Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches

Jacob T Nearing [1] , **Gavin M Douglas** [1] , **André M Comeau** [2] , **Morgan G.I Langille** Corresp. [1, 3]

[1] Department of Microbiology and Immunology, Dalhousie University, Halifax, Nova Scotia, Canada

[2] Integrated Microbiome Resource, Dalhousie University, Halifax, Nova Scotia, Canada

[3] Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada

Corresponding Author: Morgan G.I Langille
Email address: morgan.langille@dal.ca

High-depth sequencing of universal marker genes such as the 16S rRNA gene are a common strategy to profile microbial communities. Traditionally, sequence reads are clustered into operational taxonomic units (OTUs) at a defined identity threshold to avoid sequencing errors generating spurious taxonomic units. However, there have been numerous bioinformatic packages recently released that attempt to correct sequencing errors to determine real biological sequences at single nucleotide resolution by generating amplicon sequence variants (ASVs). As more researchers begin to use high resolution ASVs, there is a need for an in-depth and unbiased comparison of these novel "denoising" pipelines. In this study, we conduct a thorough comparison of three of the most widely-used denoising packages (DADA2, UNOISE3, and Deblur) as well as an open-reference 97% OTU clustering pipeline on mock, soil, and host-associated communities. We found from the mock community analyses that although they produced similar microbial compositions based on relative abundance, the approaches identified vastly different numbers of ASVs that significantly impact alpha diversity metrics. Our analysis on real datasets using recommended settings for each denoising pipeline also showed that the three packages were consistent in their per-sample compositions, resulting in only minor differences based on weighted UniFrac and Bray-Curtis dissimilarity. DADA2 tended to find more ASVs than the other two denoising pipelines when analyzing both the real soil data and two other host-associated datasets, suggesting that it could be better at finding rare organisms, but at the expense of possible false positives. The open-reference OTU clustering approach identified considerably more OTUs in comparison to the number of ASVs from the denoising pipelines in all datasets tested. The three denoising approaches were significantly different in their run times, with UNOISE3 running greater than 1200 and 15 times faster than DADA2 and Deblur, respectively. Our findings indicate that, although all pipelines result in similar general community structure, the number of ASVs/OTUs and

resulting alpha-diversity metrics varies considerably and should be considered when attempting to identify rare organisms from possible background noise.

Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches

6 Jacob T. Nearing[1], Gavin M. Douglas[1], André M. Comeau[2], Morgan G. I. Langille[1,2,3]

8 [1] Department of Microbiology & Immunology, Dalhousie University, Halifax, NS, Canada

10 [2] Integrated Microbiome Resource, Dalhousie University, Halifax, NS, Canada

12 [3] Department of Pharmacology, Dalhousie University, Halifax, NS, Canada

15 Corresponding Author:

17 Morgan Langille

21 Email address: morgan.langille@dal.ca

Abstract

High-depth sequencing of universal marker genes such as the 16S rRNA gene are a common

strategy to profile microbial communities. Traditionally, sequence reads are clustered into

operational taxonomic units (OTUs) at a defined identity threshold to avoid sequencing errors

generating spurious taxonomic units. However, there have been numerous bioinformatic

packages recently released that attempt to correct sequencing errors to determine real biological

sequences at single nucleotide resolution by generating amplicon sequence variants (ASVs). As

more researchers begin to use high resolution ASVs, there is a need for an in-depth and unbiased

comparison of these novel "denoising" pipelines. In this study, we conduct a thorough

comparison of three of the most widely-used denoising packages (DADA2, UNOISE3, and

Deblur) as well as an open-reference 97% OTU clustering pipeline on mock, soil, and host-

associated communities. We found from the mock community analyses that although they

produced similar microbial compositions based on relative abundance, the approaches identified

vastly different numbers of ASVs that significantly impact alpha diversity metrics. Our analysis

on real datasets using recommended settings for each denoising pipeline also showed that the

three packages were consistent in their per-sample compositions, resulting in only minor

differences based on weighted UniFrac and Bray-Curtis dissimilarity. DADA2 tended to find

more ASVs than the other two denoising pipelines when analyzing both the real soil data and

two other host-associated datasets, suggesting that it could be better at finding rare organisms,

but at the expense of possible false positives. The open-reference OTU clustering approach

identified considerably more OTUs in comparison to the number of ASVs from the denoising

pipelines in all datasets tested. The three denoising approaches were significantly different in

46   their run times, with UNOISE3 running greater than 1200 and 15 times faster than DADA2 and

47   Deblur, respectively. Our findings indicate that, although all pipelines result in similar general

48   community structure, the number of ASVs/OTUs and resulting alpha-diversity metrics varies

49   considerably and should be considered when attempting to identify rare organisms from possible

50   background noise.

51

52   **Introduction**

53

54   Microbiome studies often use an amplicon sequencing approach where a single genomic region

55   is sequenced at a sufficient depth to provide relative abundance profiles of the microbes present

56   in a sample. The 16S rRNA gene (16S) is usually chosen as a marker gene for sequencing of

57   bacterial communities due to its unique structure that contains both conserved and variable

58   regions and presence in all known Bacteria and Archaea species. This sequencing approach is

59   often used to avoid the high cost of shotgun metagenomic sequencing or to avoid problems with

60   sequencing non-microbial DNA from host contamination. However, sequencing errors make it

61   difficult to distinguish biologically real nucleotide differences in 16S sequences from sequencing

62   artefacts. To avoid this issue sequences are often clustered into operational taxonomic units

63   (OTUs) at a particular identity threshold (e.g. 97%) to avoid the problem of differentiating

64   biological from technical sequence variations; however, this comes at the cost of taxonomic

65   resolution. Recently, many new bioinformatic sequence "denoising" approaches have been

66   developed to address this issue by attempting to correct sequencing errors thus improving

67   taxonomic resolution. These pipelines differ in how they correct sequencing errors. DADA2

68   generates a parametric error model that is trained on the entire sequencing run and then applies

69   that model to correct and collapse the sequence errors into what the authors call amplicon

70   sequence variants (ASVs) (Callahan et al., 2016). This approach is advantageous as it builds

71   unique error models for each sequencing run. Deblur aligns sequences together into "sub-OTUs"

72   and, based on an upper error rate bound along with a constant probability of indels and the mean

73   read error rate, removes predicted error-derived reads from neighboring sequences (Amir et al.,

74   2017). Deblur employs a sample-by-sample approach which reduces both memory requirements

75   and computational demand. UNOISE3 uses a one-pass clustering strategy that does not depend

76   on quality scores, but rather two parameters with pre-set values that were curated by its author to

77   generate "zero-radius OTUs" (Edgar, 2016). The advantage of a one-pass clustering strategy is

78   that it saves on the computational time required to analyze the sequences in the provided study.

79   Note that ASVs, sub-OTUs, and zero-radius OTUs are synonymous and the term ASV will be

80   used henceforth. Denoising approaches provide improved resolution and they avoid having to

81   make a choice between various OTU strategies which may result in differing results (Edgar,

82   2017). In addition, ASVs can be identified by their unique biological sequences instead of

83   relying on per-study IDs, which allows for easier comparison across datasets (Callahan,

84   McMurdie & Holmes, 2017).

85   Although there have been several bioinformatic comparisons of OTU-based approaches

86   (Allali et al., 2017; Plummer & Twin, 2015), a thorough third-party comparison of denoising

87   pipelines has yet to be conducted. In this paper, we compare the strengths and weaknesses of the

88   DADA2, UNOISE3, and Deblur packages along with a comparison to an open-reference 97%

89   OTU-based approach (Rognes et al., 2016), while following the recommended default settings.

90   We assess the accuracy of these approaches using several mock communities including both

91 bacterial and fungal amplicons. In addition, we compare the results of the four pipelines on three

92 previously-published real human, mouse, and soil datasets.

93

94 **Methods**

95 **Sequence Acquisition**

96 The HMP mock community (which has even expected abundances) and the ZymoBIOMICS

97 Microbial Community Standard (referred to as the Zymomock community) were sequenced by

98 the Integrated Microbiome Resource at Dalhousie University using an Illumina MiSeq on

99 separate sequencing runs, as previously described using the V4-V5 16S rRNA gene region

100 (Comeau et al., 2017). Reads were then uploaded to the European Nucleotide Archive (ENA)

101 under accession number PRJEB24409. The Extreme dataset (mock-12) originally presented in

102 the DADA2 paper and the fungal ITS1 dataset (mock-9) were retrieved from the Mockrobiota

103 project (Bokulich et al., 2016). The Extreme dataset was sequenced using an Illumina MiSeq

104 (Callahan et al., 2016) and the fungal mock community was sequenced using an Illumina HiSeq

105 (Bokulich et al., 2016). The soil data set collected from blueberry fields (available under NCBI

106 SRA PRJNA389786) and the exercise dataset collected from stool of mice that exercised plus

107 controls (ENA accession PRJEB18615) (Lamoureux et al., 2017) as well as the human-

108 associated dataset of intestinal biopsies of pediatric Crohn's disease patients plus controls (ENA

109 accession PRJEB21933) (Douglas et al., 2018) were sequenced at the Integrated Microbiome

110 Resource at Dalhousie University.

111

112 **Filtering**

113 All sample data was filtered using the Microbiome Helper filtering scripts (Comeau et al., 2017).

114 In summary, primers were trimmed off all reads using Cutadapt (v 1.14) (Martin, 2011) and

115 GNU Parallel (Tange, 2011). Primer-free sequences were then input into the dada2_filter.R

116 script available in Microbiome Helper. This script takes in the maximum expected number of

117 errors allowed as well as a truncation length. The HMP mock community and the Zymomock

118 community were truncated to read lengths of 270 and 210 base pairs for the forward and reverse

119 reads respectively to remove low quality bases at the end of the reads. The shorter length for the

120 reverse reads is a result of their lower overall quality compared to the forward reads. Note that

121 for DADA2, and Deblur to work properly reads need to be of the same length. The single-end

122 reads from the Extreme mock community and the fungal mock community were truncated to 80

123 base-pairs. The three real datasets: soil, mouse, and human-associated, were truncated to 270 and

124 210 base-pairs for the forward and reverse reads, respectively. The number of expected errors

125 allowed were defined as three different filtering stringencies: 5 (low), 3 (medium), and 1 (high).

126

127 **DADA2 Pipeline**

128 The DADA2 pipeline was run using Microbiome Helper scripts, which wraps the core

129 algorithms of the DADA2 pipeline (Callahan et al., 2016). Filtered reads were input into the

130 wrapper script dada2_inference.R which runs the DADA2 inference algorithm. Once ASVs are

131 determined, they are passed into DADA2's chimera-checking algorithm which was run using the

132 wrapper script dada2_chimera_taxa.R to screen out chimeric sequences. The output objects

133 containing ASV sequences and abundances counts were then converted into BIOM table format

134 using convert_dada2_out.R. All DADA2 wrapper scripts were run with default settings.

135

136 **UNOISE3 Pipeline**

137   Filtered reads were input into USEARCH's (v 10) (Edgar, 2010) fastq_mergepairs command if

138   they were paired-end reads or concatenated together into one FASTQ if they were single-end

139   reads. Next, the merged FASTQ file was converted into a FASTA file using the Microbiome

140   Helper script run_fastq_to_fasta.pl and then used as input for USEARCH's fastx_uniques

141   command which generated a FASTA containing all the unique sequences and their abundance.

142   Finally, the FASTA containing unique sequences was used as input into USEARCH's unoise3

143   (Edgar, 2016) command generating a BIOM table and representative ASVs that were used in

144   subsequent analyses. All USEARCH scripts were run with default settings.

145

146   **Deblur Pipeline**

147   Paired-end filtered reads were stitched together using the Microbiome Helper wrapper script

148   run_pear.pl which wraps the program PEAR (v 0.9.10) (Zhang et al., 2014). This step was

149   skipped for filtered single-end reads. Next, reads were renamed to match a format that was

150   compatible with QIIME2 (Caporaso et al., 2010b) and converted into a QIIME2 artifact. Samples

151   were then run through QIIME2's built-in deblur command using the 16S setting, which uses

152   Greengenes 13_8 (DeSantis et al., 2006) for positive filtering. A separate analysis to test the

153   effect of positive filtering was conducted by using the non-positive filtered output files

154   ("all.biom" and "all.seqs.fa") from the stand alone version of Deblur. Fungal reads were run

155   using the "other" setting and the UNITE 10.10.2017 database (Kõljalg et al., 2013). Finally, the

156   representative ASV sequences and a BIOM table were exported from the QIIME2 artifact. All

157   other Deblur scripts were run with default settings.

158

159   **Open Reference OTU Pipeline**

160   Filtered reads were stitched and imported into a QIIME2 artifact as previously described in the

161   Deblur pipeline. The imported sequences were then dereplicated using the QIIME2 VSEARCH

162   plugin. Dereplicated sequences were filtered for chimeras using the QIIME2 VSEARCH

163   UCHIME reference-based chimera filtering (Edgar et al., 2011) using the Greengenes13_8 97%

164   OTU database for reference. Chimera-checked sequences were then clustered at 97% in an open-

165   reference fashion using the QIIME2 VSEARCH cluster-features-open-reference plugin. The

166   Greengenes 13_8 97% OTU database was used as the reference database during open-reference

167   OTU picking. Finally, singleton OTUs were removed from the OTU table and the table was

168   exported in BIOM format along with its representative sequences in FASTA format.

169

170   **Run Time and Memory Analysis**

171   The soil dataset was filtered using the low-stringency filter and then individual samples were

172   rarefied to either 5000, 10000, 20000 or 30000 reads per sample. The different read-depth sets

173   were then run through the three denoising pipelines and user time and maximum memory usage

174   was determined using the GNU time (v 1.7) command.

175

176   **ASV and OTU Analysis of Mock Communities**

177   ASVs and OTUs were compared against the expected sequences provided with each of the mock

178   communities. This comparison was done using the command-line BLASTN (v 2.7.10) (Altschul

179   et al., 1990) tool against the expected sequences from each community. This allowed us to

180   determine the number of full length 100% matches and 97% matches. All ASVs and OTUs that

181   did not match these criteria were then compared against the SILVA 16S rRNA gene database (v

182   128) (Pruesse et al., 2007) or the UNITE 10.10.2017 database to find all full length 100% and

183   97% matches. Any ASVs or OTUs that did not match these databases were then labeled as

184   "Unmatched". The number of unique expected sequences for each mock community was

185   determined by slicing out the amplified regions using a custom Python (v 3.6.1) script

186   (slice_amplified_region.py, available on GitHub) from the expected sequences of each mock

187   community and then finding the number of unique sequences from this output using

188   USEARCH10's fastx_uniques command. In addition to analysis done with no abundance

189   filtering, we also compared how filtering of low abundance ASVs/OTUs affected the type and

190   amount of ASVs/OTUs called by each pipeline. This was done by applying a 0.1% minimum

191   abundance filter to each approach and dataset except for the Extreme mock community where a

192   0.0004% minimum abundance filter was applied. The 0.1% minimum abundance filter was

193   chosen based on the known 0.1% bleed through between Illumina MiSeq runs. A lower

194   minimum abundance filter of 0.0004% was chosen for the Extreme mock community since some

195   of the expected sequences had a lower expected abundance than 0.1%. Histograms with

196   sequence identity to the expected sequences were generated for each pipeline and each mock

197   community except for the fungal mock community due to the community missing some expected

198   sequences from its expected sequence list. All sequences that showed less than 75% identity

199   were binned together.

200

201   **Abundance Data Analysis of Mock Communities**

202   For the HMP, Zymomock and Extreme datasets all ASVs/OTUs that matched at 97% identity or

203   greater with the provided expected sequences based on a BLASTN search were added to the

204   abundance of the corresponding expected taxa. Stacked bar charts of expected taxa relative

205   abundances were created using the ggplot2 (v 2.2.1) (Wickham, 2009) R (v 3.4.3) (R

206 Development Core Team, 2008) package and the cowplot (v 0.9.2) R package (see Data

207 Deposition for exact scripts).

208   Due to the incomplete nature of the expected sequences for the fungal mock community,

209 UNITE database hits at 97% or greater to an expected species within the community were

210 considered as expected ASVs/OTUs. All other ASVs/OTUs were classified as "Non-Reference"

211 hits.

212

213 **Analysis of Real Datasets**

214 Sequences from the three real datasets; soil, human-associated, and exercise were filtered using

215 medium stringencies (allowing up to 3 expected errors per sequence) for each package and

216 rarified to 5215, 3000 and 4259 reads, respectively. These rarefaction levels were chosen as they

217 were the lowest read count above 2000, after the DADA2 pipeline was complete. ASV

218 abundance tables output by all approaches were combined into a single table where each

219 biological sample was represented multiple times (once for each pipeline). ASVs/OTUs not

220 called by a specific pipeline were given an abundance of zero in their column (e.g. ASVs only

221 called by Deblur for sampleA were given zero abundances in the columns for DADA2's and

222 UNOISE3's outputs of sampleA). Representative sequences were aligned against the Greengenes

223 13_5 99% OTU aligned sequences and placed on the Greengenes 13_5 99% OTU tree using

224 SEPP (Mirarab et al., 2011). Weighted UniFrac and unweighted UniFrac distances at the

225 ASV/OTU level were then generated using the QIIME1 beta_diversity.py command and

226 principal coordinates were generated using the QIIME1 prinicpal_cordinates.py command.

227   The Bray-Curtis distance matrix at the genus level was generated by assigning taxonomy

228 to the resulting ASV/OTUs from each package using the RDP classifier (Cole et al., 2014) with

229 the assignTaxonomy function available in the DADA2 package and the rdp_train_set_16

230   database. Distances were then generated using the summarize_taxa.py and beta_diversity.py

231   commands in QIIME1. Ordination was generated using the metaMDS function in the vegan R

232   package (Oksanen et al., 2018).

233   Mantel correlations between each distance matrix for each pipeline and each beta-

234   diversity metric (weighted/unweighted UniFrac and Bray-Curtis) were generated using the vegan

235   R package. The number of observed OTUs/ASVs per sample for each real dataset were

236   generated using the rarified combined OTU/ASV tables and the alpha_diversity.py command in

237   QIIME1 using observed OTUs as the metric option.

238

239   **Results**

240

241   **Differences in methodology and availability between different denoising pipelines**

242   There are important aspects other than accuracy that need to be considered when determining

243   which denoising package a researcher should use for their project. Both DADA2 and UNOISE3

244   are suggested to be run in a pooled-sample workflow, where all sequences are pooled together

245   during the denoising process (**Table 1**). This allows them to better account for batch errors

246   across multi-run experiments. Deblur, on the other hand, runs its denoising process sample-by-

247   sample. This approach helps lower Deblur's computational requirements, but at the cost of

248   reducing its ability to correct multi-run batch effects. Both DADA2 and Deblur are open source

249   projects, whereas UNOISE3 is a closed-source project which has a free 32-bit academic version

250   with a 4 Gb memory cap and a full 64-bit version that costs between $885-1485 USD (**Table 1**).

251   Another major difference is that the built-in Deblur function in QIIME2 has a positive filtering

252   process. This default setting causes Deblur to discard reads that do not reach a length-scaled bit

253   score and an e-value threshold to any sequence in the 88% representative sequences Greengenes

254   database. Note the default database can be changed using the "other" version of the Deblur

255   plugin in QIIME2, an important feature when working with fungal or eukaryotic data. It is also

256   important to note that the stand-alone version of Deblur outputs both positively filtered and non-

257   filtered results by default, unlike the QIIME2 plugin which, only outputs the positively filtered

258   results. Currently, the functionality of both DADA2 and Deblur can be accessed through a

259   graphical user interface as plugins in QIIME2 Studio, whereas UNOISE3 does not support a

260   graphical user interface (**Table 1**).

261

262   **Total number of ASVs/OTUs vary across approaches in mock communities**

263   We processed four different mock communities with the DADA2, UNOISE3, and Deblur

264   denoising pipelines as well as an open-reference 97% OTU clustering pipeline to compare the

265   resulting ASVs/OTUs from each approach. Focusing on the number of called ASVs from each

266   denoising pipeline, we found no approach consistently called more ASVs. DADA2 called the

267   most ASVs in two mock communities (HMP: 42, Extreme: 78) and UNOISE3 called the most

268   ASVs in the other two mock communities (Fungal: 38, Zymomock: 43) under medium

269   stringency filtering (**Fig 1, Supp Table 1**). Overall, open-reference OTU clustering output the

270   most ASVs/OTUs for all four mock communities (HMP: 453, Extreme: 8891, Fungal: 96,

271   Zymomock: 294). None of the approaches were capable at outputting all expected sequences at

272   100% identity in any of the mock communities that were processed and, in all datasets, at least

273   one denoising pipeline output more ASVs than expected sequences. All four approaches output

274   at least one ASV/OTU at 97% or greater identity from all organisms in the HMP mock

275   community and the Zymomock community (**Supp Table 2,3**). DADA2 output nine more ASVs

276   with 97% or greater identity matches to expected sequences in the Extreme dataset (read depth of

277   11.6 million reads) than the other two denoising pipelines (**Supp Table 4)**. Six of the nine taxa

278   that DADA2 called and the other pipelines did not call, had expected relative abundances of only

279   0.000427% (**Supp Table 4**). Of the other three taxa not found by Deblur or UNOISE3, one was

280   at an expected abundance of 0.00427% and two had an expected abundance of 0.0427% (**Supp**

281   **Table 4**). Interestingly, open reference OTU clustering called OTUs corresponding to these nine

282   expected taxa along with two more expected taxa that DADA2 missed, both of which were in the

283   lowest possible expected abundance range (0.000427%). One organism was missed by all

284   pipelines in the Extreme community (*C. methylpentusum*), which was also in the lowest possible

285   expected abundance range. Sequence identity histograms were constructed for the HMP,

286   Zymomock and Extreme communities (**Supp Fig 1-3**). In the HMP community, no ASV was

287   found to have below 80% identity by the denoising pipelines, but open reference OTU clustering

288   found 21 OTUs below 80% identity. Furthermore, the OTU pipeline called 324 OTUs below

289   99% identity whereas all denoising approaches called less than 15 ASVs (**Supp Fig 1**). In the

290   Extreme community, most ASVs were found in the 99-100% identity range whereas the majority

291   of OTUs found by open-reference OTU clustering were in the 95-97% identity range (**Supp Fig**

292   **2**).

293   Due to the expected sequences for the fungal community being incomplete, we included

294   any UNITE database hits as expected sequences if the matching sequence was from a species

295   that was included in the fungal community. This resulted in almost all the fungi present in the

296   community to be found by all four pipelines except for *Penicillium allii* and *P. commune*. Open-

297   reference OTU clustering did find *P. allii*, although at an Extremely low abundance of 0.004%

298   compared to its expected abundance of 6.25%.

299   Given that some of the above potential spurious ASVs would be removed by sequence

300   bleed-through (Illumina, 2017) or low-abundance filters in typical workflows, we applied an

301   abundance cutoff filter of 0.1% to each mock community except for the Extreme community,

302   where an abundance filter of 0.0004% was applied (**Supp Fig 4**). This lower abundance filter

303   was chosen due to some expected sequences being in abundance of 0.000427%. Application of

304   the abundance filter to the HMP community resulted in all 10 unmatched ASVs (those that did

305   not match either the expected or SILVA by 97% or greater) called by DADA2 to be discarded,

306   but none of the four unmatched reads in UNOISE3 to be discarded (**Supp Fig 4**). A similar

307   phenomenon was seen in the Zymomock community with all 12 of Deblur's unmatched reads

308   being discarded (along with one database hit) and UNOISE3 only discarding one of 19

309   unmatched reads it called. The largest effect that the application of the filters had was the

310   removal of a significant amount of OTUs found in each mock community by the open-reference

311   OTU pipeline. The most drastic change was seen in the Extreme community that started with

312   8891 OTUs and was reduced to 1248 OTUs (**Supp Fig 4**). In the fungal community the number

313   of OTUs found by open-reference OTU clustering (27) became less than the number of ASVs

314   found by UNOISE3 (36).

315   To determine how read quality filtering affects the number of ASVs called by each

316   pipeline, we ran each denoising pipelines using two additional quality filtering stringencies, low

317   and high (see Methods). The different quality filter stringencies used made only small impacts on

318   the numbers of ASVs called by each pipeline for the HMP, Extreme and fungal datasets (**Supp**

319   **Table 1**). A difference of six ASVs was the largest between the high and medium stringencies on

320   the HMP community and was output by UNOISE3 (**Supp Table 1**). In the Zymomock

321   community, the number of ASVs called by DADA2 only varied by one for all three stringencies,

322   but Deblur varied by as much as 12 ASVs and UNOISE3 varied by as much as 16 ASVs being

323   outputted between the high and medium filter stringencies (**Supp Table 1**).

324

325   **Denoising pipelines are consistent in determining mock community composition**

326   Despite the different ASV counts between each pipeline in the mock community, the relative

327   abundances of the expected taxa are strikingly similar (**Fig. 2**). In both the HMP and Zymomock

328   communities, only a small portion of ASVs called by DADA2 and Deblur did not match the

329   expected sequences by 97% identity or greater. In contrast, UNOISE3 identified multiple (8 in

330   HMP, 20 in Zymomock) sequences that summed together to make up 2.9% and 4.8% of the

331   relative abundance in the HMP and Zymomock communities, respectively (**Supp Table 1,2,3**).

332   Open-reference OTU clustering found 6783 non-reference OTUs (**Supp Table 1**) in the Extreme

333   community that summed together to make up 2.6% of the community, whereas the denoising

334   approaches all found non-reference abundances less than or equal to 0.3% (**Supp Table 4**). None

335   of the approaches were good at distinguishing the proper abundances of the two *Parabacteroides*

336   *distasonis* strains with denoising pipelines finding similar proportions of both strains and the

337   open-reference OTU pipeline finding dominance of the 13400 strain and not the 13401 strain

338   (**Fig 2b, Supp Table 4**). None of the approaches performed well at matching the expected

339   abundance of the Zymomock community or the fungal community (**Fig 2**). All three denoising

340   pipelines called over-abundances of *Lactobacillus fermentum* in the Zymomock community,

341   with Deblur calling the most (44.5%) and UNOISE3 calling the least (37.6%) (**Supp Table 3**).

342   Similarly, all approaches called non-reference hits in greater than 9% abundance in the fungal

343   community (**Supp Table 5**), with UNOISE3 calling the most non-reference hits (12.4%) and

344   Deblur calling the least (9.8%). Due to all four pipelines producing similar proportions of non-

345   reference hits in the fungal community, this could suggest that either the mock compositions are

346   not in the expected proportions, the four pipelines are similarly biased, or that an upstream

347   process during sequencing caused the introduction of unexpected sequences.

348

349   **Weighted beta-diversity results from different approaches are indistinguishable in real soil**

350   **and host-associated communities**

351   After comparing each pipeline using mock communities, we next wanted to investigate how

352   comparable the results between pipelines were for real 16S datasets. We compared the pipelines

353   on a soil dataset (soil) due to its high diversity (Fierer & Jackson, 2006) a mouse exercise stool

354   dataset (exercise) and a gut biopsy sample dataset from Crohn's Disease patients plus controls

355   (human-associated). The intra-sample distances based on weighted UniFrac measurements were

356   comparable between each approach among all three datasets (**Fig 3**). In general, we found the

357   intra-samples distances to be small, ranging between median values of ~0.09 and ~0.15. There

358   were no consistent differences seen between all three datasets. Looking at the soil dataset, we

359   found that all three denoising pipelines had similarly small intra-sample distances (~0.11) (**Fig**

360   **3A**), while the open-reference OTU clustering pipeline was slightly further away from UNOISE3

361   and Deblur (~0.13). Despite this observation, it is important to note that this difference is still

362   relatively small. Furthermore, when each sample was plotted onto a PCoA we found that samples

363   tended to group by biological origin rather than the approach used to process the sequences,

364   suggesting that a relatively similar PCoA plot would be generated by each pipeline (**Fig 3D-F**).

365   The Mantel correlations between each weighted UniFrac distance matrix were all highly

366   correlated (correlation values ranging from 0.764 to 0.975), which suggests similar weighted

367   UniFrac profiles between each approach (**Supp Table 6-8**).

368 Looking at Bray-Curtis dissimilarity, another metric that takes in abundance information,

369 we found that intra-sample distances based on genus-level assignment were similar among

370 different approaches (**Supp Fig 5**). In general, the intra-sample distances were relatively low,

371 with median values ranging from ~0.04 to ~0.20, indicating a high amount of agreement between

372 approaches. Interestingly, we found that the intra-sample distances seemed to increase across

373 datasets based on how diverse the samples within them were (**Supp Fig 5**). In two of the three

374 datasets (soil and BISCUIT), we found that DADA2 and UNOISE3 tended to be closer together

375 in distance than Deblur was to DADA2 or UNOISE3, indicating a slightly higher amount of

376 agreement between DADA2 and UNOISE3 at the genus level. Again, these differences were

377 relatively small and would have minimal impacts on biological results obtained from them.

378 Plotting the Bray-Curtis dissimilarity matrices onto an NMDS plot resulted in similar findings as

379 weighted UniFrac, with samples grouping by biological origin rather than the pipeline used to

380 process them. Furthermore, we found that the Bray-Curtis dissimilarity matrices were extremely

381 well correlated with each other, ranging in values between 0.956-0.995 (**Supp Table 6-8**).

382 We next compared unweighted UniFrac distances, which is a metric that considers the

383 presence or absence of ASVs/OTUs and their phylogenetic distance. We found that the median

384 intra-sample distance between each pipeline was much greater, ranging between ~0.40 to ~0.79

385 (**Supp Fig 6**). Similar to our results focused on Bray-Curtis dissimilarity, we also found that the

386 median intra-sample distance between samples tended to increase with sample diversity (0.72-

387 0.79 for soil; 0.40-0.55 for human-associated; 0.60-0.70 for exercise. The PCoA plot based on

388 these distances resulted in samples grouping together based on pipeline, rather than biological

389 origin, indicating large differences between the different approaches (**Supp Fig 6**).

390   To look at how much of a difference the positive filtering process had on the profiles

391   generated by Deblur and to see if the slight difference in Bray-Curtis dissimilarity among the

392   denoising approaches was caused by this filter, we ran the stand-alone version of Deblur and

393   examined the non-positive filtered results. We found that the intra-sample distances based on any

394   of the three metrics tested (weighted/unweighted UniFrac and Bray-Curtis dissimilarity) between

395   any of the pipelines did not vary (**Supp Fig 7**).

396

397   **Alpha-diversity metrics vary between denoising pipelines**

398   We next investigated how the number of ASVs/OTUs called in a real dataset differed between

399   processing pipelines. In the soil dataset, DADA2 called 16609 ASVs, UNOISE3 called 11613

400   ASVs, Deblur called 8270 ASVs and open-reference OTU clustering called 21297 OTUs (**Fig**

401   **4A**). Across all datasets, we found that DADA2 called the most ASVs among the denoising

402   pipelines and that open-reference OTU clustering called the most ASVs/OTUs overall (**Fig 4A,**

403   **4C, 4E**). On average, DADA2 called 727 more ASVs than Deblur and 532 more ASVs than

404   UNOISE3 while open-reference OTU clustering called 3135 more OTUs/ASVs than DADA2. In

405   all datasets, Deblur called the least amount of ASVs/OTUs. Looking at the number of OTUs

406   called per sample, we found that DADA2 did not correlate well with UNOISE3 or open-

407   reference OTU clustering and in the soil dataset the correlation between DADA2 and UNOISE3

408   was not found to be significant (p=0.054) (**Fig 4B,4D,4F**). This is a concerning result as it

409   indicated the possibility of different biological results based on the pipeline that was chosen to

410   process the data.

411   One interesting result was that, despite DADA2 calling the most total ASVs among the

412   denoising pipelines, it called the least amount per sample (**Supp Fig 8-10**) and, among denoising

413   pipelines, UNOISE3 tended to call the most (**Supp Fig 8-10**). However, open-reference OTU

414   clustering called the most OTUs/ASVs overall per sample among all approaches that were tested

415   with the exception of the soil dataset.

416

417   **Computational requirements are vastly different across denoising pipelines**

418   Knowing that all three of these pipelines resulted in similar relative abundance profiles on mock

419   communities and small intra-sample distances on real 16S communities, we next investigated

420   how the run time and memory usage differed between the denoising approaches. We found that

421   UNOISE3 (4.6 minutes) was 1273 times faster than DADA2 (5834 minutes) and 15 times faster

422   than Deblur (69.3 minutes) at a total read count of 1,926,000 reads evenly distributed across 103

423   samples (**Fig 5A**). Run times for all pipelines increased as the number of reads per sample

424   increased. Deblur used a static amount of memory (611 Mb) as reads per sample increased,

425   whereas in general the other two pipelines increased in memory usage as the number of reads per

426   sample increased, with the exception of DADA2 run at 1,926,000 reads (**Fig 5B**). Deblur used

427   the smallest amount of memory at the maximum read count of 1,926,000 reads. We found that

428   DADA2 had the highest amount of memory usage (4071 Mb at 1,287,000 reads) among the three

429   pipelines. Interestingly, this usage was more than the amount used at the maximum read count

430   (3600 Mb). In addition, none of the runs exceeded the 4 Gb memory cap on the 32-bit free

431   academic version of USEARCH10.

432

433   **Discussion**

434   Despite the differences between the underlying algorithms, all four approaches were

435   comparable based on weighted UniFrac and Bray-Curtis distances. However, the tools varied

436   greatly when looking at unweighted or raw number of ASVs/OTUs found. It should be noted that all

437   denoising pipelines were run using their recommended settings, limiting our comparison to only looking

438   at the default settings for each pipeline. Adjusting of parameters within each sequence analysis pipeline

439   would allow users to tailor towards more sensitivity and specificity as desired. Furthermore, each package

440   uses individual chimera filtering methods which may affect the overall accuracy of the entire pipeline.

441   During mock community data processing, no denoising approach consistently called

442   more ASVs than another, however open-reference OTU clustering found more OTUs than any of

443   the three denoising approaches before abundance filtering (**Fig 1**). In many cases, the number of

444   OTUs found was vastly greater than the number of expected sequences, which is consistent with

445   previous literature reporting that OTU clustering tends to exaggerate the number of unique

446   organisms found within a sample (Edgar, 2017). No pipeline was able to call all expected

447   sequences for any of the mock communities at 100% identity (**Fig 1**), which indicates that some

448   may not have been present or that preparation steps show large bias towards specific 16S rRNA

449   gene variants within an organism. Each sequence processing pipeline was able to detect every

450   organism in the HMP community (note *S. aureus* and *S. epidermidis* are collapsed together as

451   they have the same sequenced region) and the Zymomock community. However, one odd result

452   was that observed abundances within the HMP community were comparable to the expected

453   abundances within the HMP community, but not the Zymomock community (**Fig 2A,2D**)**. This

454   again most likely indicates bias during library preparation but could also be due to similar bias

455   among pipelines. In the Extreme dataset, all denoising pipelines missed *P. buccalis*, *C.*

456   *methylpentusum* and *P. sp._D13* (**Fig 2B, Supp Table 4**). All three of these organisms had very

457   low expected abundances (less than 0.00427%) which may explain why they were difficult to

458   detect (**Supp Table 4**). We do not believe it was an issue with sequence depth as the single

459   sample in the mock community was sequenced at a depth of 11.6 million reads and open-

460   reference OTU clustering only missed one organism (*C. methylpentusum*). Deblur and UNOISE3

461   both failed to detect 9 of the 27 expected taxa in the Extreme dataset at 97% identity, which were

462   all detected by DADA2 and open-reference OTU clustering. Again, these nine organisms were at

463   very low abundances (less than 0.05%) (**Supp Table 4**). As mentioned above, open-reference

464   OTU clustering did find 2 more expected taxa than DADA2, however, this came at the cost of

465   significantly increased numbers of non-reference sequences being detected. Differences in

466   detection between DADA2 and the other two denoising pipelines suggests that it is better at

467   detecting organisms that are very rare without the cost of finding significantly more non-

468   reference hits that plagues the open-reference OTU clustering pipeline. Whether finding rare

469   organisms is truly advantageous is debatable, as many of these low-abundance organisms would

470   be removed by typical filtering cut-offs and/or contribute little to weighted beta-diversity metrics

471   such as UniFrac or Bray-Curtis dissimilarity.

472   In all mock datasets, a large number of Unmatched OTUs were found by open-reference

473   OTU clustering and, in some of the datasets, different denoising pipelines also found a relatively

474   large amount of Unmatched ASVs (**Fig 1**). This could have been due to errors in sequencing or

475   mutations during PCR amplification of the 16S rRNA gene as most sequences did show over

476   90% identity to the expected sequences, indicating that they originated from a 16S rRNA gene

477   (**Supp Fig 1-3**). It is also possible that many of these sequences are contaminants from the

478   environment, while others may have been from chimeras that were not detected by each

479   pipelines' individual chimera-checking algorithms. We also found a high proportion of matches

480   to the reference database but not the expected sequences. There are several reasons why this

481   could have occurred. Some of these sequences may be contaminants introduced during mock

482   community preparation or sequencing as they match with 100% identity to a sequence within the

483   Greengenes database indicating that they are a real 16S rRNA gene sequence. However, due to

484   different pipelines identifying different numbers of these unexpected sequences, it could also

485   suggest that some of these contaminants are likely either chimeric sequences that can be found

486   within the Greengenes database, or sequencing errors/mutations during PCR amplification that

487   could not be corrected for by the respective pipelines. Unsurprisingly, the highest sequence depth

488   mock community (Extreme) had the most Unmatched OTUs (**Fig 1**), most likely due to the

489   increased absolute abundance of sequencing errors and the increased likelihood of finding minor

490   contaminants. Furthermore, DADA2 found a significantly greater amount of sequences under or

491   at 75% sequence identity, when compared to the other denoising pipelines, highlighting the

492   trade-off that it makes to find lower-abundance ASVs (**Supp Fig 2**).

493   In all cases, the open-reference OTU clustering had greatly increased numbers of OTUs

494   found when compared to ASVs found by denoising pipelines (**Fig 1**). To try and address this

495   issue, we applied an abundance filter of 0.1% minimum abundance (except for the Extreme

496   community 0.0004% minimum abundance) to all four different pipelines over all the datasets

497   (**Supp Fig 4**). The largest effect that this filter had was the reduction of the number of OTUs

498   called by the open-reference OTU clustering. These significant changes in the number of OTUs

499   by open-reference OTU clustering highlight the importance of removing low-abundance OTUs

500   during analysis and demonstrate the stability of denoising approaches after abundance filtering.

501   The filter cutoff did have an impact on the number of Unmatched ASVs called by DADA2 in the

502   HMP mock community (**Supp Fig 4A**) and the Unmatched ASVs called by Deblur in the

503   Zymomock community (**Supp Fig 4E**), but had little effect on the number of ASVs called by

504   UNOISE3 on these communities, indicating UNOISE3's tendency to call ASVs of higher

505   abundance. No differences were seen in the Extreme community on ASV output by denoising

506   approaches, most likely due the relatively low cutoff of 0.0004%. Overall, results on the mock

507   communities showed that Deblur tended to call the least amount of ASVs/OTUs among all

508   pipelines and open-reference OTU clustering called the most.

509   The relative abundances determined for each study on the mock communities were

510   similar to each other irrespective of which pipeline processed the data (**Fig 2**). This finding

511   suggests that biological conclusions based on microbial relative abundance data should be

512   unaffected by the choice of denoising pipeline. One trend that was noticed in the relative

513   abundance data was that UNOISE3 tended to call higher abundances of non-reference

514   ASVs/OTUs in each mock community except for the Extreme community where open-reference

515   OTU clustering found the highest abundance of non-reference hits. Interestingly, the lowest

516   identity match for any of these ASVs called by UNOISE3 in both the Zymomock and HMP

517   mock communities was still found to be above 90% identity to the expected sequences (**Supp**

518   **Fig 1,3**) and was classified as Gammaproteobacteria by the RDP classifier using a 70%

519   confidence threshold, suggesting it is a 16S rRNA sequence that may have been introduced by

520   contamination, sequencing bleed-through or acquired an error early on in PCR amplification.

521   The relative abundances determined within the Zymomock and fungal communities were

522   highly similar between pipelines, but markedly differed from the expected result. This finding

523   suggests that either the expected abundances of sequences from these communities may be

524   incorrect or all four pipelines are similarly biased. This non-agreement could also be due to steps

525   during the sequencing processes such as PCR amplification, which may be causing primer bias

526   (Aird et al., 2011) or the inclusion of contaminant organisms. In the case of the fungal

527   community, it is possible that none of these approaches work well with ITS1 data which are

528   more variable than 16S data. Additional fungal mock communities should be analyzed in the

529   future to better explore this issue.

530   Benchmarking relative abundance profiles from different pipelines with mock

531   communities can be useful, however, they tend to lack the diversity that is found in many real

532   sample datasets. To address this issue, we compared resulting microbial compositions from each

533   pipeline across three real datasets (exercise, human-associated, and soil). Weighted UniFrac,

534   unweighted UniFrac and Bray-Curtis dissimilarity distances between the same biological

535   samples for each approach were examined. In both cases the weighted UniFrac and Bray-Curtis

536   intra-sample distances between all pipelines for all three datasets were small (less than a median

537   of 0.21) (**Fig 3, Supp Figure 5**). This complemented our previous results, showing that each

538   pipeline had comparable microbial compositions for the mock communities. Furthermore,

539   plotting the samples on a PCoA or NMDS resulted in the same biological samples from each

540   pipeline grouping together (**Fig 3, Supp Figure 5**). This indicated that a similar plot would be

541   observed whether the researcher was using the Deblur, UNOISE3, DADA2 or open-reference

542   OTU clustering. Interestingly, Deblur did not agree with the DADA2 or UNOISE3 as much as

543   they agreed with each other on multiple occasions, based on Bray Curtis dissimilarity at the

544   genus level (**Supp Fig 5A,5B**). This result is interesting, as one of the main differences between

545   Deblur and the other two denoising pipelines is its positive filtering feature, and so we expected

546   this feature to be driving this difference. However, when we compared the other three

547   approaches to Deblur and Deblur without positive filtering, we found no difference (**Sup Fig 7**).

548   Due to the similar weighted UniFrac results (**Fig 3**) between the denoising pipelines, we believe

549   that this difference is most likely due to highly similar sequences being classified into slightly

550   different genera.

551   Comparing the three denoising pipelines to the open-reference OTU picking pipeline, we

552   found that Deblur agreed the least with the OTU pipeline of all the denoising pipelines (**Supp**

553   **Fig 5**). In most cases, OTUs had similar intra-sample distances between pipelines in both

554   weighted UniFrac and Bray Curtis dissimilarity as denoising pipelines had amongst each other.

555   The relatively small differences seen in the abundance-based metrics indicates that choice of

556   pipeline would have minimal impact when looking at weighted beta-diversity metrics. Although,

557   due to denoising pipelines capability of single nucleotide resolution, they may provide more

558   strain information than OTU clustering at 97% would.

559   Unweighted UniFrac beta-diversity metrics were highly variable between pipelines

560   indicating different bias between pipelines when determining low abundance sequence (**Supp**

561   **Fig6A,6B,6C**). When pipelines were plotted together on a PCoA plot we found that the samples

562   separated by approach, rather than sample, indicating that interpretation from unweighted

563   UniFrac data would most likely be impacted based on the pipeline chosen to process any set of

564   given data. We wish to highlight, however, that recent evidence has emerged that the unweighted

565   version of UniFrac analysis can give misleading results (Wong et al. 2016), therefore these

566   patterns should be interpreted with caution.

567   To follow up on the vastly different unweighted UniFrac profiles of samples, we looked

568   at the alpha-diversity between the same samples run by each pipeline and the total number of

569   ASVs/OTUs called by each pipeline in the real datasets. We found that DADA2 called the most

570   ASVs among all denoising approaches, but overall open-reference OTU clustering found the

571   most (**Fig 4**). This was not surprising, as it agreed with the analysis of the mock communities.

572   Interestingly, despite DADA2 finding the most ASVs overall for all three denoising pipelines, it

573   found the least amount of ASVs per sample (**Supp Fig 8-10**). We suspect this is due to

574   DADA2's ability to create pooled error profiles and then pick ASVs sample by sample. Overall,

575   open-reference OTU clustering found more OTUs per sample than ASVs found by any denoising

576   pipelines. This indicated that the number of different organisms found within a sample is directly

577   impacted by the choice of processing pipeline, emphasizing the difficulty in determining the true

578   number of different organisms a sample contains. It should be noted that all of the denoising

579   pipelines provide parameters that could be altered to increase or decrease sensitivity in

580   identifying rare/spurious ASVs depending on a user's targeted application. One alarming result

581   we found was the poor correlation between the number of observed OTUs/ASVs found by

582   DADA2 and both open-reference OTU clustering and UNOISE3 (**Fig 4**). This is concerning as

583   major differences in biological signal would be seen depending on the approach that was chosen

584   to process the data (i.e. a sample could have relatively low numbers of ASVs based on DADA2

585   analysis, but have relatively high numbers of ASVs/OTUs based on UNOISE3 or an OTU

586   analysis). This issue highlights that the approaches are all very good at identifying highly

587   abundant sequences but vary when identifying low-abundance sequences which will impact

588   metrics that do not take into account the abundance of OTUs/ASVs.

589   A major difference between the three denoising pipelines was their computational run

590   time. UNOISE3 was magnitudes faster than both DADA2 and Deblur. This is most likely due to

591   both the programming language that UNOISE3 is implemented in (C++), as well as its simple

592   one-pass denoising pipeline. DADA2 was the slowest pipeline and, although computation time

593   could be inconvenient for those with limited computational power, it did not reach times that

594   were impractical even when running almost 2 million total reads. Memory usage for each

595   program also did not reach impractical amounts when running close to 2 million reads, with

596   DADA2 using a maximum amount of 1024 Mb of memory which is a reasonable amount for

597   modern computers. Memory usage by UNOISE3 did not come close to reaching the 4 Gb

598   memory cap on the 32-bit version, even after running 103 samples at a total read depth of 2

599   million reads. This suggests that for most moderately sized 16S datasets the 32bit version of

600   UNOISE3 should be sufficient.

601

602   **Conclusion**

603   In conclusion, all four pipelines are comparable when looking at weighted results that are

604   based on the relative abundances of ASVs/OTUs while using default settings. However, the

605   approaches do vary when looking at the number of ASVs/OTUs found and unweighted metrics

606   such as unweighted UniFrac, while using the default settings for each denoising pipeline. The

607   number of ASVs called did not differ between denoising pipelines in a consistent way across

608   mock communities, suggesting that determining species richness within low-diverse samples

609   could be problematic. However, we did find that open-reference OTU clustering consistently

610   called more OTUs than ASV-calling pipelines. Analysis of the real datasets showed that DADA2

611   consistently called more ASVs than the other two denoising pipelines and that, again, open-

612   reference OTU clustering called the most overall. More importantly, in the soil dataset and in the

613   Extreme dataset DADA2 and the open-reference OTU clustering pipeline were capable of

614   finding more low-abundance organisms, but DADA2 could do this without the cost of

615   significantly increasing the number of non-reference hits. The most alarming result was the poor

616   correlation in the number of ASVs/OTUs per sample found between DADA2 and UNOISE3 or

617   open-reference OTU clustering. From this, we believe that choice of approach will have large

618   impacts on determining the alpha-diversity between different samples. Looking at computational

619   run time, we found that DADA2 was by far the slowest denoising pipeline, whereas UNOISE3

620   was the fastest, processing datasets more than 1200 times faster than DADA2. In the end, the

621   choice of approach did not play a large role in weighted analyses based on microbial abundances,

622   but did have implications on unweighted results and alpha-diversity metrics. We believe this is

623   promising, as it indicates that no matter the choice of approach, a similar weighted biological

624   signal will be seen. On the other hand, extreme caution is required when looking at unweighted

625   results and alpha-diversity metrics between different pipelines.

626

627   **Acknowledgements**

628

629   We would like to thank members of the Langille lab for providing feedback and suggestions for

630   additional analysis during lab meetings. We would especially like to thank Karl Leuschen for

631   providing previous data on a comparison between QIIME OTU clustering and DADA2

632   denoising.

633

634

635

636   **References**

637

638   Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B.,

639   Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in

640   Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. https://doi.org/10.1186/gb-

641   2011-12-2-r18

642   Allali, I., Arnold, J. W., Roach, J., Cadenas, M. B., Butz, N., Hassan, H. M., Koci, M., Ballou,

643   A., Mendoza, M., Ali, R., & Azcarate-Peril, M. A. (2017). A comparison of sequencing

644 platforms and bioinformatics pipelines for compositional analysis of the gut microbiome.

645 *BMC Microbiology*, *17*(1), 194. https://doi.org/10.1186/s12866-017-1101-8

646 Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local

647 alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

648 https://doi.org/10.1016/S0022-2836(05)80360-2

649 Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z.,

650 Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur

651 Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems*, *2*(2).

652 https://doi.org/10.1128/mSystems.00191-16

653 Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., Dutton,

654 R. J., Turnbaugh, P. J., Knight, R., & Caporaso, J. G. (2016). mockrobiota: A Public

655 Resource for Microbiome Bioinformatics Benchmarking. *mSystems*, *1*(5).

656 https://doi.org/10.1128/mSystems.00062-16

657 Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace

658 operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–

659 2643. https://doi.org/10.1038/ismej.2017.119

660 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P.

661 (2016). DADA2: High resolution sample inference from Illumina amplicon data. *Nature*

662 *Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

663 Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R.

664 (2010a). PyNAST: A flexible tool for aligning sequences to a template alignment.

665 *Bioinformatics*, *26*(2), 266–267. https://doi.org/10.1093/bioinformatics/btp636

666   Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K.,

667   Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights,

668   D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M.,

669   Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T.,

670   Zaneveld, J., & Knight, R. (2010b). QIIME allows analysis of high-throughput community

671   sequencing data. *Nature Methods*, *7*, 335. http://dx.doi.org/10.1038/nmeth.f.303

672   Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-

673   Alfaro, A., Kuske, C. R., & Tiedje, J. M. (2014). Ribosomal Database Project: Data and

674   tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(Database issue),

675   D633–D642. https://doi.org/10.1093/nar/gkt1244

676   Comeau, A. M., Douglas, G. M., & Langille, M. G. I. (2017). Microbiome Helper: A Custom

677   and Streamlined Workflow for Microbiome Research. *mSystems*, *2*(1).

678   https://doi.org/10.1128/mSystems.00127-16

679   DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T.,

680   Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA

681   gene database and workbench compatible with ARB. *Applied and Environmental*

682   *Microbiology*, *72*(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05

683   Douglas, G. M., Hansen, R., Jones, C. M. A., Dunn, K. A., Comeau, A. M., Bielawski, J. P.,

684   Tayler, R., El-Omar, E. M., Russell, R. K., Hold, G. L., Langille, M. G. I., & Van

685   Limbergen, J. (2018). Multi-omics differentially classify disease state and treatment

686   outcome in pediatric Crohn's disease. *Microbiome*, *6*(1), 13.

687   https://doi.org/10.1186/s40168-018-0398-3

688   Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST.

689   *Bioinformatics*, *26*(19), 2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461

690   Edgar, R. C. (2016). UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon

691   sequencing. *bioRxiv*. http://biorxiv.org/content/early/2016/10/15/081257.abstract

692   Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-

693   reference OTUs. *PeerJ*, *5*, e3889. https://doi.org/10.7717/peerj.3889

694   Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves

695   sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200.

696   https://doi.org/10.1093/bioinformatics/btr381

697   Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial

698   communities. *Proceedings of the National Academy of Sciences of the United States of*

699   *America*, *103*(3), 626–631. https://doi.org/10.1073/pnas.0507535103

700   Illumina. (n.d.). Effects of Index Misassignment on Multiplexing and Downstream Analysis.

701   https://www.illumina.com/content/dam/illuminamarketing/documents/products/whitepapers

702   /index-hopping-white-paper-770-2017-004.pdf

703   Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., Bates, S.

704   T., Bruns, T. D., Bengtsson-Palme, J., Callaghan, T. M., Douglas, B., Drenkhan, T.,

705   Eberhardt, U., Dueñas, M., Grebenc, T., Griffith, G. W., Hartmann, M., Kirk, P. M.,

706   Kohout, P., Larsson, E., Lindahl, B. D., Lücking, R., Martín, M. P., Matheny, P. B.,

707   Nguyen, N. H., Niskanen, T., Oja, J., Peay, K. G., Peintner, U., Peterson, M., Põldmaa, K.,

708   Saag, L., Saar, I., Schüßler, A., Scott, J. A., Senés, C., Smith, M. E., Suija, A., Taylor, D.

709   L., Telleria, M. T., Weiss, M., & Larsson, K.-H. (2013). Towards a unified paradigm for

710   sequence-based identification of fungi. *Molecular Ecology*, *22*(21), 5271–5277.

711   https://doi.org/10.1111/mec.12481

712   Lamoureux, E. V, Grandy, S. A., & Langille, M. G. I. (2017). Moderate Exercise Has Limited

713   but Distinguishable Effects on the Mouse Microbiome. *mSystems*, *2*(4), e00006-17.

714   https://doi.org/10.1128/mSystems.00006-17

715   Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.

716   *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*.

717   http://dx.doi.org/10.14806/ej.17.1.200

718   MIRARAB, S., NGUYEN, N., & WARNOW, T. (2011). SEPP: SATé-Enabled phylogenetic

719   placement. In *Biocomputing 2012* (pp. 247–258). WORLD SCIENTIFIC.

720   https://doi.org/doi:10.1142/9789814366496_0024

721   Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H.

722   (2018). vegan: Community ecology package. Retrieved from https://cran.r-

723   project.org/package=vegan

724   Plummer, E., & Twin, J. (2015). A Comparison of Three Bioinformatics Pipelines for the

725   Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. *Journal of*

726   *Proteomics & Bioinformatics*, *8*(12). https://doi.org/doi: 10.4172/jpb.1000381

727   Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O.

728   (2007). SILVA: A comprehensive online resource for quality checked and aligned

729   ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, *35*(21),

730   7188–7196. https://doi.org/10.1093/nar/gkm864

731   R Development Core Team. (2008). R: A Language and Environment for Statistical Computing.

732   Vienna, Austria. http://www.r-project.org

733   Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open

734   source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

735   Tange, O. (2011). GNU Parallel: the command-line power tool. *The USENIX Magazine*, *36*(1),

736   42–47. https://doi.org/10.5281/zenodo.16303

737   Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

738   http://ggplot2.org

739   Wong, R.G., Wu, J.R., & Gloor, G.B. (2016). Expanding the UniFrac Toolbox. *PLoS ONE, 11*,

740   e0161196.

741   Yurgel, S. N., Douglas, G. M., Comeau, E. M., Mammoliti, M., Dusault, A., Percival, D., &

742   Langille, M. G. I. (2017). Variation in Bacterial and Eukaryotic Communities Associated

743   with Natural and Managed Wild Blueberry Habitats. *Phytobiomes*, *1*, 102–113.

744   https://doi.org/10.1094/PBIOMES-03-17-0012-R

745   Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina

746   Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620.

747   https://doi.org/10.1093/bioinformatics/btt593

748

749

750

751

752

753
754

755

Qualitative Comparison of DADA2, Deblur, and UNOISE3

1   **Table 1:**

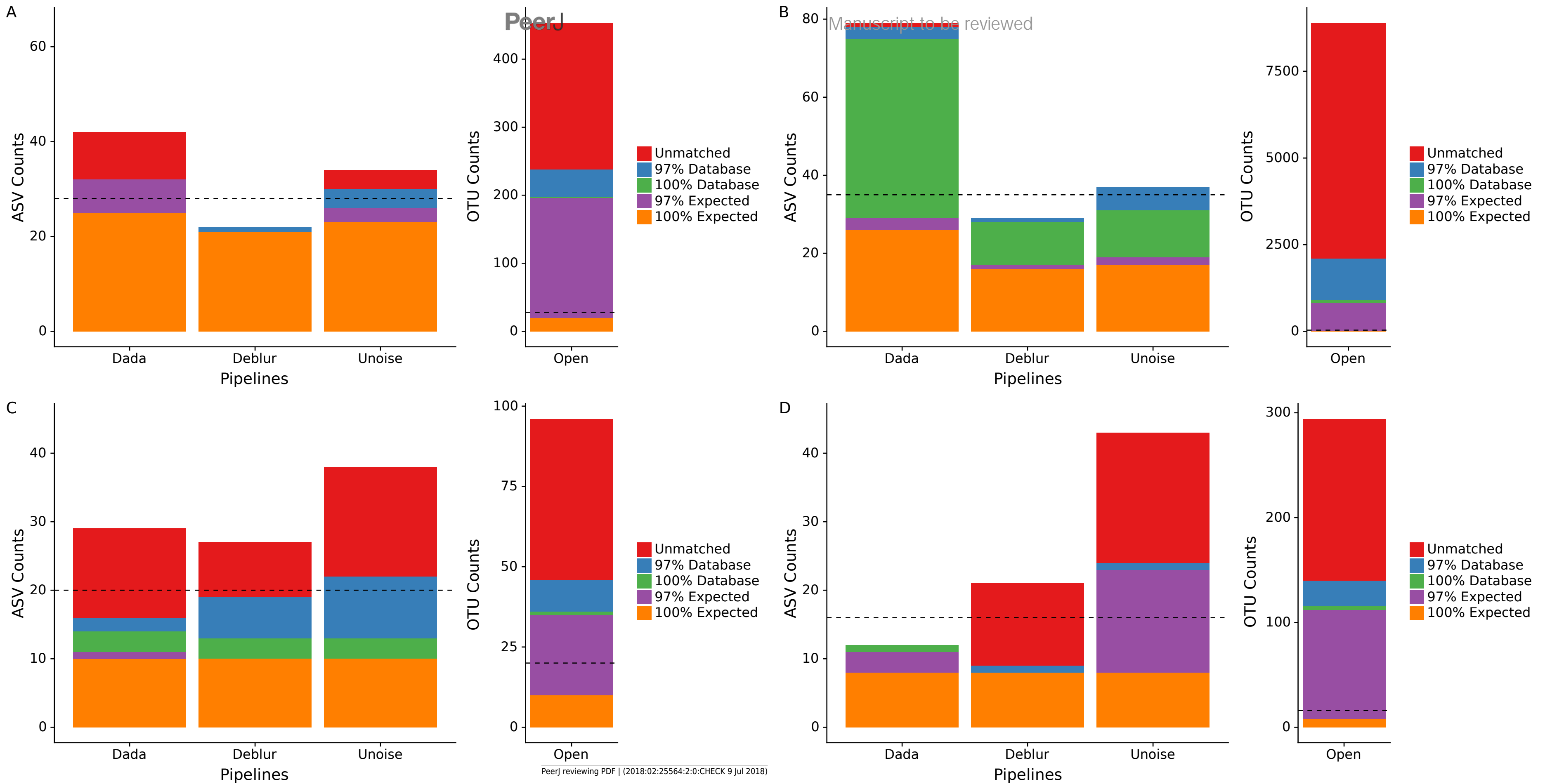
2   **Qualitative Comparison of DADA2, Deblur, and UNOISE3**

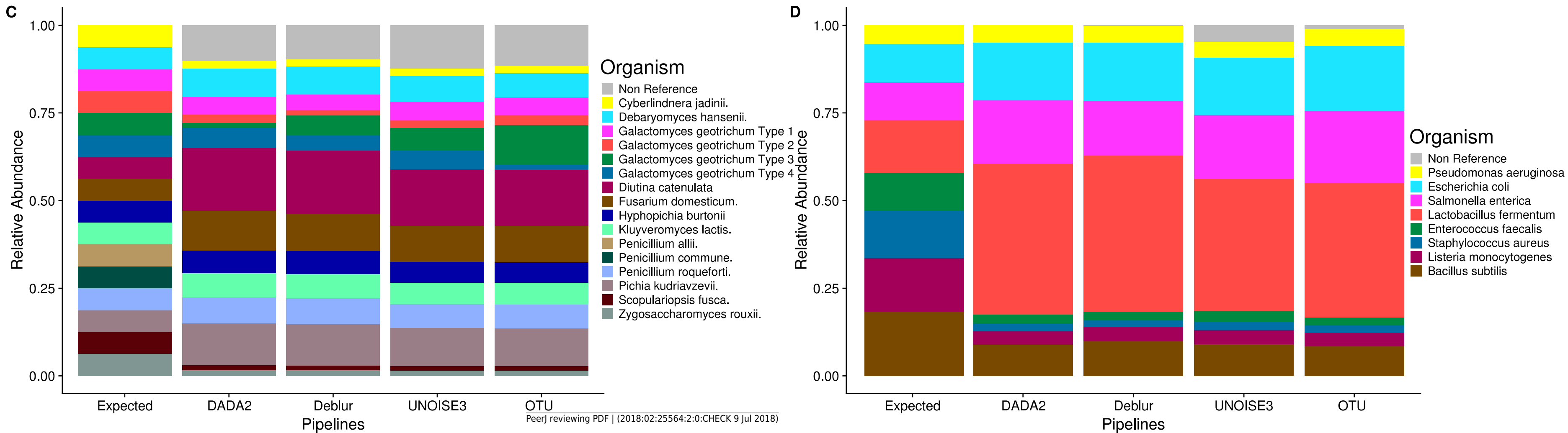
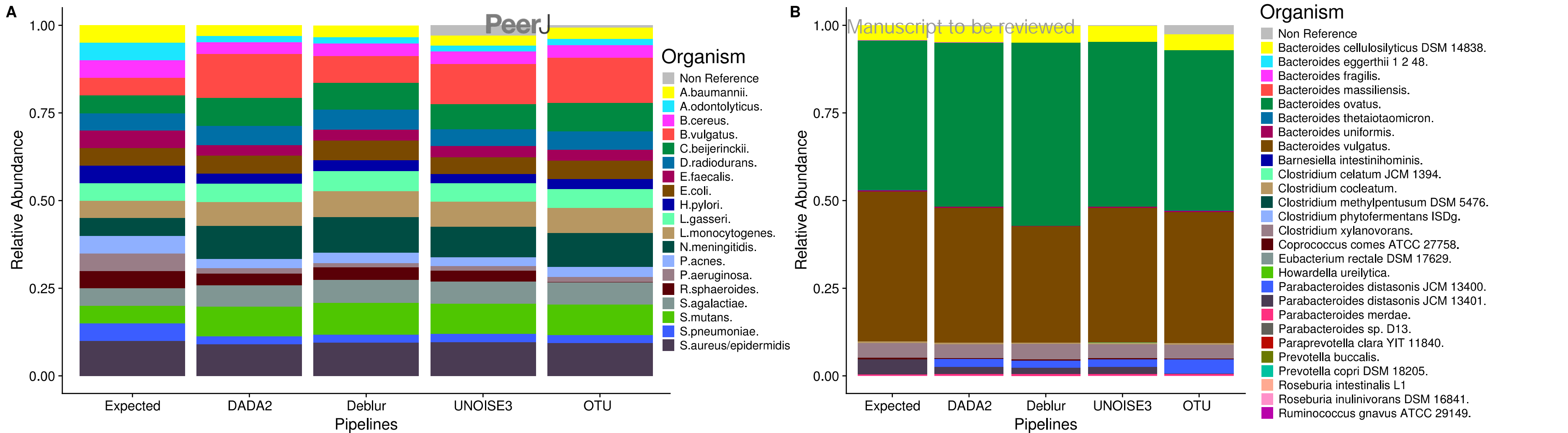
Pipeline	Implemented In	Open Source	*Pooled Sampling	**Positive Filtering	Version Tested	GUI via QIIME2	Publication Date
DADA2	R	Yes	Yes	No	1.6	Yes	April 13 2016
Deblur	Python	Yes	No	Yes	1.0.2	Yes	March 7, 2017
UNOISE3	C++	No	Yes	No	3	No	Oct 15, 2016

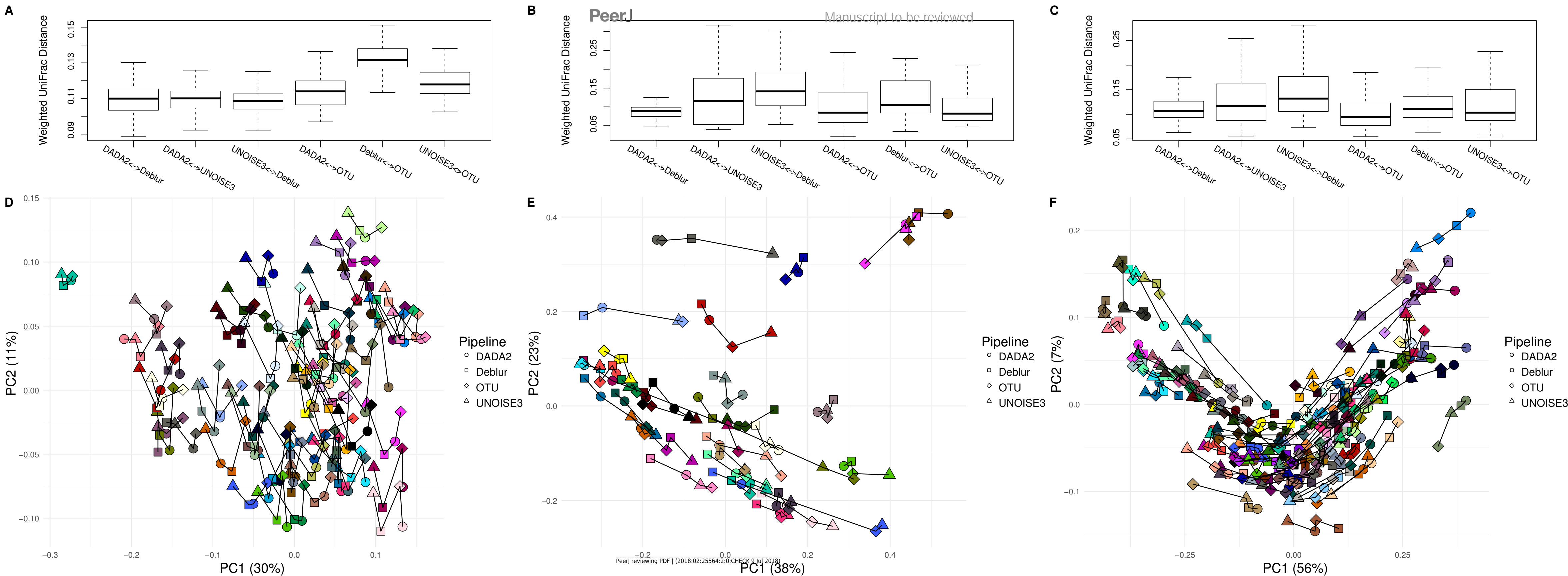
6   * When all sequences from all samples are denoised at the same time (in contrast to running each
7   sample separately).

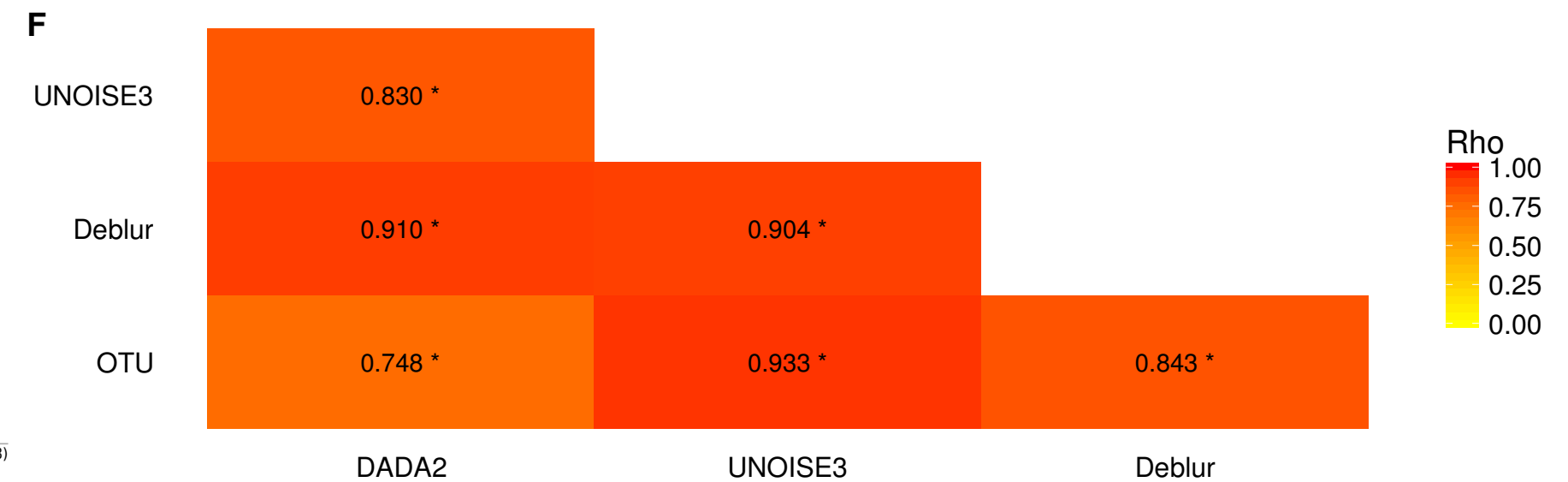
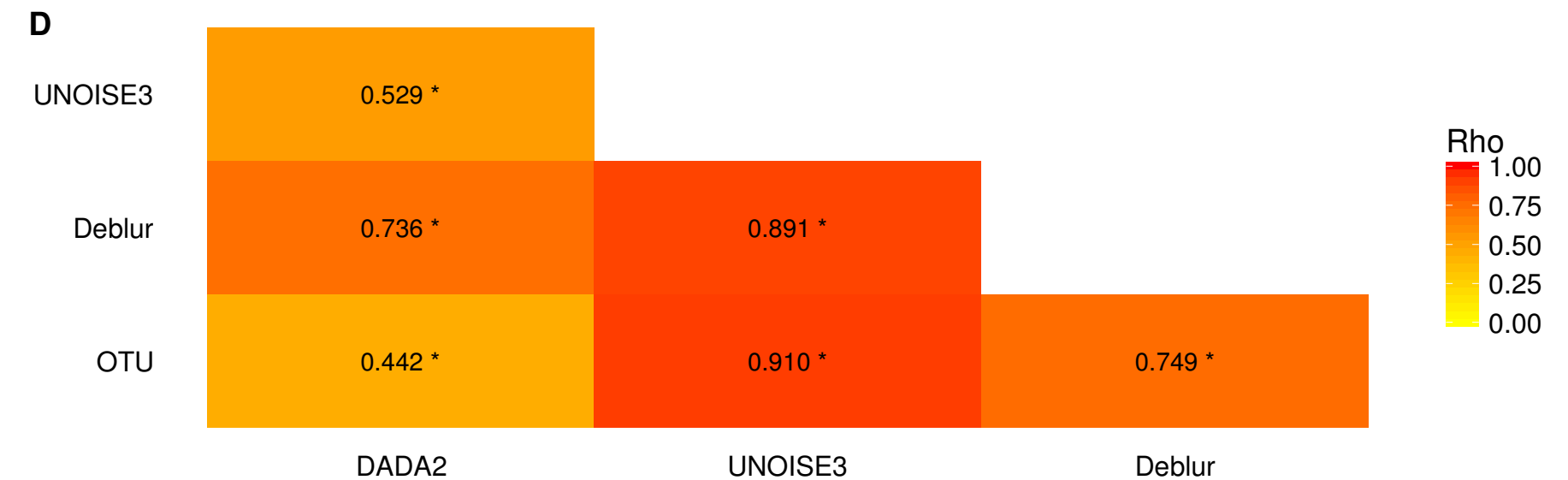
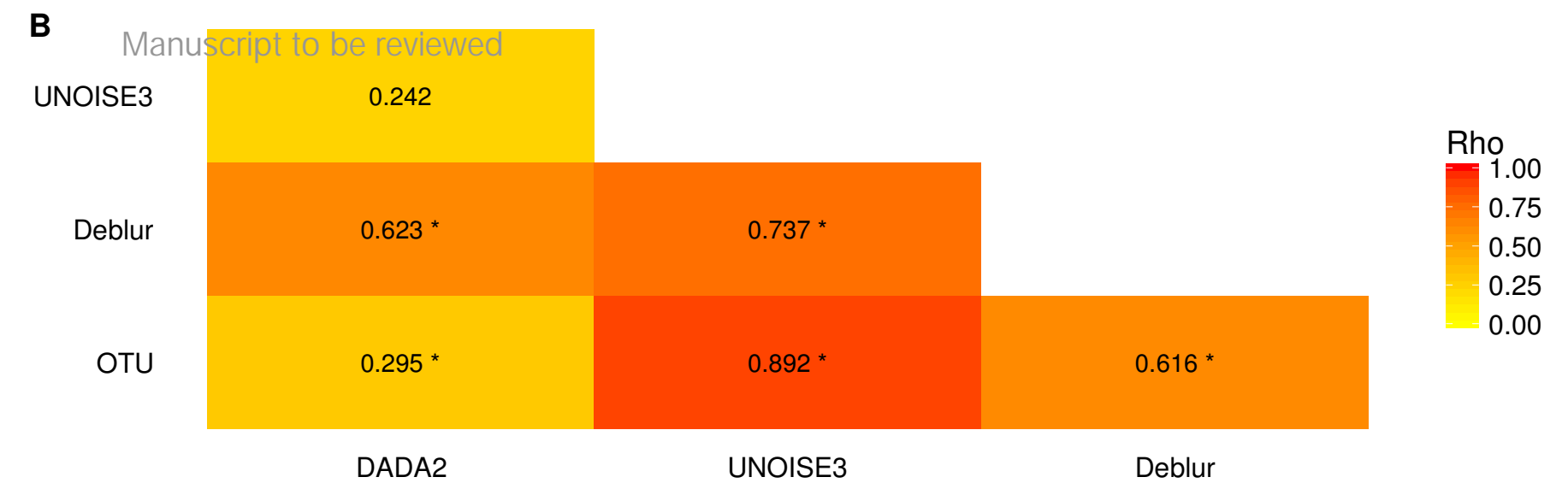
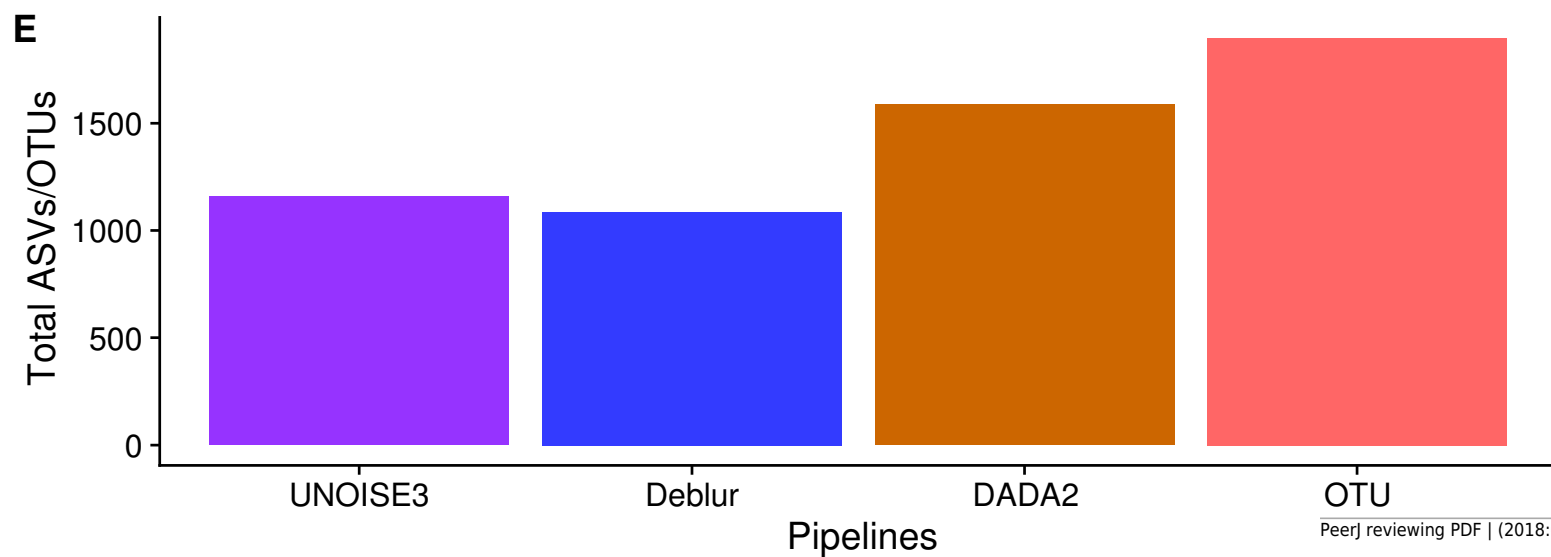
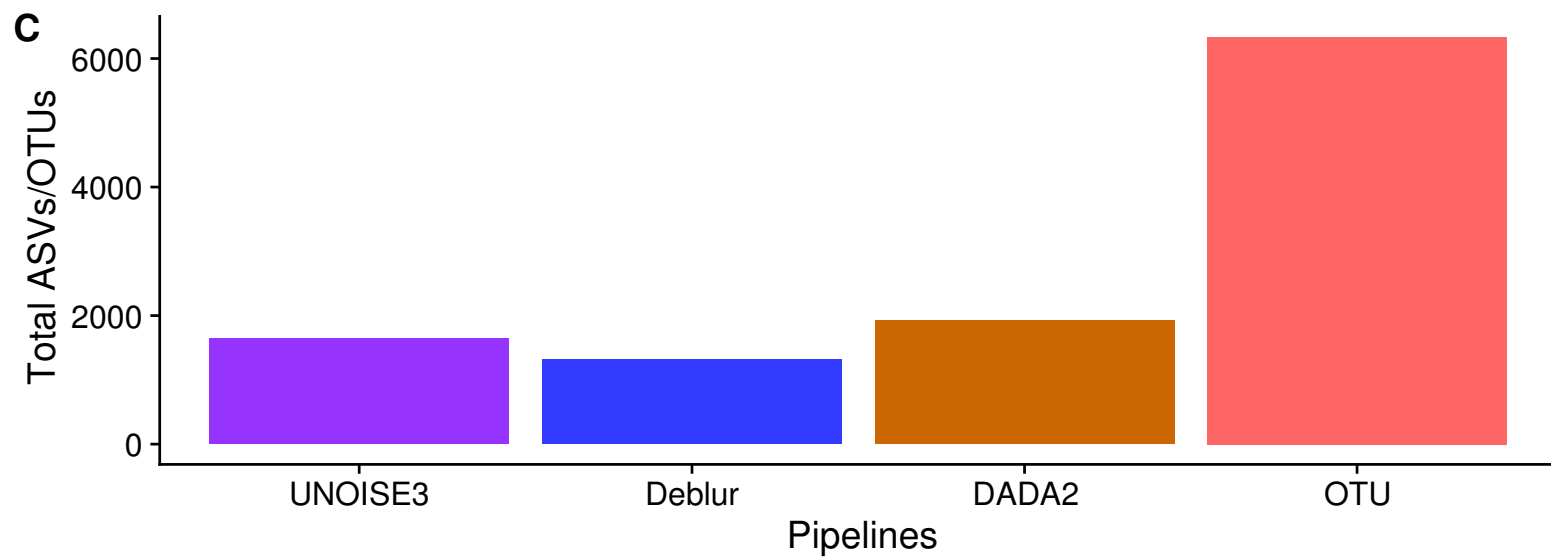
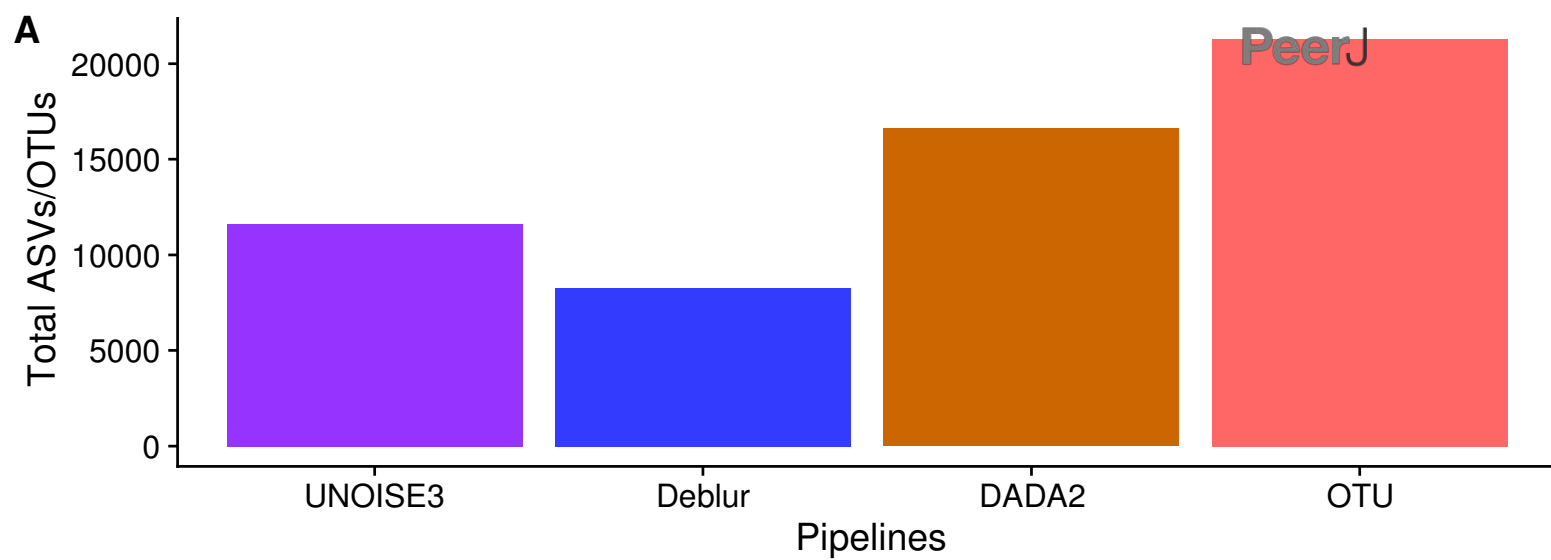
8   ** Compares resulting ASVs to a database (Greengenes for Deblur) and discards reads if they do
9   not match a certain identity threshold (88% for Deblur).

Total number of ASVs/OTUs identified by each sequence processing method for four different mock communities.

Amplicon sequence variants/Operational taxonomic units (ASVs/OTUs) were compared to a database of full-length amplicon sequences for just the microbes supposedly in the community ("Expected") and against the full SILVA or ITS databases ("Database") using BLASTN at 97% and 100% identity cut-offs. "Unmatched" sequences did not match an expected sequence or the SILVA/ITS databases at 97% identity or greater. Dotted lines indicate the total number of ASVs/OTUs expected, accounting for 16S copy variation within genomes. Note that the y-axis for open-reference OTU clustering is different than the y-axis on the denoising methods. A) Human Microbiome Project equal abundance mock community; B) Extreme dataset; C) Fungal ITS1 mock community; D) Zymomock community.

Relative abundances of taxa generated by each sequence processing method for four different mock communities.

All ASVs/OTUs that matched with expected sequences at 97% or greater identity were assigned taxonomy using a BLASTN search against the expected sequences provided for the Extreme, Human Microbiome Project, and Zymomock mock communities. All ASVs/OTUs that matched an expected species with 97% or greater identity to the UNITE database were classified as expected sequences in the fungal mock community. Non-reference refers to the abundance of ASVs/OTUs that did not match expected sequences with 97% or greater identity. A) Human Microbiome Project equal abundance mock community; B) Extreme dataset – it is important to note that some organisms are not displayed in this figure due to their very low abundances; C) Fungal ITS1 mock community; D) Zymomock community.

Weighted UniFrac intra-sample distances between sequence processing methods based on three real datasets.

A-C) The weighted UniFrac distances between the same biological samples based on ASVs/OTUs outputted by each of the different sequence processing methods on the soil, human associated and Exercise datasets, respectively. D-F) Principal coordinates analysis of the weighted UniFrac distances of all the samples in the soil, human associated, and Exercise datasets, respectively. The four different sample profiles generated for each biological sample are colour-coded and are joined by an interconnecting line.

Total number of ASVs/OTUs called by each processing method and the per-sample observed ASV/OTUs correlation between each sequence processing method.

A,C,E) The total numbers of ASVs/OTUs determined by each method on the soil, Exercise, and human associated datasets, respectively. B,D,F) Heatmaps of the Spearman correlations between the numbers of observed ASVs/OTUs per sample between different sequence processing methods. Significant p-values ($p < 0.05$) are indicated by *.

Run time and memory usage of each denoising pipeline on a dataset of varying size

The time in seconds A) and memory in megabytes B) to run varying amounts of reads through the three different denoising methods. Note time is on a $\log_{10}$ scale.