., *Escherichia coli*, *Salmonella* spp.,

147   *Campylobacter* spp.), Acidobacteria, Bacteroidetes, Chloroflexi, Gemmatimonadetes,

148   Nitrospirae, and Spirochaetae (Zhang et al., 2018). The V4 region of the 16S rRNA gene is

7

149   frequently targeted using specific primers designed to minimize amplification bias while

150   accounting for common aquatic bacteria (Walters et al., 2015) and is frequently used in aquatic

151   studies (Zhang et al., 2018). It is important to consider suitability of a 16S rRNA region for the

152   habitat (Escapa et al., 2020) and the taxa present in the microbial community due to potential

153   bias of analyzing differing subregions of the 16S rRNA gene (Johnson et al., 2019; Zhang et al.,

154   2018).

155   The use of amplicon sequencing of partial sequences of the 16S rRNA gene allows

156   examination of microbial community composition and the exploration of shifts in community

157   structure in response to environmental conditions (Hodkinson and Grice, 2015), and

158   identification of differentially abundant taxa between samples (Hugerth and Andersson, 2017).

159   Amplicon sequencing datasets can be analyzed using a variety of bioinformatics pipelines for

160   sequence analysis (e.g., sequence denoising, taxonomic classification, diversity analysis)

161   including *mothur* (Schloss et al., 2009) and *QIIME2* (Bolyen et al., 2019). Previously,

162   sequencing analysis involved the creation of dataset-dependent operational taxonomic units

163   (OTUs) by clustering sequences into groups that met a certain similarity threshold, resulting in a

164   loss of representation of variation in sequences and precluding cross-study comparison (Callahan

165   et al., 2017). Advances in computational power have allowed a shift from use of OTUs to

166   amplicon sequence variants (ASVs) representative of each unique sequence in a sample, which

167   allows for the comparison of sequence variants generated in different studies and retains the full

168   observed biological variation (Callahan et al., 2017). The implementation of tools included

169   bioinformatics pipelines, such as *DADA2* (Callahan et al., 2016) or *Deblur* (Amir et al., 2017),

170   allows quality control of sequencing through the removal of sequencing errors and for the

171   creation of ASVs.

8

172   Quality controlled sequencing data for a particular run is then organized into large matrices

173   where columns represent experimental samples and rows contain counts for different ASVs

174   (Weiss et al., 2017). Amplicon sequencing samples have a total number of sequencing reads

175   known as the library size (McMurdie and Holmes, 2014), but do not provide information on the

176   absolute abundance of sequence variants (Gloor et al., 2016, 2017). This data can be used for

177   studies on taxonomic composition, differential abundance analysis and diversity analyses (Figure

178   1). Taxonomic classification of 16S rRNA sequences using rRNA databases including SILVA

179   (Quast et al., 2013), the Ribosomal Database Project (Cole et al., 2014) and GreenGenes

180   (DeSantis et al., 2006) allows for construction of taxonomic community profiles (Bartram et al.,

181   2011). Taxonomic composition analysis allows for characterization of microbial communities by

182   classifying sequence variants based on similarities to sequences in online databases. The creation

183   of taxonomic composition graphs frequently expresses community composition in proportions.

184   Differential abundance analysis is utilized to explore whether specific sequence variants are

185   found in significantly different proportions between samples (Weiss et al., 2017) to identify

186   potential biological drivers for these differences. This application is outside the scope of this

187   work and is frequently performed using programs initially designed for transcriptomics, such as

188   *DESeq2* (Love et al., 2014) and *edgeR* (Robinson et al., 2009), or programs designed to account

189   for the compositional structure of sequence data *ALDeX2* (Fernandes et al., 2014). The final

190   potential application of this data is diversity analyses, which can be evaluated on varying scales

191   from within sample (alpha) to between samples (beta; Sepkoski, 1988) but is associated with the

192   challenge of the true diversity of environmental sources largely remaining unknown (Hughes et

193   al., 2001).

194 Alpha diversity serves to identify richness (e.g., number of observed sequence variants) and

195 evenness (e.g., allocation of read counts across observed sequence variants) within a sample

196 (Willis, 2019). Comparison of alpha diversity among samples of differing library sizes may

197 result in inherent biases, with samples having larger library sizes appearing more diverse due to

198 the potential presence of more sequence variants in samples with larger libraries (Hughes and

199 Hellmann, 2005; Willis, 2019). This has commonly required samples to have equal library sizes

200 before comparison to prevent bias fabricated only from differences in initial library size.

201 Diversity indices used to characterize the alpha diversity of samples include but are not limited

202 to the Shannon index (Shannon, 1948), Chao1 index (Chao and Bunge, 2002), and Simpson

203 index (Simpson, 1949), but unique details of such indices should be understood for correct

204 usage. For example, Chao1 relies on the observation of singletons in data to estimate diversity

205 (Chao and Bunge, 2002), but denoising processes for sequencing data may remove singleton

206 reads making the Chao1 estimator invalid for accurate analysis. The Shannon index, used in this

207 study, is affected by differing library sizes because the contribution of rare sequences to total

208 diversity is progressively lost with smaller library sizes.

209 Similar to alpha diversity, samples with differing library sizes in beta diversity analyses may

210 produce erroneous results due to the potential for samples with larger library sizes to have more

211 unique sequences simply due to the presence of more sequence variants (Weiss et al., 2017). A

212 variety of beta diversity metrics can be used to compare sequence variant composition between

213 samples including Bray-Curtis (Bray and Curtis, 1957) or Unifrac (Lozupone and Knight, 2007)

214 distances, which can then be visualized using ordination techniques (e.g., PCA, PCoA, NMDS).

215 Bray-Curtis dissimilarity, used in this study, includes pairwise comparison of the numbers for

216   each ASV between two samples, which are expected to be quite dissimilar (even if the

217   communities they represent are not) if library sizes vary substantially.

218   *2.2 Limitations of Library Size Normalization Techniques*

219   Diversity analysis, as it is presently applied, usually requires library size normalization to

220   account for bias introduced through varying read counts in samples. For example, samples with

221   larger library sizes may appear more diverse simply due to the presence of more sequences.

222   Normalization techniques that feature various statistical transformations have been proposed for

223   use in place of rarefying or proportions (McKnight et al., 2019), including upper-quartile log fold

224   change (e.g., Robinson et al., 2009), centered log-ratio transformations (e.g., Gloor et al., 2017),

225   geometric mean pairwise ratios (e.g., Chen et al., 2018), variance stabilizing transformations

226   (e.g., Love et al., 2014) or relative log expressions (e.g., Badri et al., 2018). McKnight et al.

227   (2019) noted that the failure of most normalization techniques to transform data to equal library

228   sizes for diversity analysis "is discouraging, as standardizing read depths are the initial impetus

229   for normalizing the data (i.e., if all samples had equal read depths after sequencing, there would

230   be no need to normalize".

231   These proposed alternatives to rarefying are also often compromised by the presence of large

232   proportions of zero count data in tabulated amplicon sequencing read counts. Zero counts

233   represent a lack of information (Silverman et al., 2018) and may arise from true absence of the

234   sequence variant in the sample or a loss resulting in it not being detected when it was actually

235   present (Tsilimigras and Fodor, 2016; Wang and LêCao, 2019). Nonetheless, many

236   normalization procedures for amplicon sequencing datasets require zero counts to be omitted or

237   modified, especially when applying transformations that utilize logarithms (e.g., centered log-

238   ratio, relative log expressions, geometric mean pairwise ratios). Methods that utilize logarithms

11

239   involve fabricating count values (pseudocounts) for the many zeros of which amplicon

240   sequencing datasets are comprised and selecting a pseudocount value is an additional challenge

241   (Weiss et al., 2017) that may be accomplished using probabilistic arguments (Gloor et al., 2016;

242   2017). Zeros are a natural occurrence in discrete, count-based data such as the counting of

243   microorganisms or amplicon sequences and adjusting or omitting them can introduce substantial

244   bias into microbial analyses (Chik et al., 2018).

245   McMurdie and Holmes (2014) noted that use of proportions is problematic due to

246   heteroscedasticity: for example, one sequence read in a library size of 100 is a far less precise

247   representation of source composition than 100 sequence reads in a library size of 10,000, even

248   though both comprise 1% of the observed sequences. McKnight et al. (2019) favour use of

249   proportions in diversity analysis without noting how precision of proportions, and the degree to

250   which alpha diversity in the source is reflected (Willis, 2019), varies with library size. Willis

251   (2019) also points towards a conceptually better approach to diversity analysis that accounts for

252   measurement error and the difference between the sample data and the population

253   (environmental source) of which the sample data are only a partial representation. Diversity

254   analysis in general does not do this, as it applies a set of calculations to sample data (or some

255   transformation thereof) to obtain one value of alpha diversity or one point on an ordination plot.

256   Pending further development of such approaches, this study revisits rarefying because of the

257   practical simplicity of comparing diversity among samples of equal library size.

258   McMurdie and Holmes (2014) propose that rarefying is not a statistically valid normalization

259   technique due to the omission of valid data, which may be resolved for the purposes of diversity

260   analysis by rarefying repeatedly to represent all sequences in the proportions with which they

261   were observed and compare sample-level microbial community diversity at a particular library

262   size. In addition, McMurdie and Holmes (2014) dismissed repeatedly rarefying as a

263   normalization technique, in part because repeatedly rarefying an artificial library consisting of a

264   50:50 ratio of two sequence variants does not yield a 50:50 ratio at the rarefied library size and

265   this added noise could affect downstream analyses. However, such error is inherent to

266   subsampling, whether from a population or from a larger sequence library and has thus already

267   affected samples with smaller library sizes; it is the reason why simple proportions are less

268   precise in samples with smaller library sizes.

269   McMurdie and Holmes (2014), also cited the investigation of Navas-Molina et al. (2013) as

270   an example of repeatedly rarefying to normalize library sizes and used it to support their

271   dismissal of this technique due to the omission of valid data and added variability. However, it is

272   critical to note that the work in Navas-Molina et al. (2013) reported using jackknife resampling

273   of sequences, which cannot be equated to repeatedly rarefying (random resampling with or

274   without replacement). Hence, it is necessary to build upon preliminary analysis of repeatedly

275   rarefying as a normalization technique and to explore the impact of subsampling approach and

276   normalized library size on diversity analysis results.

277   **3. Methods**

278   *3.1 Example Data – DNA Extraction and Amplicon Sequencing*

279   Samples used in this study are part of a larger study at Turkey Lakes Watershed (North Part,

280   ON), but only an illustrative subset of samples is considered for the purpose of evaluating

281   rarefaction rather than for ecological interpretation. This allows evaluation of repeated rarefying

282   as a normalization technique without utilizing simulated data. DNA extracts isolated from

283   environmental samples were submitted for amplicon sequencing using the Illumina MiSeq

284   platform (Illumina Inc., San Diego, California) at the commercial laboratory Metagenom Bio

285   Inc. (Waterloo, Ontario). Primers designed to target the 16S rRNA gene V4 region [515FB

286   (GTGYCAGCMGCCGCGGTAA) and 806RB (GGACTACNVGGGTWTCTAAT; Walters et

287   al., 2015)] were used for PCR amplification.

288   *3.2 Sequence Processing and Library Size Normalization*

289   The program *QIIME2* (v. 2019.10; Bolyen et al., 2019) was used for bioinformatic processing of

290   sequence reads. Demultiplexed paired-end sequences were trimmed and denoised, including the

291   removal of chimeric sequences and singleton sequence variants to avoid sequences that may not

292   be representative of real organisms, using *DADA2* (Callahan et al., 2016) to construct the ASV

293   table. Zeroing all singleton sequences could erroneously remove legitimate sequences,

294   particularly if the sequence in question is detected in large numbers in other similar samples;

295   however, the potential effect of such error upon diversity analysis is beyond the scope of this

296   work. Output files from *QIIME2* were imported into R (v. 4.0.1; R Core Team, 2020) for

297   community analyses using *qiime2R* (v. 0.99.23; Bisanz, 2018). Initial sequence libraries were

298   further filtered using *phyloseq* (v. 1.32.0; McMurdie and Holmes, 2013) to exclude amplicon

299   sequence variants that were taxonomically classified as mitochondria or chloroplast sequences.

300   We developed a package called *mirlyn* (Multiple Iterations of Rarefaction for Library

301   Normalization; Cameron and Tremblay, 2020) that facilitates implementation of techniques used

302   in this study built from existing R packages (Table S1). Using the output from *phyloseq*, *mirlyn*

303   was used to (1) generate rarefaction curves, (2) repeatedly rarefy libraries to account for

304   variation in library sizes among samples, and (3) plot diversity metrics given repeated

305   rarefaction.

306 *3.3 Community Diversity Analyses on Normalized Libraries*

307 The impact of normalized library size on the Shannon index (Shannon, 1948), an alpha diversity

308 metric, was evaluated. Normalized libraries were also used for beta diversity analysis. A

309 Hellinger transformation was applied to normalized libraries to account for the arch effect

310 regularly observed in ecological count data and Hellinger-transformed data were then used to

311 calculate Bray-Curtis distances (Bray and Curtis, 1957). Principal component analysis (PCA)

312 was conducted on the Bray-Curtis distance matrices.

313 *3.4 Study Approach*

314 Typically, rarefaction has only been conducted a single time in microbial community analyses,

315 and this omits a random subset of observed sequences, introducing a possible source of error. To

316 examine this error, samples were repeatedly rarefied 1000 times. This repetition provides a

317 representative suite of rarefied samples capturing the randomness in sequence variant

318 composition imposed by rarefying. The sections below address the various decisions that must be

319 made by the analyst and factors affecting reliability of results when rarefaction is used.

320 *3.4.1 The Effects of Subsampling Approach – With or Without Replacement*

321 Rarefying library sizes may be performed with or without replacement. To evaluate the effects of

322 subsampling replacement approaches, we repeatedly rarefied filtered sequence libraries with and

323 without replacement. Results of the two approaches were contrasted in diversity analyses to

324 evaluate the impact of subsampling approach on interpretation of results.

325 *3.4.2 The Effects of Normalized Library Size Selection*

326 Rarefying involves the selection of an appropriate sampling depth to be shared by each sample.

327 To evaluate the effects of different rarefied library sizes, filtered sequence libraries were rarefied

15

328   repeatedly to varying depths. Results for various sampling depths were contrasted in diversity

329   analyses to evaluate the impact of normalized library size selection on interpretation of results.

330   **4. Results and Discussion**

331   *4.1 Use of Rarefaction Curves to Explore Suitable Normalized Library Sizes*

332   Rarefying requires the selection of a potentially arbitrary normalized library size, which

333   can impact subsequent community diversity analyses and therefore presents users with the

334   challenge of making an appropriate decision of what size to select (McMurdie and Holmes,

335   2014). Suitable sampling depths for groups of samples can be determined through the

336   examination of rarefaction curves (Figure 2). By selecting a library size that encompasses the

337   flattening portion of the curve for each sample, it is generally assumed that the normalized

338   library size will adequately capture the diversity within the samples despite the exclusion of

339   sequence reads during the rarefying process (i.e., there are progressively diminishing returns in

340   including more of the observed sequence variants as the rarefaction curve flattens).

341   Suggestions have previously been made encouraging selection of a normalized library

342   size that is encompassing of most samples (e.g., 10,000 sequences) and advocation against

343   rarefying below certain depths (e.g., 1,000 sequences) due to decreases in data quality (Navas-

344   Molina et al., 2013). However, generic criteria may not be applicable to all datasets and

345   exploratory data analysis is often required to make informed and appropriate decisions on the

346   selection of a normalized library size. Although previous research advises against rarefying

347   below certain thresholds, users may be presented with the dilemma of selecting a sampling depth

348   that either does not capture the full diversity of a sample depicted in the rarefaction curve (Figure

349   2 – I) or would require the omission of entire samples with smaller library sizes (Figure 2 – III).

350   The implementation of multiple iterations of rarefying library sizes will aid in alleviating this

351   dilemma by capturing the potential losses in community diversity for samples that are rarefied to

352   lower than ideal depth. Doing so with two or more normalized library sizes may reveal

353   differences in diversity attributable to relatively rare variants that could be suppressed by

354   normalizing to too small of a library size.

355   *4.2 The Effects of Subsampling Approach and Normalized Library Size Selection on Alpha*

356   *Diversity Analyses*

357   The differences in how rarefying samples may be carried out requires users to be diligent

358   in the selection of appropriate tools and commands for their analysis. The R package *phyloseq*, a

359   popular tool for microbiome analyses, has default settings for rarefying including sampling with

360   replacement to optimize computational run time and memory usage (McMurdie and Holmes,

361   2013). Sampling without replacement, however, is more appropriate statistically because it draws

362   a subset from the observed set of sequences (as though the sample had yielded only the specified

363   library size), whereas sampling with replacement fabricates a set of sequences in similar

364   proportions to the observed set of sequences (Figure 3). Sampling with replacement can

365   potentially cause a rare sequence variant to appear more frequently in the rarefied sample than it

366   was in the original library.

367   Rarefying libraries with or without replacement was not found to substantially impact the

368   Shannon index in the scenarios considered in this study (Figure 4-A), but users should still be

369   aware of potential implications of sampling with or without replacement when rarefying

370   libraries. Libraries rarefied with replacement are observed to have a slightly reduced Shannon

371   index relative to libraries rarefied without replacement at many library sizes because rare

372   sequences are excluded more often when sampling with replacement.

373   The conservation of larger normalized library sizes allows detection of more diversity

374   with minimal variation observed between the iterations of rarefaction (Figure 4-A). The largest

375   considered normalized library size (the sample with the smallest library size has 11,213

376   sequences) captured the highest Shannon index values, while the Shannon index diminishes for

377   all samples at lower normalized library sizes. The use of repeated iterations of rarefying allows

378   variation introduced through subsampling to be represented in the diversity metric, which is

379   small at larger library sizes. While there was only slight disparity in the Shannon index values

380   between the largest library size and unnormalized data, this may not always be the case and is

381   dependent on the sequence variant composition of the samples. Samples dominated by a large

382   number of low-abundance sequence variants are more likely to have a substantially reduced

383   Shannon index value at a larger normalized library size. Alternatively, samples dominated by

384   only a few highly abundant sequence variants will be comparatively robust to rarefying. A plot

385   of the Shannon index as a function of rarefied library size (Figure 4-B) demonstrates the overall

386   robustness of the Shannon index of these samples for larger library sizes (e.g., > 5,000

387   sequences) and the increased variation and diminishing values when proceeding to smaller

388   rarefied library sizes. When the normalized library size was decreased to 5,000, the Shannon

389   index is still only slightly reduced by the rarefaction but there is greater variability introduced

390   from rarefying.

391   The consistency of the diversity metric when rarefying repeatedly is extremely degraded

392   when libraries were rarefied to the smallest considered library size of 500 sequences. It illustrates

393   the potential to reach incorrect conclusions if rarefying is completed only once. When rarefying

394   repeatedly to a small library size, however, diversity index values that are both highly

395   inconsistent and suppressed relative to the diversity of the unrarefied data may lead to

18

396   inappropriate claims of identical diversity values between samples (e.g., samples A, B, and C

397   become indistinguishable). The extreme reduction and introduced variation of the Shannon index

398   suggests that the selection of smaller rarefied library sizes should be approached with caution

399   when using alpha diversity metrics, while larger normalized library sizes prevent loss of

400   precision and reduction of the Shannon index value. However, as previously noted, the reduction

401   in the value of the Shannon index will be dependent on the sequence variant composition of the

402   samples.

403   Previous research evaluating normalization techniques has focused on beta diversity

404   analysis and differential abundance analysis (Gloor et al., 2017; McMurdie and Holmes, 2014;

405   Weiss et al., 2017), but the appropriateness of library size normalization techniques for alpha

406   diversity metrics must be evaluated due to the prerequisite of having equal library sizes for

407   accurate calculation. Utilization of unnormalized library sizes with alpha diversity metrics may

408   generate bias due to the potential for samples with larger library sizes to inherently reflect more

409   of the diversity in the source than a sample with a small library size. The repeated iterations of

410   rarefying library sizes allow characterization of the variability introduced to sample diversity by

411   rarefying at any rarefied library size (Figure 4) but does not allow evaluation of uncertainty

412   about the diversity in the source from which the sample was taken, as is the case for all

413   normalization-based approaches.

414   *4.3 The Effects of Subsampling Approach and Normalized Library Size Selection on Beta*

415   *Diversity Analysis*

416   When samples were repeatedly rarefied to a common normalized library size with and

417   without replacement, similar amounts of variation in the Bray-Curtis PCA ordinations were

418   observed between the sampling approaches (Figure 5). This indicates that although rarefying

19

419   with replacement seems potentially erroneous due to the fabrication of count values that are not

420   representative of actual data, the impact on the variation introduced into the Bray-Curtis

421   dissimilarity distances is not large and will likely not interfere with the interpretation of results.

422   However, rarefying without replacement should be encouraged because it is more theoretically

423   correct to represent possible data if only the smaller library size had been obtained, and it has not

424   been comprehensively demonstrated that sampling with replacement is a valid approximation for

425   all types of diversity analysis or library compositions.

426   When larger normalized library sizes are maintained through rarefaction, there is less

427   potential variation introduced into beta diversity analyses, including Bray-Curtis dissimilarity

428   PCA ordinations. For example, in the largest normalized library size possible for these data

429   (Figure 5A), a minimal amount of variation was observed within each community, indicating

430   that the preservation of higher sequence counts minimizes the amount of artificial variation

431   introduced into datasets by rarefaction (including no variation for Sample F because it is not

432   actually rarefied in this scenario). For this reason, rarefying to the smallest library size of a set of

433   samples is a sensible guideline. Although, a normalized library size of 5,000 is lower than the

434   flattening portion of the rarefaction curve for samples A, B, and C (Figure 2), the selection of

435   this potentially inappropriate normalized library size (Figure 5C) can still accurately reflect the

436   diversity between samples without excess artificial variation introduced through rarefaction. Due

437   to the variation introduced to the Bray-Curtis dissimilarity ordinations in the smaller rarefied

438   library sizes (Figure 5E/G), it is critical to include computational replicates of rarefied libraries

439   to fully characterize the introduced variation in communities. As discussed above, it has been

440   suggested that repeatedly rarefying is inappropriate due to the introduction of "added noise".

441   However, as demonstrated, the maintenance of larger rarefied library sizes when repeatedly

20

442   rarefying does not impact interpretation of beta-diversity analysis results. Without this

443   replication, rarefaction to small, normalized library sizes could result in artificial similarity or

444   dissimilarity identified between samples.

445   Beta diversity analysis of very small, rarefied library sizes (Figure 6A, B, C) can still

446   reflect similar clustering patterns observed in larger library sizes but with a much lower

447   resolution of clusters. Rarefying has previously been shown to be an appropriate normalization

448   tool for samples with low sequence counts (e.g., <1,000 sequences per sample) by Weiss et al.

449   (2017), which is promising for datasets containing samples with small initial library sizes or

450   potentially analyzing subsets of data to explore diversity within specific phyla (e.g.,

451   Cyanobacteria). Caution must be taken to avoid selection of an excessively small, normalized

452   library size due to the introduction of extreme levels of artificial variation that compromises

453   accurate depiction of diversity (Figure 6D) and suppresses the contribution of rare variants to

454   overall diversity. The tradeoff between rarefying to a smaller than advisable library size or

455   excluding entire samples with small library sizes remains and can possibly be resolved by

456   analyzing results with all samples and a small, rarefied library size as well as with some omitted

457   samples and a larger rarefied library size.

458   Although rarefying has the potential to introduce artificial variation into data used in beta

459   diversity analyses, these results suggest that rarefying repeatedly does not become problematic

460   until normalized library sizes are very small (e.g., 500 sequences or less) for the samples

461   considered. While we saw a degradation of the consistency and value of the alpha diversity

462   Shannon index at 500 sequences, beta diversity analyses may be more robust to rarefaction and

463   capable of reflecting qualitative clusters in ordination as previously discussed in Weiss et al.

464   (2017). The artificial variation introduced to beta diversity analyses by rarefaction could lead to

21

465   erroneous interpretation of results, but the implementation of multiple iterations of rarefying

466   library sizes allows a full representation of this variation to aid in determining if apparent

467   similarity or dissimilarity is a chance result of rarefying.

468   The use of non-normalized data has been shown to be more susceptible to the generation

469   of artificial clusters in ordinations, and rarefying has been demonstrated to be an effective

470   normalization technique for beta diversity analyses (Weiss et al., 2017). However, the use of a

471   single iteration of rarefying does result in the omission of valid data (McMurdie and Holmes,

472   2014). Repeated iterations of rarefying in this study demonstrated that rarefying repeatedly does

473   not substantially impact the output and interpretation of beta diversity analyses unless rarefying

474   to sizes that are inadvisably small to begin with. McMurdie and Holmes (2014) were dismissive

475   of rarefying repeatedly due to the variability it introduces, but such repetition was not evaluated

476   in the context of beta-diversity analysis. In the case of differential abundance analysis, the added

477   variability of rarefying would be statistically inappropriate relative to generalized linear

478   modelling that can account for varying library sizes. Additionally, repeatedly rarefying allows

479   for characterization of variation introduced through subsampling while accounting for

480   discrepancies in library size, supporting the potential utility of the normalization technique for

481   beta diversity analyses. McKnight et al. (2019) preferred use of proportions in diversity analysis

482   over rarefying (arguing that both were superior to other normalization approaches). While

483   proportions normalize the sum of the ASV weights to one for each sample, we note that the

484   approach does not normalize the library size in terms of sequence counts. This is important

485   because sample proportions will provide a more precise reflection of the true proportions of

486   which the set of sequences is believed to be representative in samples with larger libraries than in

487   samples with smaller libraries. In particular, using proportions of unnormalized sequence count

488 libraries in beta diversity analysis overlooks the loss of alpha diversity associated with smaller

489 library sizes when comparing samples with different library sizes.

490 *4.4 Perspectives on Library Size Normalization*

491 The increasing popularity and accessibility of amplicon sequencing has enabled the

492 scientific community to gain access to a wealth of microbial community data that would

493 otherwise not have been accessible. However, despite amplicon sequencing of taxonomic marker

494 genes being the gold standard approach for microbial community analysis, the data handling and

495 statistical analysis is still in the early stages of development. The diversity analyses that the

496 scientific community desires to perform on amplicon sequencing data require library sizes to be

497 normalized across samples, which creates the challenge of determining appropriate

498 normalization techniques. New normalization techniques and tools are constantly being

499 developed and released to the community with claims that the newest technique is the best and

500 only solution that should be utilized for analysis, but they may be associated with data handling

501 limitations, be too specifically tailored to a particular type of analysis or desired property, or not

502 normalize the library sizes that motivated the need for normalization (McKnight et al., 2018).

503 For example, the centered-log ratio transformation (Gloor et al., 2016) cannot be used with zero

504 count data and amplicon sequencing datasets must be augmented with an artificial pseudocount

505 to apply the normalization technique. The limitations of normalization techniques may affect

506 downstream analyses, making it critical to understand the implications of the technique chosen.

507 Further discussion within the scientific community is needed to ensure rigorous interpretation

508 of amplicon sequencing data without unwarranted bias introduced by the normalization

509 technique. Approaches to microbiome data analysis that recognize data as samples from a source

510 population and seek to draw inference about diversity in the source rather than just calculating

23

511   diversity in the (transformed) sample are desirable. Random errors are inherent to sample

512   collection, handling, processing, amplification, and sequencing and should be reflected in how

513   resulting data are analyzed. Pending further research on such approaches, rarefying remains

514   common in current research requiring library size normalization despite potential limitations,

515   especially for diversity analysis. The implementation of a single iteration of rarefying is

516   problematic due to the omission of valid data and should not be used for library size

517   normalization. Conducting repeated iterations of rarefying, however, does not discard valid

518   sequences and allows for the characterization of variation introduced through random

519   subsampling in diversity analyses.

520   **Conclusions**

521   ▪ Repeated rarefying (e.g., 1000 times if computationally feasible) statistically describes

522   possible realizations of the data if the number of sequences read had been limited to the

523   normalized library size, thus allowing diversity analysis using samples of equal library

524   size in a way that accounts for the data loss in rarefying.

525   ▪ Rarefying with or without replacement did not substantially impact the interpretation of

526   alpha (Shannon index) or beta (Bray-Curtis dissimilarity) diversity analyses considered in

527   this study, but rarefying without replacement is theoretically more appropriate and will

528   provide more accurate reflection of sample diversity.

529   ▪ The use of larger normalized library sizes when rarefying minimizes the amount of

530   artificial variation introduced into diversity analyses but may necessitate omission of

531   samples with small library sizes (or analysis at both inclusive low library sizes and

532   restrictive higher library sizes).

24

533 • Ordination patterns are relatively well preserved down to small, normalized library sizes

534 with increasing variation shown by repeatedly rarefying, whereas the Shannon index is

535 very susceptible to being impacted by small, normalized library sizes both in declining

536 values and variability introduced through rarefaction.

537 • Even though repeated rarefaction can characterize the error introduced by excluding some

538 fraction of the sequence variants, rarefying to extremely small sizes (e.g., 100 sequences)

539 is inappropriate because the substantial introduced variation leads to an inability to

540 differentiate between sample clusters and suppresses contribution of rare variants to

541 diversity.

542 • Further development of strategies (e.g., data handling, library size normalization for

543 diversity analyses) for ensuring rigorous interpretation of amplicon sequencing data is

544 required.

Acknowledgements

5. References

551 Amir, A., Daniel, M., Navas-Molina, J., Kopylova, E., Morton, J., Xu, Z.Z., Eric, K., Thompson,
552 L., Hyde, E., Gonzalez, A., Knight, R., 2017. Deblur rapidly resolves single-nucleotide
553 community sequence patterns. mSystems 2:e00191, e00191-16.
554 https://doi.org/10.1128/mSystems.00191-16

555 Badri, M., Kurtz, Z., Muller, C., Bonneau, R., 2018. Normalization methods for microbial
556 abundance data strongly affect correlation estimates. bioRxiv 406264.
557 https://doi.org/10.1101/406264

558 Bartram, A.K., Lynch, M.D.J., Stearns, J.C., Moreno-Hagelsieb, G., Neufeld, J.D., 2011.
559 Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial
560 communities by assembling paired-end Illumina reads. Appl. Environ. Microbiol. 77, 3846–
561 3852. https://doi.org/10.1128/AEM.02772-10

562 Bisanz, J.E., 2018. qiime2R: Importing QIIME2 artifacts and associated data into R sessions.
563   https://github.com/jbisanz/qiime2R.

564 Bodor, A., Bounedjoum, N., Vincze, G.E., Erdeiné Kis, Á., Laczi, K., Bende, G., Szilágyi, Á.,
565   Kovács, T., Perei, K., Rákhely, G., 2020. Challenges of unculturable bacteria:
566   environmental perspectives. Rev. Environ. Sci. Biotechnol. 19, 1–22.
567   https://doi.org/10.1007/s11157-020-09522-4

568 Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A.,
569   Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K.,
570   Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M.,
571   Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall,
572   D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M.,
573   Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S.,
574   Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler,
575   B.D., Kang, K. Bin, Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek,
576   T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.X., Loftfield, E., Lozupone, C.,
577   Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A. V., Metcalf,
578   J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F.,
579   Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E.,
580   Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer,
581   A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh,
582   P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-
583   Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J.,
584   Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu,
585   Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible
586   microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857.
587   https://doi.org/10.1038/s41587-019-0209-9

588 Bray, J.R., Curtis, J.T., 1957. An Ordination of the Upland Forest Communities of Southern
589   Wisconsin. Ecol. Monogr. 27, 325–349. https://doi.org/10.2307/1942268

590 Bukin, Y.S., Galachyants, Y.P., Morozov, I. V., Bukin, S. V., Zakharenko, A.S., Zemskaya, T.I.,
591   2019. The effect of 16S rRNA region choice on bacterial community metabarcoding results.
592   Sci. Data 6, 1–14. https://doi.org/10.1038/sdata.2019.7

593 Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace
594   operational taxonomic units in marker-gene data analysis. ISME J. 11, 2639–2643.
595   https://doi.org/10.1038/ismej.2017.119

596 Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016.
597   DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods 13,
598   581–583. https://doi.org/10.1038/nmeth.3869

599 Cameron, E.S., Tremblay, B.J-M., 2020. mirlyn: Multiple iterations of rarefying for library
600   normalization. http://github.com/escamero/mirlyn

601 Case, R.J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W.F., Kjelleberg, S., 2007. Use of
602   16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl.
603   Environ. Microbiol. 73, 278–288. https://doi.org/10.1128/AEM.01177-06

Chao, A., Bunge, J., 2002. Estimating the number of species in a stochastic abundance model. Biometrics 58, 531–539. https://doi.org/10.1111/j.0006-341X.2002.00531.x

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 2018, 1–20. https://doi.org/10.7717/peerj.4600

Chik, A.H.S., Schmidt, P.J., Emelko, M.B., 2018. Learning something from nothing: The critical importance of rethinking microbial non-detects. Front. Microbiol. 9, 1–9. https://doi.org/10.3389/fmicb.2018.02304

Chik, A.H.S., Emelko, M.B., Anderson, W.B., O'Sullivan, K.E., Savio, D., Farnleitner, A.H., Blaschke, A.P., Schijven, J.F., 2020. Evaluation of groundwater bacterial community composition to inform waterborne pathogen vulnerability assessments. Sci. Total Environ. 743, 140472. https://doi.org/10.1016/j.scitotenv.2020.140472

Cho, J.C., Giovannoni, S.J., 2004. Cultivation and Growth Characteristics of a Diverse Group of Oligotrophic Marine Gammaproteobacteria. Appl. Environ. Microbiol. 70, 432–440. https://doi.org/10.1128/AEM.70.1.432-440.2004

Clooney, A.G., Fouhy, F., Sleator, R.D., O'Driscoll, A., Stanton, C., Cotter, P.D., Claesson, M.J., 2016. Comparing apples and oranges?: Next generation sequencing and its impact on microbiome analysis. PLoS One 11, 1–16. https://doi.org/10.1371/journal.pone.0148028

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M., 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. Nucleic Acids Res. 42, 633–642. https://doi.org/10.1093/nar/gkt1244

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72, 5069–5072. https://doi.org/10.1128/AEM.03006-05

Escapa, I.F., Huang, Y., Chen, T., Lin, M., Kokaras, A., Dewhirst, F.E., Lemon, K.P., 2020. Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. Microbiome 8, 65. https://doi.org/10.1186/s40168-020-00841-w

Ferguson, R.L., Buckley, E.N., Palumbo, A. V., 1984. Response of marine bacterioplankton to differential filtration and confinement. Appl. Environ. Microbiol. 47, 49–55. https://doi.org/10.1128/aem.47.1.49-55.1984

Fernandes, A.D., Reid, J.N.S., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., Gloor, G.B., 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, 1–13. https://doi.org/10.1186/2049-2618-2-15

Field, K.G., Olsen, G.J., Lane, D.J., Giovannoni, S.J., Ghiselin, M.T., Raff, E.C., Pace, N.R., Raff, R.A., 1988. Molecular phylogeny of the animal kingdom. Science (80-.). 239, 748–753. https://doi.org/10.1126/science.3277277

643 Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome datasets
644   are compositional: And this is not optional. Front. Microbiol. 8, 1–6.
645   https://doi.org/10.3389/fmicb.2017.02224

646 Gloor, G.B., Macklaim, J.M., Vu, M., Fernandes, A.D., 2016. Compositional uncertainty should
647   not be ignored in high-throughput sequencing data analysis. Austrian J. Stat. 45, 73–87.
648   https://doi.org/10.17713/ajs.v45i4.122

649 Gray, M.W., Sankoff, D., Cedergren, R.J., 1984. On the evolutionary descent of organisms and
650   organelles: A global phylogeny based on a highly conserved structural core in small subunit
651   ribosomal RNA. Nucleic Acids Res. 12, 5837–5852. https://doi.org/10.1093/nar/12.14.5837

652 Hodkinson, B.P., Grice, E.A., 2015. Next-Generation Sequencing: A Review of Technologies
653   and Tools for Wound Microbiome Research. Adv. Wound Care 4, 50–58.
654   https://doi.org/10.1089/wound.2014.0542

655 Hugerth, L.W., Andersson, A.F., 2017. Analysing microbial community composition through
656   amplicon sequencing: From sampling to hypothesis testing. Front. Microbiol. 8, 1561.
657   https://doi.org/10.3389/fmicb.2017.01561

658 Hughes, J.B., Hellmann, J.J., 2005. The application of rarefaction techniques to molecular
659   inventories of microbial diversity. Methods Enzymol. 397, 292–308.
660   https://doi.org/10.1016/S0076-6879(05)97017-1

661 Hughes, J.B., Hellmann, J.J., Ricketts, T.H., Bohannan, B.J.M., 2001. Counting the
662   Uncountable: Statistical Approaches to Estimating Microbial Diversity. Appl. Environ.
663   Microbiol. 67, 4399–4406. https://doi.org/10.1128/AEM.67.10.4399-4406.2001

664 Johnson, J.S., Spakowicz, D.J., Hong, B.Y., Petersen, L.M., Demkowicz, P., Chen, L., Leopold,
665   S.R., Hanson, B.M., Agresta, H.O., Gerstein, M., Sodergren, E., Weinstock, G.M., 2019.
666   Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.
667   Nat. Commun. 10, 1–11. https://doi.org/10.1038/s41467-019-13036-1

668 Kim, M., Morrison, M., Yu, Z., 2011. Evaluation of different partial 16S rRNA gene sequence
669   regions for phylogenetic analysis of microbiomes. J. Microbiol. Methods 84, 81–87.
670   https://doi.org/10.1016/j.mimet.2010.10.020

671 Kirisits, M.J., Emelko, M.B., Pinto, A.J., 2019. Applying biotechnology for drinking water
672   biofiltration: advancing science and practice. Curr. Opin. Biotechnol. 57, 197–204.
673   https://doi.org/10.1016/j.copbio.2019.05.009

674 Lam, T.Y.C., Mei, R., Wu, Z., Lee, P.K.H., Liu, W.T., Lee, P.H., 2020. Superior resolution
675   characterisation of microbial diversity in anaerobic digesters using full-length 16S rRNA
676   gene amplicon sequencing. Water Res. 178, 115815.
677   https://doi.org/10.1016/j.watres.2020.115815

678 Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A.,
679   Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G., Huttenhower,
680   C., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker
681   gene sequences. Nat. Biotechnol. 31, 814–821. https://doi.org/10.1038/nbt.2676

682 Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for

683   RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-
684   0550-8

685 Lozupone, C.A., Knight, R., 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci.
686   U. S. A. 104, 11436–11440. https://doi.org/10.1073/pnas.0611525104

687 McKnight, D.T., Huerlimann, R., Bower, D.S., Schwarzkopf, L., Alford, R.A., Zenger, K.R.,
688   2019. Methods for normalizing microbiome data: An ecological perspective. Methods Ecol.
689   Evol. 10, 389–400. https://doi.org/10.1111/2041-210X.13115

690 McMurdie, P.J., Holmes, S., 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is
691   Inadmissible. PLoS Comput. Biol. 10. https://doi.org/10.1371/journal.pcbi.1003531

692 McMurdie, P.J., Holmes, S., 2013. Phyloseq: An R Package for Reproducible Interactive
693   Analysis and Graphics of Microbiome Census Data. PLoS One 8, e61217.
694   https://doi.org/10.1371/journal.pone.0061217

695 Navas-Molina, J.A., Peralta-Sánchez, J.M., González, A., McMurdie, P.J., Vázquez-Baeza, Y.,
696   Xu, Z., Ursell, L.K., Lauber, C., Zhou, H., Song, S.J., Huntley, J., Ackermann, G.L., Berg-
697   Lyons, D., Holmes, S., Caporaso, J.G., Knight, R., 2013. Advancing our understanding of
698   the human microbiome using QIIME. Methods Enzymol. 531, 371–444.
699   https://doi.org/10.1016/B978-0-12-407863-5.00019-8

700 Paranjape, K., Bédard, É., Whyte, L.G., Ronholm, J., Prévost, M., Faucher, S.P., 2020. Presence
701   of Legionella spp. in cooling towers: the role of microbial diversity, Pseudomonas, and
702   continuous chlorine application. Water Res. 169, 115252.
703   https://doi.org/10.1016/j.watres.2019.115252

704 Perrin, Y., Bouchon, D., Delafont, V., Moulin, L., Héchard, Y., 2019. Microbiome of drinking
705   water: A full-scale spatio-temporal study to monitor water quality in the Paris distribution
706   system. Water Res. 149, 375–385. https://doi.org/10.1016/j.watres.2018.11.013

707 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O.,
708   2013. The SILVA ribosomal RNA gene database project: Improved data processing and
709   web-based tools. Nucleic Acids Res. 41, 590–596. https://doi.org/10.1093/nar/gks1219

710 R Core Team 2020. R: A language and environment for statistical computing. R Foundation for
711   Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

712 Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2009. edgeR: A Bioconductor package for
713   differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.
714   https://doi.org/10.1093/bioinformatics/btp616

715 Sanders, H.L., 1968. Marine Benthic Diversity: A Comparative Study. Am. Nat. 102, 243–282.

716 Schloss, P.D., Handelsman, J., 2004. Status of the Microbial Census. Microbiol. Mol. Biol. Rev.
717   64, 686–691. https://doi.org/10.1128/mmbr.68.4.686-691.2004

718 Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski,
719   R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van
720   Horn, D.J., Weber, C.F., 2009. Introducing mothur: Open-source, platform-independent,
721   community-supported software for describing and comparing microbial communities. Appl.

29

722   Environ. Microbiol. 75, 7537–7541. https://doi.org/10.1128/AEM.01541-09

723 Sepkoski, J.J., 1988. Alpha, beta, or gamma: Where does all the diversity go? Paleobiology 14,
724   221–234. https://doi.org/10.1017/S0094837300011969

725 Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Techinical
726   Journal, 27:369-423, 623-656.

727 Shaw, J.L.A., Monis, P., Weyrich, L.S., Sawade, E., Drikas, M., Cooper, A.J., 2015. Using
728   amplicon sequencing to characterize and monitor bacterial diversity in drinking water
729   distribution systems. Appl. Environ. Microbiol. 81, 6463–6473.
730   https://doi.org/10.1128/AEM.01297-15

731 Shokralla, S., Spall, J.L., Gibson, J.F., Hajibabaei, M., 2012. Next-generation sequencing
732   technologies for environmental DNA research. Mol. Ecol. 21, 1794–1805.
733   https://doi.org/10.1111/j.1365-294X.2012.05538.x

734 Silverman, J., Roche, K., Mukherjee, S., David, L., 2018. Naught all zeros in sequence count
735   data are the same. bioRxiv 477794. https://doi.org/10.1101/477794

736 Simpson, E.H., 1949. Measurment of Diversity. Nature 163, 688.

737 Thomas, T., Gilbert, J., Meyer, F., 2012. Metagenomics - a guide from sampling to data analysis.
738   Microb. Inform. Exp. 2, 3. https://doi.org/10.1186/2042-5783-2-3

739 Tromas, N., Fortin, N., Bedrani, L., Terrat, Y., Cardoso, P., Bird, D., Greer, C.W., Shapiro, B.J.,
740   2017. Characterising and predicting cyanobacterial blooms in an 8-year amplicon
741   sequencing time course. ISME J. 11, 1746–1763. https://doi.org/10.1038/ismej.2017.58

742 Tsilimigras, M.C.B., Fodor, A.A., 2016. Compositional data analysis of the microbiome:
743   fundamentals, tools, and challenges. Ann. Epidemiol. 26, 330–335.
744   https://doi.org/10.1016/j.annepidem.2016.03.002

745 Tsukuda, M., Kitahara, K., Miyazaki, K., 2017. Comparative RNA function analysis reveals high
746   functional similarity between distantly related bacterial 16 S rRNAs. Sci. Rep. 7, 1–8.
747   https://doi.org/10.1038/s41598-017-10214-3

748 Vierheilig, J., Savio, D., Farnleitner, A.H., Reischer, G.H., Ley, R.E., Mach, R.L., Farnleitner,
749   A.H., Reischer, G.H., 2015. Potential applications of next generation DNA sequencing of
750   16S rRNA gene amplicons in microbial water quality monitoring. Water Sci. Technol. 72,
751   1962–1972. https://doi.org/10.2166/wst.2015.407

752 Walters, W., Hyde, E.R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert,
753   J.A., Jansson, J.K., Caporaso, J.G., Fuhrman, J.A., Apprill, A., Knight, R., 2015. Improved
754   Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker
755   Gene Primers for Microbial Community Surveys. mSystems 1, e0009-15.
756   https://doi.org/10.1128/mSystems.00009-15.Editor

757 Wang, Y., LêCao, K.-A., 2019. Managing batch effects in microbiome data. Brief. Bioinform.
758   https://doi.org/10.1093/bib/bbz105

759

760 Weisburg, W.G., Barns, S.M., Pelletier, D.A., Lane, D.J., 1991. 16S Ribosomal DNA
761   Amplification for Phylogenetic Study. J. Bacteriol. 173, 697–703.

762 Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld,
763   J.R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E.R., Knight, R., 2017. Normalization and
764   microbial differential abundance strategies depend upon data characteristics. Microbiome 5,
765   1–18. https://doi.org/10.1186/s40168-017-0237-y

766 Willis, A.D., 2019. Rarefaction, alpha diversity, and statistics. Front. Microbiol. 10.
767   https://doi.org/10.3389/fmicb.2019.02407

768 Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms:
769   Proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A.
770   87, 4576–4579. https://doi.org/10.1073/pnas.87.12.4576

771 Yang, B., Wang, Y., Qian, P.Y., 2016. Sensitivity and correlation of hypervariable regions in
772   16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17, 1–8.
773   https://doi.org/10.1186/s12859-016-0992-y

774 Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., Zhang, H., Xiong, Z., Xue, Y., Tu, J.,
775   Lu, Z., 2018. Evaluation of different 16S rRNA gene V regions for exploring bacterial
776   diversity in a eutrophic freshwater lake. Sci. Total Environ. 618, 1254–1267.
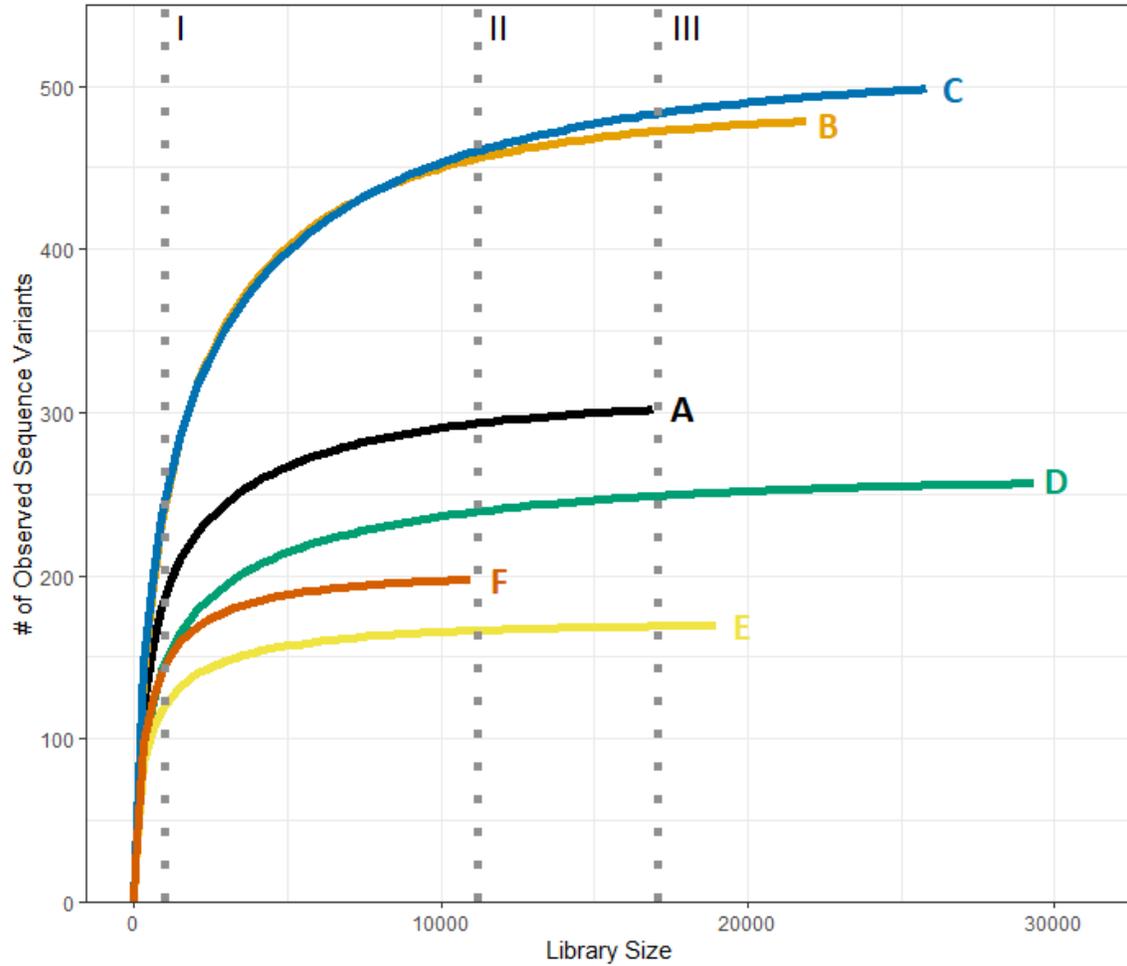777   https://doi.org/10.1016/j.scitotenv.2017.09.228

778 Zhang, L., Fang, W., Li, X., Lu, W., Li, J., 2020. Strong linkages between dissolved organic
779   matter and the aquatic bacterial community in an urban river. Water Res. 184, 116089.
780   https://doi.org/10.1016/j.watres.2020.116089

781

782
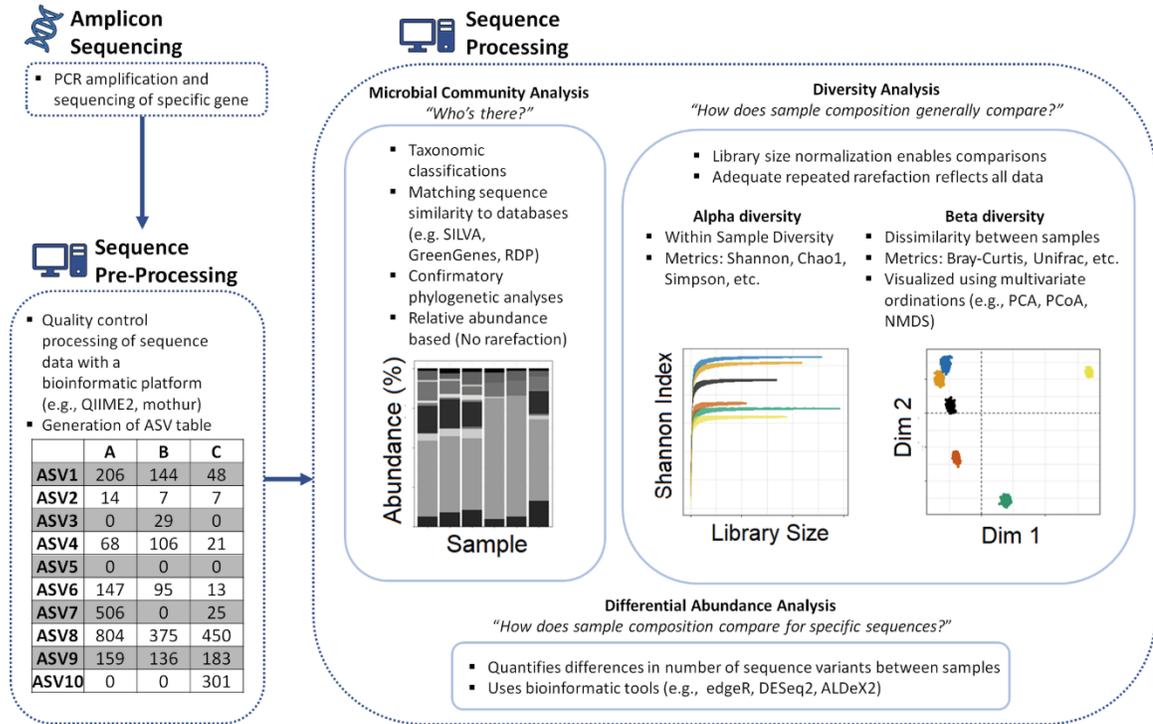
**Figure 1 Schematic of general workflow in amplicon sequencing of samples.**

Figure 2 Rarefaction curves showing the number of unique sequence variants as a function of normalized library size for six samples (labelled A – F) of varying diversity and initial library size. Selection of unnecessarily small library sizes (I) omits many sequence variants. Rarefying to the smallest library size (II) omits fewer sequences and variants. While selection of a larger normalized library size (III) would omit even less sequences, it is necessary to omit entire samples (e.g., Sample F) that have too few sequences)
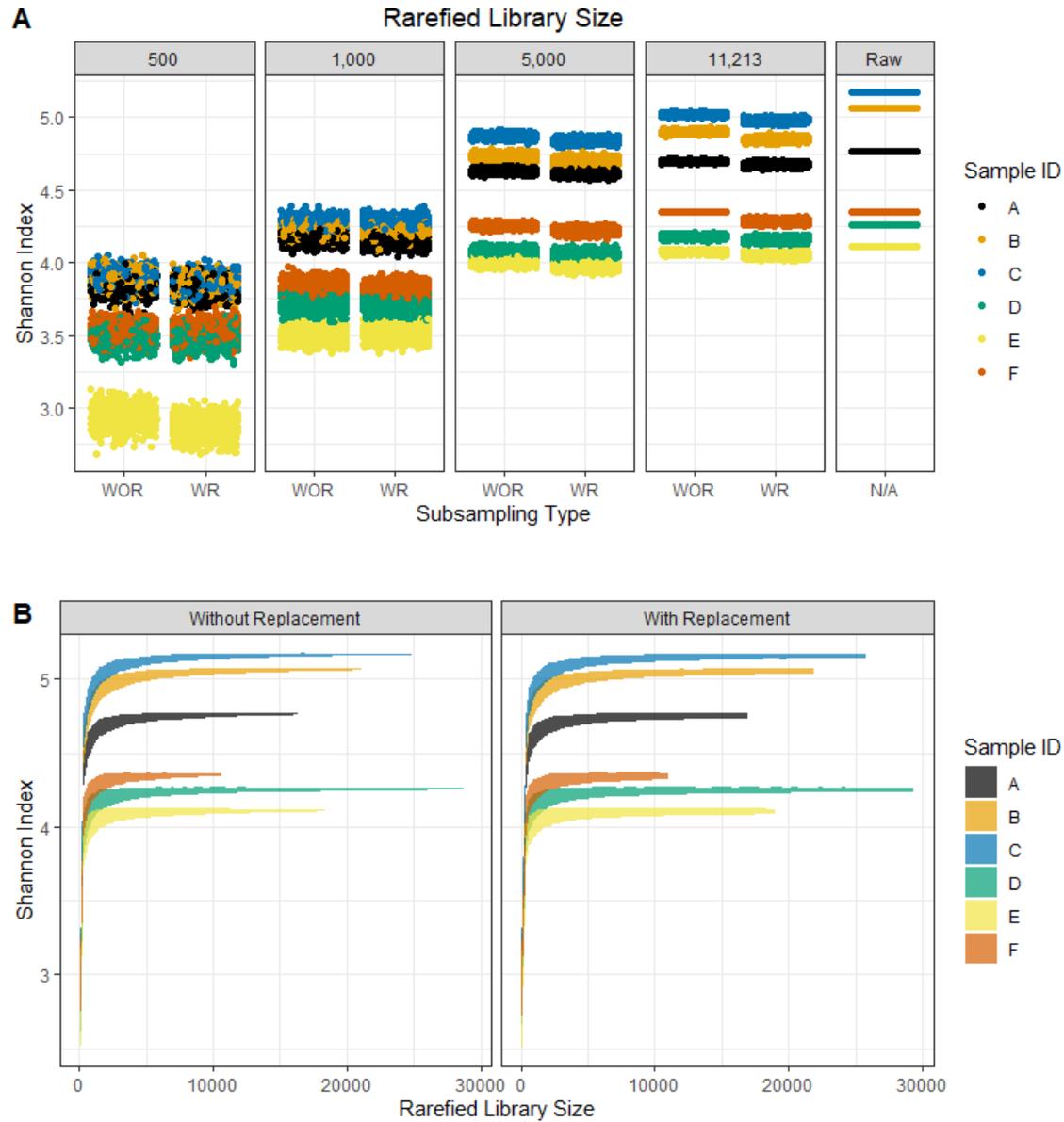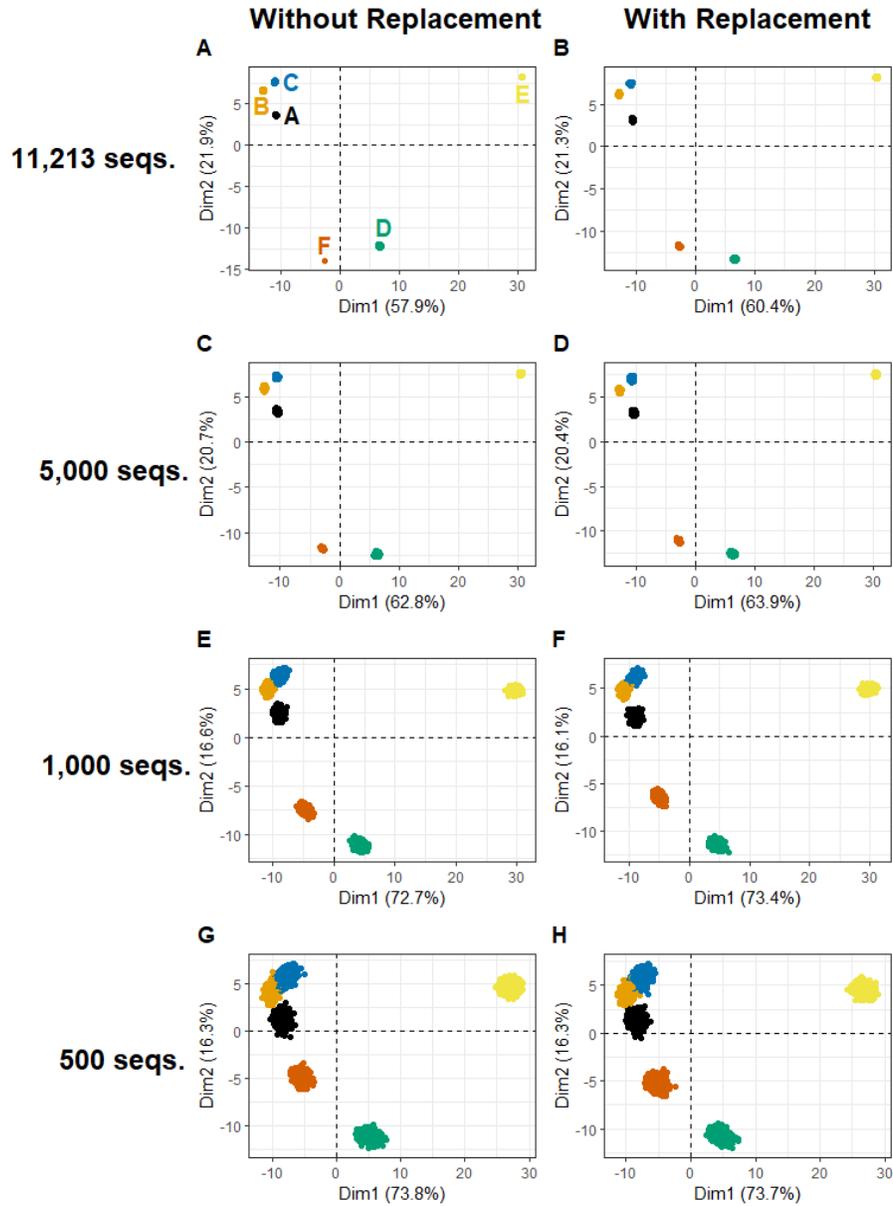
a) Without Replacement
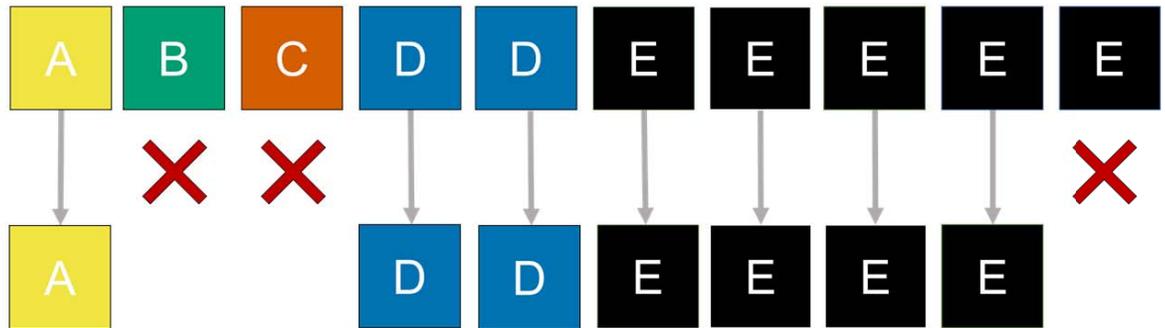
b) With Replacement
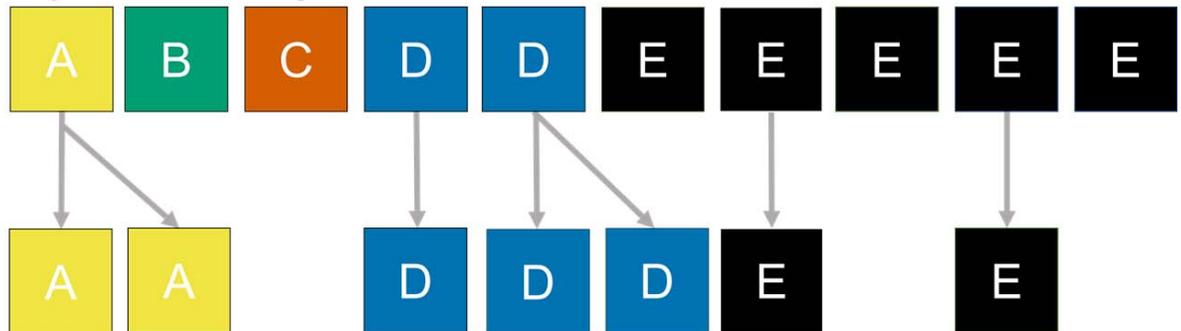
Figure 3 The mechanics of rarefying with or without replacement for a hypothetical sample with a library size of ten composed of five sequence variants (A – E). Rarefying without replacement (a) draws a subset from the observed library excluding the complementary subset, while rarefying with replacement (b) has the potential to artificially inflate the numbers of some sequence variants beyond what was observed.

Figure 4 Effect of chosen rarefied library size and sampling with (WR) or without (WOR) replacement upon the Shannon Diversity Index. Six microbial communities were rarefied repeatedly (A) at specific rarefied library sizes of 11,213 sequences, 5,000 sequences, 1,000 sequences, and 500 sequences and (B) to evaluate the Shannon Index as a function of rarefied library size.

Figure 5 Variation in PCA ordinations (using the Bray-Curtis dissimilarity on Hellinger transformed rarefied libraries) of six microbial communities repeatedly rarefied with and without replacement to varying library sizes.
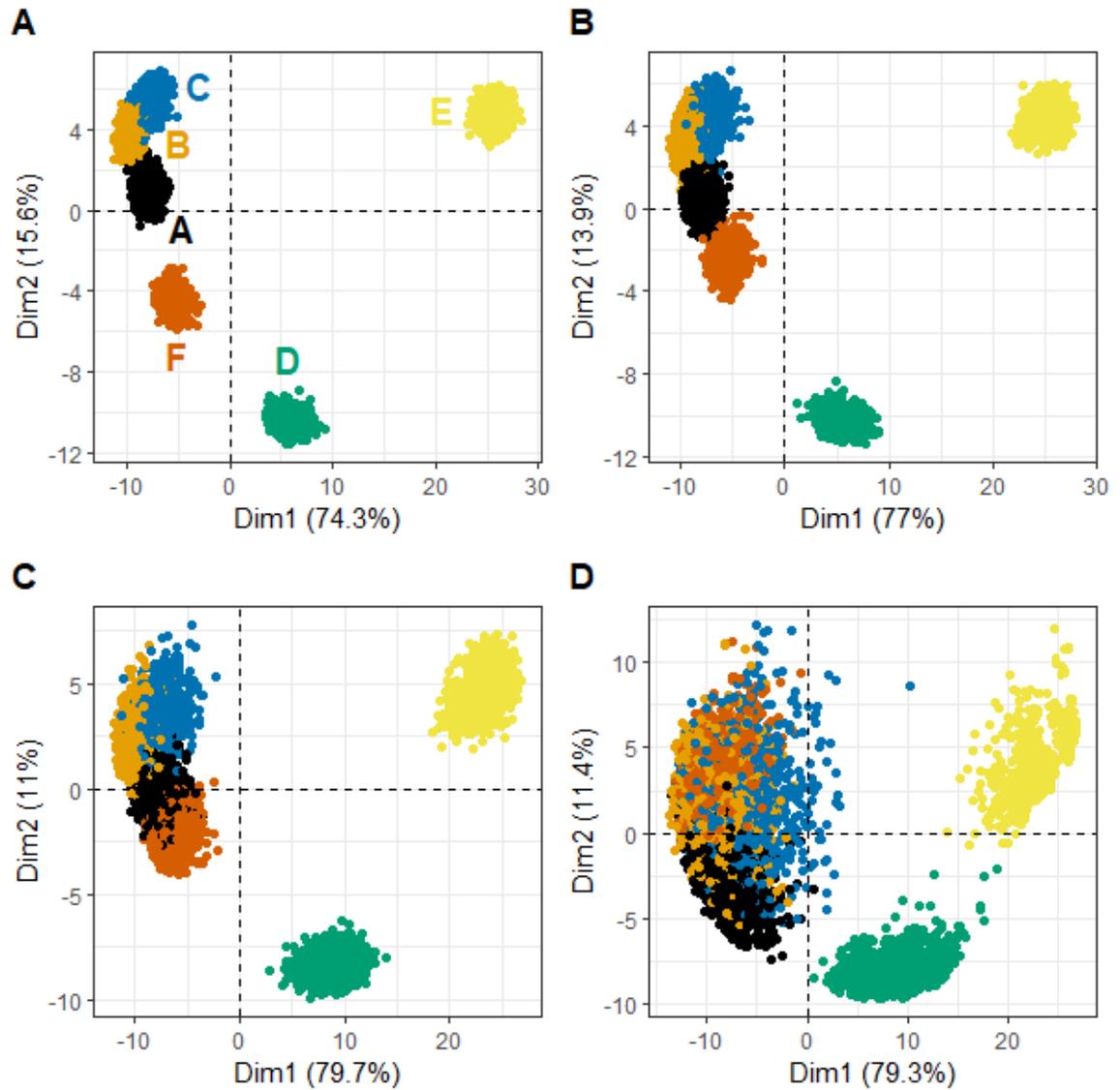
Figure 6 Variation in PCA ordinations (using the Bray-Curtis dissimilarity on Hellinger transformed rarefied microbial communities) of six microbial communities repeatedly rarefied to very small library sizes of (A) 400, (B) 300, (C) 200 and (D) 100 sequences.