

A comparison of bioinformatics pipelines for compositional analysis of the human gut microbiome

Joanna Szopinska-Tokov^{1,2}, Mirjam Bloemendaal^{1,2}, Jos Boekhorst^{3,4}, Gerben DA Hermes⁵, Thomas HA Ederveen⁶, Priscilla Vlaming¹, Jan K Buitelaar⁷, Barbara Franke^{1,2}, Alejandro Arias-Vasquez^{1,2,*}

5 ¹ Department of Psychiatry, Radboud University Medical Center, Donders Institute for Brain,
6 Cognition and Behaviour, Nijmegen, The Netherlands

7 ² Department of Human Genetics, Radboud University Medical Center, Donders Institute for Brain,
8 Cognition and Behaviour, Nijmegen, The Netherlands

9 ³ NIZO Food Research BV, Ede, The Netherlands

⁴ Host-Microbe Interactomics Group, Department of Animal Sciences, Wageningen University & Research, Wageningen, The Netherlands

12 ⁵ Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands

13 ⁶ Center for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences,
14 Radboud University Medical Center, Nijmegen, The Netherlands

15 ⁷Department of Cognitive Neuroscience, Radboud University Medical Center, Donders Institute for
16 Brain, Cognition and Behaviour & Karakter Child and Adolescent Psychiatry University Centre,
17 Nijmegen, The Netherlands

18 *Correspondence:

19 Alejandro Arias Vasquez

20 Email: Alejandro.AriasVasquez@radboudumc.nl (AAV)

21

22 **Abstract**

23 Investigating the impact of gut microbiome on human health is a rapidly growing area of research. A
24 significant limiting factor in the progress in this field is the lack of consistency between study results,
25 which hampers the correct biological interpretation of findings. One of the reasons is variation of the
26 applied bioinformatics analysis pipelines. This study aimed to compare five frequently used
27 bioinformatics pipelines (NG-Tax 1.0, NG-Tax 2.0, QIIME, QIIME2 and mothur) for the analysis of 16S
28 rRNA marker gene sequencing data and determine whether and how the analytical methods affect the
29 downstream statistical analysis results. Based on publicly available case-control analysis of ADHD and
30 two mock communities, we show that the choice of bioinformatic pipeline does not only impact the
31 analysis of 16S rRNA gene sequencing data but consequently also the downstream association results.
32 The differences were observed in obtained number of ASVs/OTUs (range: 1,958 - 20,140), number of
33 unclassified ASVs/OTUs (range: 210 - 8,092) or number of genera (range: 176 - 343). Also, the case
34 versus control comparison resulted in different results across the pipelines. Based on our results we
35 recommend: i) QIIME1 and mothur when interested in rare and/or low-abundant taxa, ii) NG-Tax1 or
36 NG-Tax2 when favouring stringent artefact filtering, iii) QIIME2 for a balance between two
37 abovementioned points, and iv) to use at least two pipelines to assess robustness of the results. This
38 work illustrates the strengths and limitations of frequently used microbial bioinformatics pipelines in
39 the context of biological conclusions of case-control comparisons. With this, we hope to contribute to
40 “best practice” approaches for microbiome analysis, promoting methodological consistency and
41 replication of microbial findings.

42 **Keywords:** bioinformatics, 16S rRNA gene, microbiome, mothur, QIIME, NG-Tax, comparison

43 **Author Summary**

44 Studies increasingly demonstrate the relevance of gut microbiota in understanding human health and
45 disease. However, the lack of consistency between study results is a significant limiting factor of
46 progress in this field. The reasons for this include variation in study design, sample size, bacterial DNA
47 extraction and sequencing method, bioinformatics analysis pipeline and statistical analysis
48 methodology. This paper focuses on the variation generated by bioinformatics pipelines. A choice of a
49 bioinformatic pipeline can influence the assessment of microbial diversity. However, it is unclear to
50 what extent and how the results and conclusion of a case-control study can be influenced. Therefore,
51 we compared the results of a case-control study across different pipelines (applying default settings)
52 while using the same dataset. Our results indicate a lack of consistency across the pipelines. We show
53 that the choice of bioinformatic pipeline not only affects the analysis results of 16S rRNA gene
54 sequencing data from the gut microbiome, but also the associated conclusions for the case-control
55 study. This means different conclusions would be drawn from the same data analysed with different
56 bioinformatic pipeline.

57 **1. Introduction**

58 Investigations of the role of the human gut microbiota have attracted much attention in the last 15
59 years [1]. Specifically, results of studies of the 16S rRNA marker gene (16S) have been crucial in
60 understanding the role the gut microbiota play in multiple common diseases, such as irritable bowel
61 syndrome [2], autism [3], depression [4] or attention deficit hyperactivity disorder (ADHD) [5].
62 Although a few papers suggested best practice for microbiome analysis [6, 7], still there is a broad
63 choice in microbiome methods. This affects the consistency across the studies. So far, 16S rRNA gene
64 sequencing is one of the most commonly used method to study bacterial phylogeny and genus/species
65 classification [8]. 16S rRNA gene sequencing is used as a tool to identify multiple bacterial taxa and
66 assist with differentiating between closely related bacteria.

67 The classification of microbial taxonomy using the 16S rRNA gene is influenced by several factors,
68 ranging from study design, sample size, the choice of variable region of 16S rRNA gene to sequence
69 [9], collection and storage procedure, wet lab approaches, such as DNA extraction [10], sequencing
70 technique and bioinformatic pipeline settings, such as frequency filters, and the taxonomic
71 classification database [11]. Bioinformatics pipelines differ in approaches, such as quality control and
72 filtering of the sequenced data (i.e., chimera detection, filtering sequences, denoising), Operational
73 Taxonomic Units (OTUs) clustering algorithms or Amplicon Sequence Variant (ASV), and statistical
74 analysis (when a statistical analysis step is included in the pipeline). All these choices may result in
75 differences in the (observed) distribution of taxonomic groups, significantly affecting the putative
76 relationships between the gut microbiota and disease outcomes. This limits the precision of biological
77 and statistical conclusions, resulting in a lack of consistency between studies [5, 8, 9, 12].

78 In this paper, we focused on comparing bioinformatics pipelines, as their contribution to biological
79 conclusions of microbiome studies is not sufficiently quantified. So far, studies investigating
80 differences between bioinformatics pipelines have focused on general characteristics of the
81 OTUs/ASVs/reads, such as richness, diversity and microbial compositional profiles, rather than on the

82 biological conclusions that could be drawn from analyzing these characteristics [6, 13, 14]. Recently,
83 Ducarmon et al. (2020) showed that the NG-Tax 1.0 [15] and QIIME2 [16] bioinformatics pipelines
84 performed equally well in terms of microbial diversity and compositional profiles for 24 samples across
85 eight types of biological specimens from human niches [13]. Poncheewin et al. (2020) compared NG-
86 Tax 2.0 with QIIME2 using 14 mock community samples [17]. Precision of NG-Tax 2.0 (0.95) was
87 significantly higher compared to QIIME2 (0.58). Prodan et al. (2020) used a large dataset of 2,170
88 samples and one mock community of 16S rRNA data to compare QIIME-uclust [18], mothur [19],
89 USEARCH-UPARSE [20], DADA2 [21], QIIME2-Deblur [16, 22] and USEARCH-UNOISE3 [23] pipelines,
90 and concluded that "*DADA2 is the best choice for studies requiring the highest possible biological*
91 *resolution (e.g. studies focused on differentiating closely related strains)*" [6]. López-García et al. (2018)
92 showed that when the SILVA reference database was used in combination with QIIME [24] or mothur
93 [19] pipelines, richness and composition of 18 samples were highly similar [14]. However, beta-
94 diversity clustered by pipelines, which they attributed to differences in less abundant bacteria. While
95 this was not tested by López-García et al., this description hints at the possibility of different biological
96 conclusions depending on a choice of pipeline. Only one study, Allali et al. (2017), investigated whether
97 the same biological conclusion was reached when using different pipelines based on 14 chicken cecum
98 16S rRNA samples across three different treatment groups. They tested different settings of QIIME1,
99 UPARSE and DADA2 and concluded that, despite differences in diversity and abundance, they could
100 discriminate samples by treatment, leading to similar biological conclusions [25]. This conclusion was
101 limited to beta-diversity (global microbiome community differences), not including a comparison of
102 individual genera. As they reported differences in relative abundances of specific genera between
103 pipelines, their data suggests that different pipelines could result in different lists of genera being
104 significantly associated with a treatment.

105 While the existing comparisons have been essential for the field, they fall short in contributing highly-
106 needed conclusions on how the choice of bioinformatic pipeline affects downstream statistical
107 comparisons of microbial composition of groups (for example, humans with and without a disease).

108 Such comparisons are also vital for the growth and stability of the field [12]. Moreover, frequently used
109 pipelines, NG-Tax1, NG-Tax2, QIIME1, QIIME2 and mothur, have not yet been compared using the
110 same dataset. Based on these gaps and limitations in the state of the art of the field, we aimed to
111 determine the differences in taxonomic output between these five pipelines and how such differences
112 affect downstream statistical analyses and interpretation of the observed results. We used the V4 16S
113 rRNA gene sequencing data of a human case-control study of attention-deficit/hyperactivity disorder
114 (ADHD) as well as two mock communities. We would like to highlight that our aim is not to draw
115 biological conclusions from these analyses (for this we refer to [26]), but rather highlight differences
116 brought in by the choice of bioinformatic pipeline.

117 **2. Materials and Methods**

118 **2.1. Dataset**

119 The material and methods and the results sections are divided into two parts: (i) results based on
120 clinical samples (NeuroIMAGE dataset [26]) and (ii) results based on mock communities (MC), which
121 allow us to better interpret the results based on the clinical samples.

122 **2.1.1. NeuroIMAGE dataset**

123 We used the clinical and microbial information from a group of samples belonging to a case-control
124 sample (case, n=42; control, n=50) reported in the NeuroIMAGE study [26]. For an exhaustive
125 description of the sample, inclusion criteria, ADHD analysis methods, diagnostic procedures, and study
126 design used in this study, see Szopinska-Tokov et al., 2021 [26], of which this study is an extension.

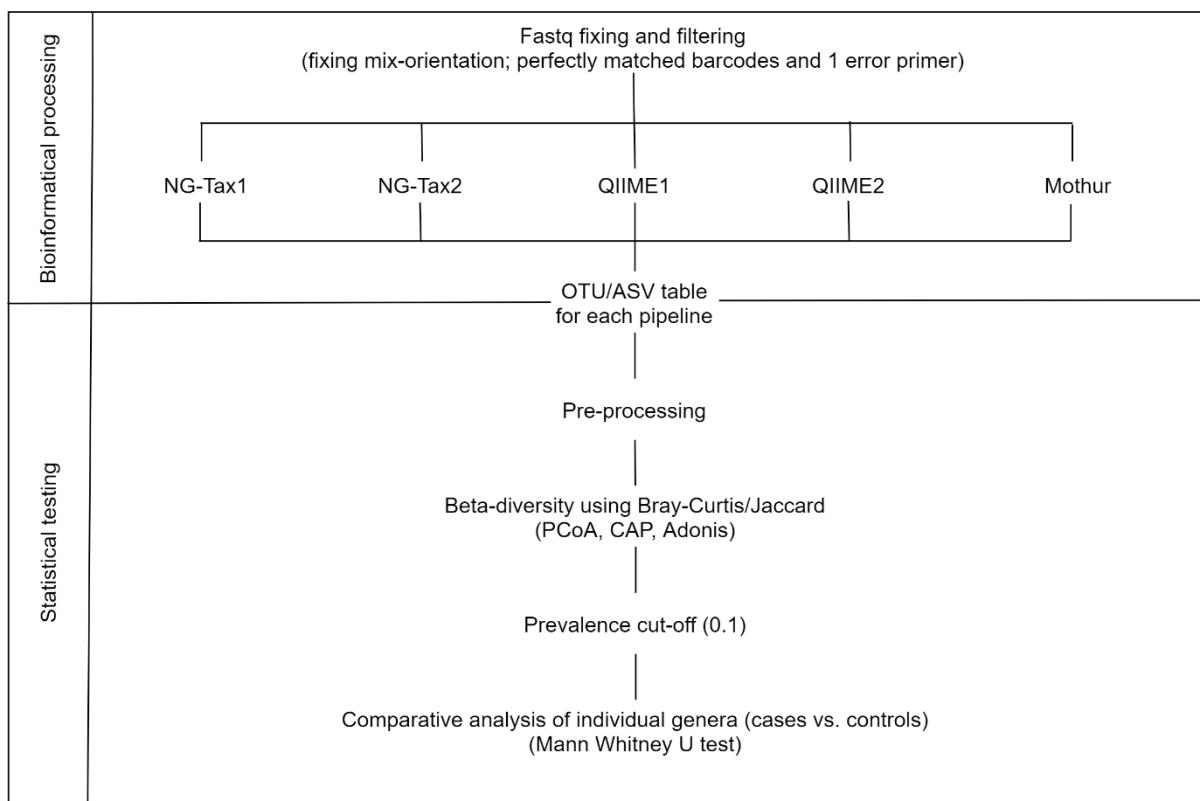
127 **2.1.2. Mock communities**

128 In addition to the case-control dataset, we analyzed eight samples based on two defined Mock
129 Communities (MCs; MC3, n=4; MC4, n=4), of which one (MC4) included taxa with very low abundances
130 (0.1%, 0.01% and 0.001%). Both MCs included the same 36 genera, but in different distributions. The
131 laboratory processing and evaluation of the observed MC composition was done exactly the same as

132 for the clinical samples [26]. The laboratory processing and evaluation of the expected microbial
133 communities' composition was carried out as described previously [15]. In short, the bacteria were
134 grown as pure cultures and their DNA was then mixed in specific amounts for each community (the
135 process was carried at the Laboratory of Microbiology, Wageningen University, The Netherlands). The
136 bacterial composition of the MCs was determined with HiSeq2000, and for each bacterium used in the
137 MCs, the full length 16S gene was sequenced with Sanger sequencing to confirm their identity.

138 **2.2. Bioinformatics pipelines and their evaluation**

139 We investigated five different pipelines: both versions of the NG-Tax pipeline (NG-Tax v.1.0 [15] and
140 v.2.0 [17], here named NG-Tax1 and NG-Tax2), adapted QIIME (v.1.8.0; here called QIIME1) [18],
141 QIIME2-DADA2 (v.2019.7.0; here called QIIME2) [16], and mothur (v.1.43.0) [19]. NG-Tax1, NG-Tax2
142 and QIIME2 are ASV-based methods, whereas QIIME1 and mothur are OTU-based methods.
143 The bioinformatic pipeline evaluation involved two steps: (i) bioinformatical processing and (ii)
144 statistical testing, involving data analysis and quantification (Figure 1).



164 input for all the pipelines. Furthermore, we used the default setting of the pipelines, except for
165 taxonomic database where we used SILVA (v.132) database for all pipelines, changing the default
166 option for NG-Tax1 and QIIME1. We used the Galaxy platform to run NG-Tax1 and NG-Tax2
167 (<http://wurssb.gitlab.io/ngtax/galaxy.html>). QIIME1 was run according to the in-house (NIZO, Ede, The
168 Netherlands) protocol as described previously [10]. For QIIME2, we followed the “Moving Pictures”
169 tutorial (<https://docs.qiime2.org/2019.4/tutorials/moving-pictures/>), and for mothur the “MiSeq SOP”
170 (https://mothur.org/wiki/miseq_sop/).

171 **2.2.2. Statistical testing**

172 **2.2.2.1. NeuroIMAGE dataset**

173 **2.2.2.1.1. Pre-processing**

174 Taxonomical names were formatted across the pipelines, e.g., D_0_Bacteria was changed into Bacteria
175 in order to align the format of taxonomic names across the pipelines. The original sample contained a
176 subthreshold-ADHD group [26], which was removed in the current analysis. Furthermore, we
177 determined a threshold of total read counts based on rarefaction plots (data not shown), in order to
178 exclude samples with small number of total reads while keeping the maximum number of samples (as
179 explained in the ‘Moving pictures’ QIIME2 tutorial [28]). Thus, samples below 1000 total reads were
180 not included in further analysis; this resulted in removal of two samples across all pipelines, which had
181 on average 11 (range: 4-21) and 255 (range: 150-341) total read counts across the pipelines. The final
182 dataset included 40 cases and 50 controls.

183 **2.2.2.1.2. OTU/ASV/reads table characteristics**

184 As a first part of the analysis, we compared the results of the pipelines in terms of characteristics and
185 distribution of reads, OTUs/ASVs, singletons (a single sequence), unclassified reads, and taxa. The
186 analyses were focused on the genus level, since this is the level at which most (clinical) studies focus
187 to identify an association with a disease/disorder status. This is due to the fact that analysis based on
188 16S rRNA gene hypervariable region(s) limits the taxonomic resolution to family- or genus-level [29].

189 We visualized overlapping genera between the pipelines using a Venn Diagram. In order to see how
190 the percentage of overlapping genera changed based on different filtering thresholds, we compared
191 the gut microbiome composition of: A) all the genera, B) genera after applying a 10% prevalence cut-
192 off, C) genera with relative abundance >0.1%, and D) genera with relative abundance <0.1%.

193 **2.2.2.1.3. Beta-diversity**

194 While beta diversity analysis is typically performed at the level of OTU/ASV, we did it at the genus level
195 in order to be able to compare the microbial composition (relative abundance; Bray-Curtis dissimilarity
196 metric) and structure (presence/absence; Jaccard similarity index) [30] across different bioinformatics
197 pipelines. The statistical significance of this comparison was determined using Permutational
198 Multivariate Analysis of Variance (PERMANOVA) using the R package 'adonis' for all pipelines; as a post
199 hoc analysis, we performed pairwise analysis between all pipelines [31]. The results were visualized by
200 unconstrained (Principal Coordinate Analysis, PCoA) and constrained (Canonical Analysis of Principal
201 coordinates, CAP) ordination methods [32] by applying following formula: ordinate(ps.merged.rel,
202 "CAP", "bray", ~ Pipeline). Additionally, we computed Tukey Honest Significant Differences (TukeyHSD;
203 calculated based on betadisper using the R package 'vegan' [31, 33, 34]) to expand the PCoA analysis
204 and to investigate intra-sample variation in a pairwise comparison manner.

205 **2.2.2.1.4. Comparative analysis at the genus level**

206 In order to obtain a more detailed overview of microbiome composition differences, we compared the
207 pipelines (i) in terms of the relative abundance of the ten most abundant genera (in order to maximize
208 our ability to find differences between the groups) and (ii) between cases and controls. At this stage,
209 we filtered out unclassified genera and applied a prevalence cut-off of 10% (at the genus level),
210 meaning that only genera present in >10% of the total number of samples were kept, in order to keep
211 the most informative data for the downstream statistical analysis [26]. Next, given the zero-inflated
212 nature of the data, a non-parametric (rank-based) test (Mann-Whitney U) was applied to evaluate
213 significant differences in relative abundances of bacterial genera between cases and controls. As we

214 aimed to evaluate the effects of the different pipelines rather than scale and significance of the
215 differences between them, this method seemed appropriate (see [12] for an extensive comparison of
216 abundance testing methods).

217 In analysing the consistency pattern of the case-control association results across pipelines, we
218 assigned a bioinformatics pipelines P-value Consistency Score (PCS, ranging from zero to five) to score
219 the number of pipelines showing statistically significant differences between groups per each genus
220 ($P<0.05$ unadjusted). A PCS=5 meant that all pipelines found significant differences ($P<0.05$
221 unadjusted) between cases and controls for a particular taxonomic group. Additionally, we calculated
222 a genus relative abundances case/control ratio (called Fold-Change, FC) and compared it (as an effect
223 measure) between the pipelines. The FC was calculated by using the foldchange() function from the
224 “gtools” package (v.3.8.1) [35]. FC was computed as follows: case/control if case>control, and as -
225 control/case otherwise. Furthermore, we tested the correlation between the PCS and the average
226 relative abundance (RA; per genus for all the pipelines) and average percentage of zeros of each genus
227 based on all pipelines.

228 All analyses were performed in RStudio (v.1.2.5033; R v.3.6.3) [36] using “phyloseq” (v.1.28.0) [37],
229 “microbiome” (v.1.6.0) [38], and “vegan” R packages [34], visualized by using “ggplot2” [39] (v.3.3.0),
230 “VennDiagram” [40] (v.1.6.20), “ggpubr” [41] (v.0.2.4), and “heatmaply” [42] (v.1.1.0) R packages;
231 statistical analyses where performed by using the “stats” R package (v.3.6.3) [39].

232 **2.2.2.2. Mock communities**

233 The main focus of the MC analysis was to compare observed to expected MC composition in order to
234 further evaluate the reliability and comparability of the pipelines. First, we compared the number of
235 genera observed to the expected MC composition. Second, beta-diversity was analysed as described
236 above. Third, we calculated Spearman's rho statistic via “stats” R package (v.3.6.3) [39] to (i) compare
237 the observed to the expected MC composition (relative abundance), and to (ii) compare the pipelines
238 against each other. In this way, we could identify the strength of correlation between the pipelines,

239 and identify strength of correlation between the pipelines and the expected MC composition. The
240 results were visualized by a heatmap using the “heatmaply” (v.1.1.0) R package [42] to identify any
241 inconsistencies across the pipelines.

242 **3. Results**

243 **3.1. NeuroIMAGE dataset**

244 **3.1.1. OTU/ASV/reads table characteristics**

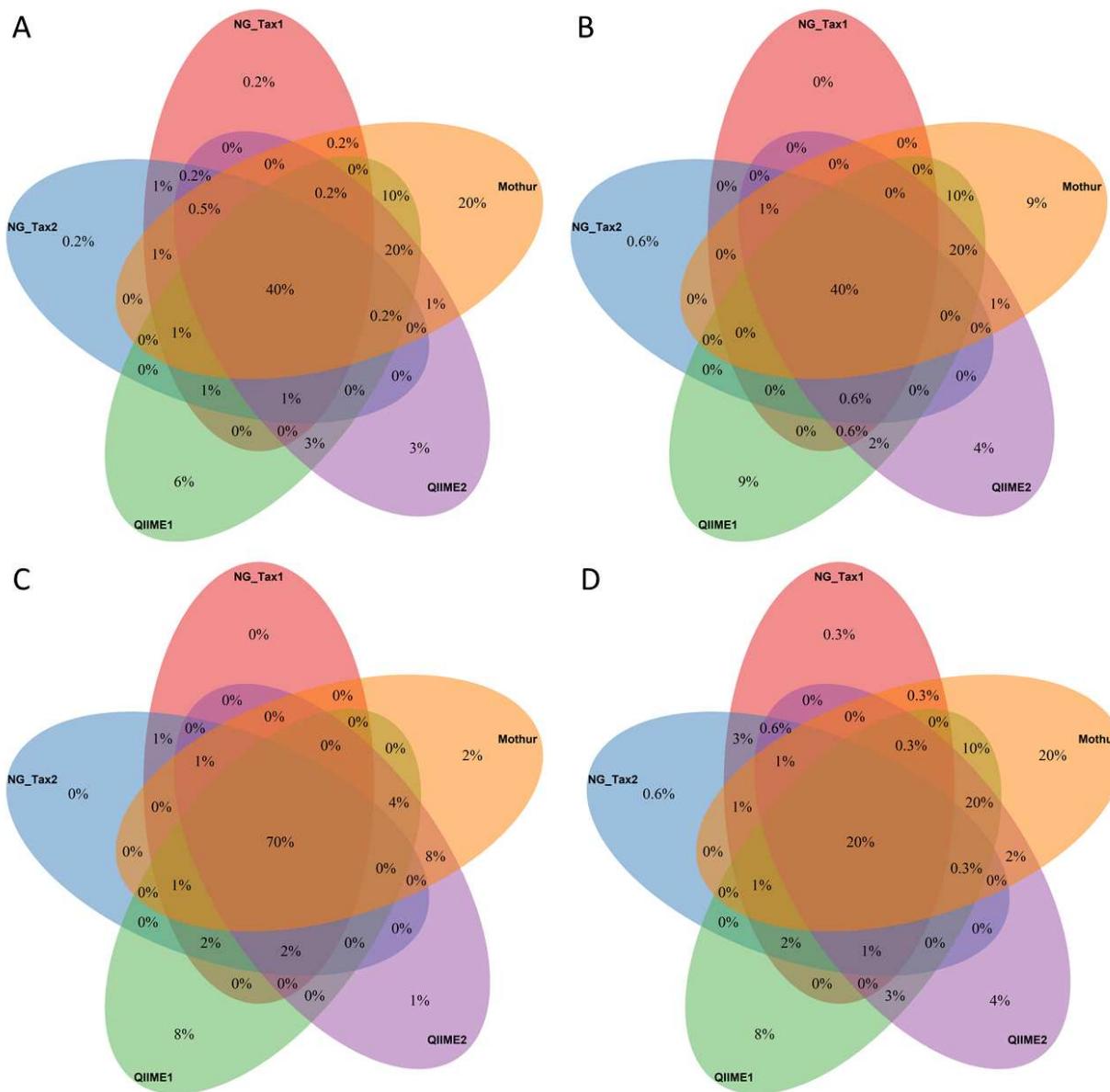
245 Table 1 shows the characteristics and distribution of OTUs/ASVs/reads per bioinformatic pipeline for
246 the complete study (N=90). We observed a high degree of variation across the pipelines for all the
247 variables. The total number of reads varied across the pipelines with QIIME1 showing the highest and
248 QIIME2 the lowest number of reads (percentage difference = 38.2%). Moreover, QIIME1 and mothur
249 showed the highest number of OTUs/ASVs, NG-Tax1 and NG-Tax2 showed the lowest (relative
250 difference ranging from 77.9% to 164.6%). Mothur showed the highest number of singletons (69.2%
251 of the total OTUs), but these only accounted for 0.67% of the total reads; these singletons did not
252 influence significantly the total relative abundance (when singletons were removed, the relative
253 abundance of other taxa was not influenced, data not shown). Furthermore, mothur and QIIME1
254 detected the biggest percentage of unclassified OTUs/ASVs (46.1% and 40.2%, respectively, at the
255 genus level), QIIME2 the lowest (4.7%).

256 **Table 1.** Summary of OTU/ASV characteristics between bioinformatics pipelines.

	NG-Tax1	NG-Tax2	QIIME1	QIIME2	mothur
Total number of final reads	1,414,916	1,357,891	1,692,581	1,149,886	1,390,041
Median of final reads per sample (IQR)	14,619 (7,648-20,997)	13,925 (7,411-19,998)	17,315 (8,783-25,819)	11,519 (5,385-17,515)	14,200 (7,173-20,742)
Total number of identified OTUs/ASVs	1,958	1,958	20,140	4,458	13,392
Number of singletons (% of total number of OTUs/ASVs)	0	0	1,291 (6.41)	3 (0.07)	9,269 (69.21)
Number of singletons (% of total number of OTUs/ASVs) before pre-processing step	0	0	0	7 (0.14)	10,206 (69.50)

Number of unclassified reads at the genus level (% of total reads)	202,165 (14.3)	193,698 (14.3)	427,601 (25.3)	23,091 (2.0)	23,404 (1.7)
Number of unclassified OTUs/ASVs at the genus level (% of total number of OTUs/ASVs)	321 (16.4)	321 (16.4)	8,092 (40.2)	210 (4.7)	6,170 (46.1)
Number of genera	177	176	312	254	343
Number of genera remaining after using a prevalence cut-off of 10% (% of total genera)	74 (41.8)	74 (42.1)	145 (46.5)	115 (45.3)	142 (41.4)
Number of genera below 0.1% relative abundance (% of total genera)	115 (65)	115 (65.3)	243 (77.8)	186 (73.2)	275 (80.2)
Number of phyla	10	10	13	14	15

257 IQR = interquartile range
258 Important to mention, the number of singletons for QIIME1 was the effect of pre-processing (removal of the subthreshold
259 group and samples having > 1000 reads). As a default setting, all the pipelines, except QIIME2 and mothur, remove
260 singletons (see Number of singletons (% of total number of OTUs/ASVs) before pre-processing step).
261 Of the genera detected by NG-Tax1, NG-Tax2, QIIME1, QIIME2 and mothur, only 40% overlapped
262 between all pipelines (Figure 2A). After applying the 10% prevalence cut-off to preserve the most
263 informative data for the downstream statistical analysis, 41.4% to 46.5% of the genera remained (Table
264 1). The prevalence cut-off did not improve the percentage of overlapping genera (Figure 2B), indicating
265 that more prevalent genera are not necessarily shared across the results from the different pipelines.
266 The relative abundance threshold did improve the percentage of overlapping genera; genera above
267 0.1% were more commonly shared across pipelines (70%) than genera below 0.1% (20%) (Figure 2C,D).



268

Figure 2. Venn diagram showing overlap between genera produced by five different bioinformatics pipelines. A) represents the overlap of genera based on raw data (based on 413 genera across pipelines), B) represents the overlap of genera after a 10% prevalence cut-off across samples (based on 171 genera across pipelines), C) overlap of genera with relative abundance >0.1% (N=80, genera across pipelines), and D) overlap of genera with relative abundance <0.1% (N=357 genera across pipelines).

274 3.1.2. Beta-diversity

275 Unconstrained PCoA plots based on the Bray-Curtis measure revealed that samples clustered based
276 on the sample ID rather than the bioinformatics pipelines (Figure 3A). However, the constrained
277 ordination method, CAP analysis, indicated relevant differences between the pipelines in terms of
278 microbial composition (Bray-Curtis index) at the genus level (Figure 3B). The CAP analyses captured
279 the variation in community structure in the first two components (CAP 1 and CAP 2) accounting for

280 11.1% of the total variance (Figure 3B). The same results were observed in terms of microbiome
281 structure using Jaccard's similarity index (Figure S1). PERMANOVA analysis supported the results by
282 revealing that microbial composition (Bray-Curtis: $R^2=13.9\%$, $p<0.001$) and structure (Jaccard: $R^2=9.5\%$,
283 $p<0.001$) differed significantly between the pipelines and, as expected, more variability was explained
284 by the same sample ID (Bray-Curtis: $R^2=89.5\%$; $p<0.001$ and Jaccard: $R^2=82.8\%$; $p<0.001$). Additionally,
285 we performed a pairwise comparison of group means dispersions (TukeyHSD). The analysis confirmed
286 that the intra-sample variation is quite similar across the pipelines, except for QIIME1 (Figure 3C).

287 The CAP analysis also showed that NG-Tax1 and NG-Tax2 clustered together, and QIIME2 clustered
288 with mothur (Figure 3C,D). We investigated these results in more detail, by running PERMANOVA
289 again, this time only with NG-Tax1 and NG-Tax2 or with QIIME2 and mothur, to investigate how
290 statistically different these clusters were. The results indicated statistically significant differences
291 between the pipelines, however, with very small percentages of explained variation (NG-Tax1/NG-tax2
292 $R^2=0.016\%$, $p<0.001$; QIIME2/mothur $R^2=0.9\%$, $p<0.001$; the results of pairwise PERMANOVA analyses
293 for other combinations can be found in Supplementary Table S1).

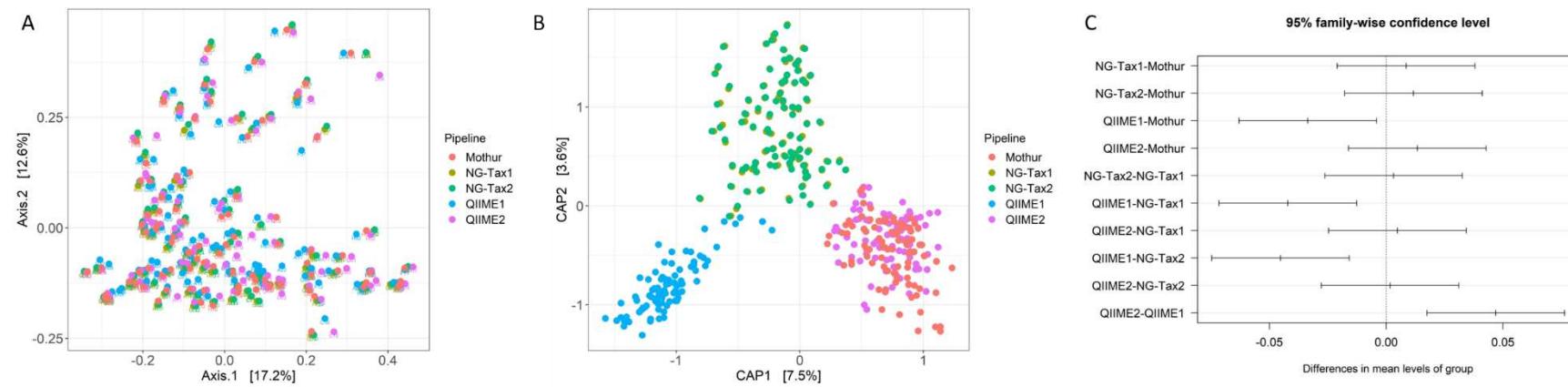
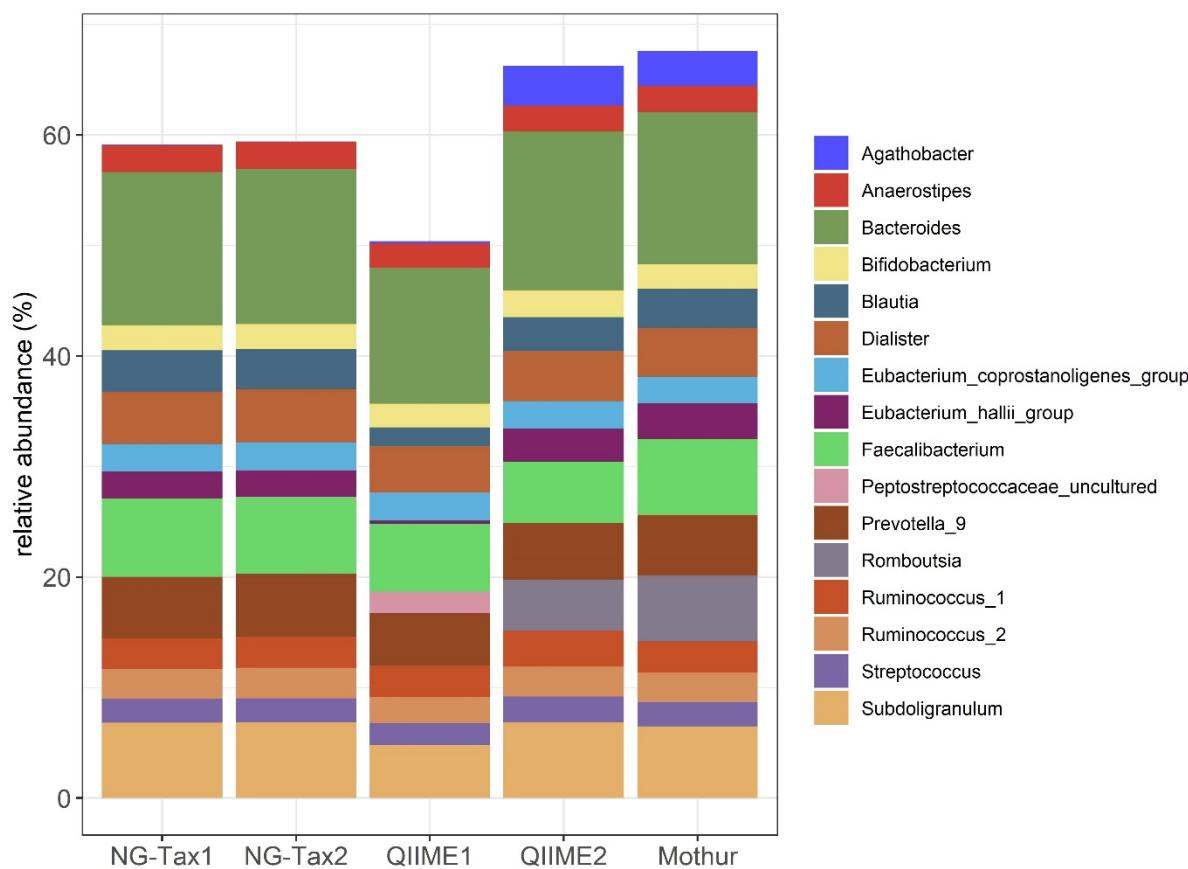


Figure 3. Results for the Bray-Curtis dissimilarity metric. A) Principal Coordinates Analysis (PCoA) plots with the percentage explained variance by the principal coordinates. B) Canonical Analysis of Principal coordinates (CAP) ordination plot of structure in microbial communities associated with bioinformatics pipelines. C) TukeyHSD, a pairwise comparison of group mean dispersions revealed that the intra-sample variation was quite similar across pipelines, with QIIME1 forming the exception.

295 **3.1.3. Comparative analysis of individual genera**

296 We also compared the distribution of the ten most abundant genera found by each pipeline (Figure 4).
297 These genera were not identical across the pipelines: across the five pipelines, 16 unique genera were
298 observed. The RA values for all of the 16 unique genera were statistically significantly different
299 between pipelines (Friedman test, Bonferroni-adjusted p-values <0.001). The descriptive statistics of
300 this data can be found in Supplementary Table S2.



301
302 **Figure 4.** Bacterial genera profile. Top 10 most abundant bacterial genera per pipeline resulted in a total of 16 unique genera.
303 We excluded unclassified genera, since they represent a group of genera rather a single genus.

304 **3.1.4. Taxonomic differences between cases and controls across pipelines**

305 We carried out univariate testing of the relative abundance of individual genera between ADHD cases
306 (N=40) and controls (N=50) in order to investigate if the downstream statistical conclusions were
307 consistent across the pipelines. In total, 10 genera showed nominally significant differences ($p < 0.05$)

308 between cases and controls in at least one pipeline (Table 2), but these differences were not consistent
309 across all pipelines. Based on the P-value consistency score (PCS), only one of the 10 genera showed
310 total agreement in terms of PCS (PCS=5), none showed high agreement (PCS=4), three genera showed
311 moderate agreement (PCS=3), and two genera showed partial agreement (PCS=2). The rest of the
312 genera (N=4) showed no agreement (PCS=1) (Table 2). The descriptive statistics of the 10 genera can
313 be found in the Supplementary Table S3.

314 In order to determine the effect of the differences in genus abundance on the case-control comparison
315 between the pipelines, we compared Fold Change (FC) based on genera relative abundance (Table 2).
316 Three observations stand out. First, the FC differs between the pipelines. For example, for
317 *Clostridiales_vadinBB60_group uncultured_bacterium*, QIIME1 resulted in a case/control ratio of 1.19,
318 whereas QIIME2 resulted in a ratio of 2.97. Second, for both versions of QIIME, the FC of *Coprococcus_2*
319 was in the opposite direction compared to the other three pipelines. Third, in some cases (e.g.,
320 *Prevotella_9*, *Ruminococcus_1*), the FC was almost the same between the pipelines, but still only one
321 pipeline indicated nominal significance.

322 In general, some genera were missing in some pipelines, and there were differences in effect size or
323 even in direction between pipelines for genera that were nominally significant different between cases
324 and controls. The non-parametric rank test indicated that genera present in all pipelines (N=6) differed
325 statistically in their relative abundance among the pipelines (Friedman test, Bonferroni-adjusted p-
326 values <0.002, Supplementary Table S3).

327 Testing the correlation between PCS and two measures of frequency, relative abundance and the
328 percentage of zeros, we found the correlation coefficient between PCS and relative abundance to be
329 $r_{PCS-RA}=0.58$ and the one between PCS and percentage of zeros to be $r_{PCS-\%0}=-0.24$ (Figure S2A,B). Both
330 correlations were non-significant ($p>0.05$), however, suggesting that the consistency across the
331 pipelines was independent of bacterial relative abundance and the observed percentage of zeros. The

332 lack of significance should be treated with caution, as it could be a result of the low number of features
333 included in the analysis (n=10 genera).

334 **Table 2.** The table represent a fold change (case/control ratio), p-value consistency score (PCS), and percentage of zeros for genera which were nominally significant ($p<0.05$) different
 335 between cases and controls by at least one pipeline. Values highlighted in red indicate nominal significance ($p<0.05$). A negative value indicates that the cases' mean is lower than the controls'
 336 mean.

Genera	NG-Tax1	NG-Tax2	QIIME1	QIIME2	mothur	PCS	% of zeros
Fold Change							
Coprococcus_2	1.09	1.12	-1.24	-1.06	1.05	5	40
Prevotella_9	-1.83	-1.81	-2.02	-1.76	-1.87	3	35
Ruminococcus_1	-1.51	-1.50	-1.49	-1.49	-1.55	3	8
Eubacterium_eligens_group	-1.61	-1.48	-1.58	-1.92	-1.65	3	62
Tyzzerella_3	1.02	-1.02	NA	1.88	1.77	2	74
Howardella	NA	NA	4.45	4.88	NA	2	82
Eubacterium_ventriosum_group	-2.32	-2.34	-1.93	-2.31	-2.02	1	17
Fusicatenibacter	-1.66	-1.69	1.10	1.24	1.03	1	47
Clostridiales_vadinBB60_group_uncultured_bacterium	NA	NA	1.19	2.97	NA	1	74
Lachnospiraceae_UCG_004	NA	NA	-1.56	NA	-1.13	1	49

339 **3.2. Mock communities**

340 **3.2.1. Genus richness**

341 Mothur identified the highest and NG-Tax1 and NG-Tax2 the lowest number of genera in both MCs.

342 NG-Tax1 ($N_{MC3}=31$, $N_{MC4}=25$), NG-Tax2 ($N_{MC3}=31$, $N_{MC4}=25$) and QIIME2 ($N_{MC3}=39$, $N_{MC4}=36$) approached

343 the expected genus richness ($N_{MC3}=36$, $N_{MC4}=36$) closer than QIIME1 ($N_{MC3}=64$, $N_{MC4}=67$) and mothur

344 ($N_{MC3}=84$, $N_{MC4}=101$) (Table S4).

345 **3.2.2. Beta-diversity**

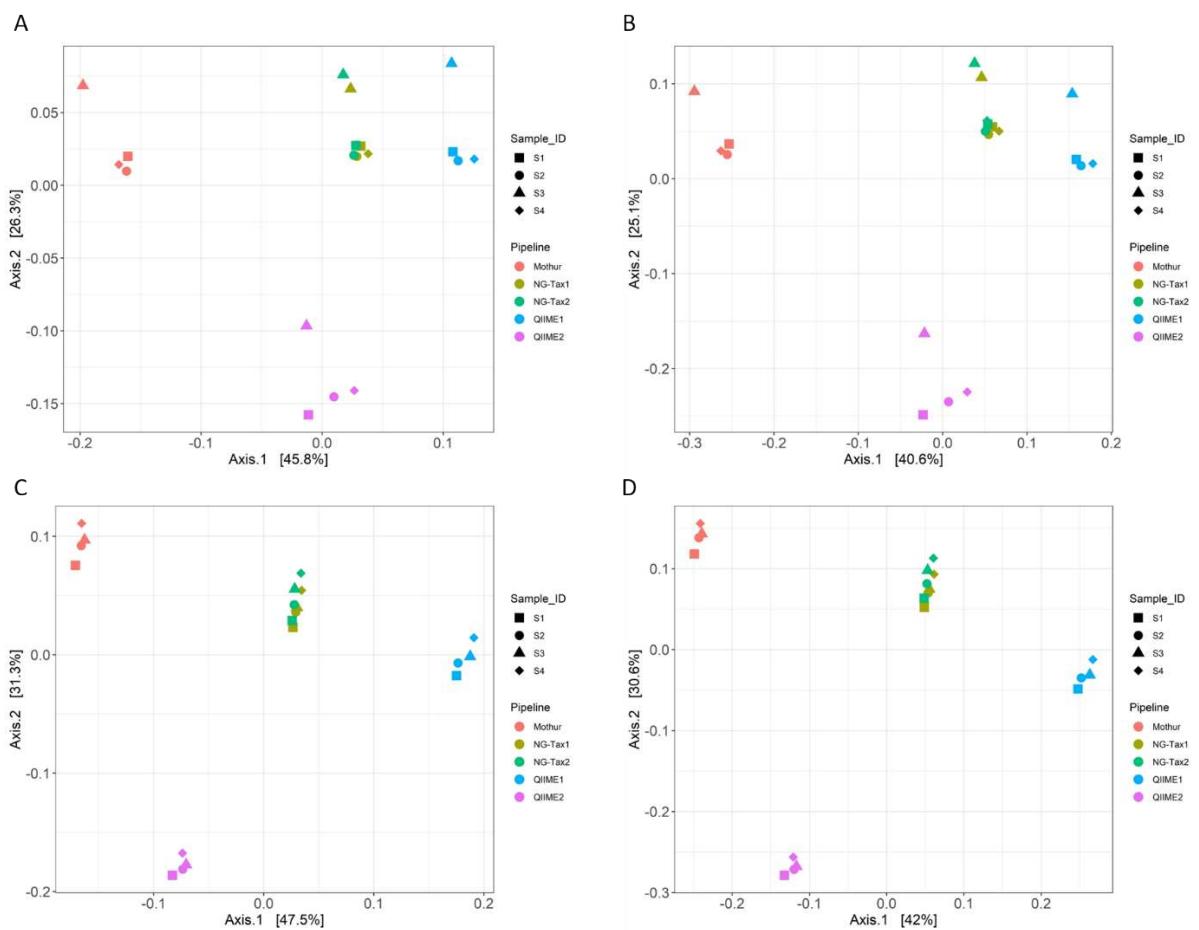
346 We also compared the observed and expected beta-diversity (at genus level) in the MCs. PCoA plots

347 based on Bray-Curtis and Jaccard measures revealed that samples clustered based on the pipelines

348 (Figure 5). 90% (for MC3) and 98% (for MC4) of total microbial composition variance (Bray-Curtis,

349 $p_{MC3}<0.001$ and $p_{MC4}<0.001$) and 87% (in case of MC3) and 97% (in case of MC4) of total microbial

350 structure variance was explained by pipelines (Jaccard, $p_{MC3}<0.001$ and $p_{MC4}<0.001$).



351

352 **Figure 5.** PCoA of MC composition was affected by the choice of bioinformatics pipelines. Results of the Bray-Curtis
353 dissimilarity metric and Jaccard similarity index based on MC3 are shown in panel A and B, respectively, and based on MC4
354 are shown in C and D, respectively. S1 = Sample 1.

355

356 3.2.3. Correlation analysis

357 The correlation of observed and expected MC relative abundance (based on N=36 genera) showed that
358 QIIME2 had the highest correlation coefficient ($r_{MC3}=0.70$, $r_{MC4}=0.76$), followed by mothur ($r_{MC3}=0.67$,
359 $r_{MC4}=0.65$), QIIME1 ($r_{MC3}=0.61$, $r_{MC4}=0.64$), NG-Tax1 ($r_{MC3}=0.56$, $R_{MC4}=0.61$) and NG-Tax2 ($r_{MC3}=0.56$,
360 $r_{MC4}=0.61$) (Figure S3 A,B).

361 3.2.4. Comparative analysis of individual genera

362 Comparison of individual genera showed inconsistencies across pipelines for both MCs (Figure S4, S5).
363 For example, NG-Tax1 and NG-Tax2 did not detect *Enterobacter* and *Dorea*, while QIIME2 did not

364 detect *Serratia*, mothur did not detect *Klebsiella*, while QIIME1 did not detect *Anaerostipes* from either
365 MCs. All pipelines failed to classify *Salmonella*. Some pipelines under/overrepresented certain genera;
366 for example, QIIME1 overrepresented *Enterobacter* and *Pseudomonas*; NG-Tax1 and NG-Tax2
367 overrepresented *Klebsiella*. As expected, NG-Tax1 and NG-Tax2 did not detect genera below 0.1%
368 abundance included in MC4 (due to the abundance cut-off setting) (Figure S5), whereas QIIME2 did
369 not detect genera below 0.01%.

370 **4. Discussion**

371 *Summary*

372 In this study, we compared five frequently used bioinformatics pipelines for the processing of 16S rRNA
373 gene amplicon sequencing data, NG-Tax1, NG-Tax2, QIIME1, QIIME2 and mothur, to determine
374 whether and in which way the analytical methods of each of these pipelines affect the downstream
375 statistical analysis results. For this purpose, we used a clinical (case-control) dataset as well as two
376 mock communities. Based on the clinical sample, we found that NG-Tax1 and NG-Tax2 were strikingly
377 similar in terms of the number of reads/OTUs/ASVs, number of singletons, number of unclassified
378 reads/OTUs/ASVs at the genus level, and number of phyla and genera. This abundance table
379 characteristics were reflected in the results of the beta-diversity analysis, where NG-Tax1 and NG-Tax2
380 clustered together based on the genera relative abundance. In both versions of NG-Tax, the same
381 genera were indicated as nominally significantly different, and the FC was almost the same. While
382 output of both NG-Tax versions largely overlapped, output varied greatly compared to the other
383 pipelines (QIIME 1, QIIME2, mothur) in terms of, amongst others, the number of singletons, number
384 of unclassified reads/OTUs/ASVs at the genus level and number of genera. Consequently, we showed
385 that only 40% of genera overlap between all the pipelines. The percentage increased to 70% when
386 applying a 10% prevalence cut-off, thereby only comparing genera with RA > 0.1%. The beta-diversity
387 results indicated that, although the samples cluster better according to sample ID than bioinformatics
388 pipelines, all pipelines detected different patterns of microbial composition (Bray-Curtis) and structure

389 (Jaccard), where QIIME1 diverged the most from the other pipelines. In terms of taxonomy, the most
390 abundant genera across the pipelines differed significantly between the pipelines. More importantly,
391 the conclusions of the case-control comparison varied; out of 10 unique genera showing a case-control
392 difference, only one overlapped between all 5 pipelines. Pipelines differed not only in the number of
393 genera showing a case-control difference, but also in the magnitude and even direction of this effect.
394 Overall, the results indicate a clear lack of consistency across the pipelines.

395 Based on the MCs, we found that QIIME1 and mothur overestimated genus richness, where NG-Tax1,
396 NG-Tax2 and QIIME2 approached the expected genus richness. Beta-diversity analyses indicated that
397 the pipelines differed in representing expected microbial composition and structure, with NG-Tax1 and
398 NG-Tax2 clustering together. Furthermore, correlation analysis between observed and expected MC
399 indicated that, of all pipelines, QIIME2 came closest to the expected microbiome composition.
400 Comparative analysis of individual genera showed that the average relative abundance of specific taxa
401 varied depending on the bioinformatic pipeline. Overall, MC-based results confirmed that the output
402 of pipelines differed in terms of microbiome composition and structure. These results show how the
403 choice of bioinformatic pipeline not only impacts the analysis of 16S rRNA gene sequencing data but
404 also the downstream association results.

405 *Pipeline characteristics*

406 QIIME1 yielded different results compared to its successor QIIME2 and the other pipelines, mainly
407 regarding the highest number of total and median reads per sample, (unclassified) OTUs and
408 prevalence-filtered genera. Since January 2018, QIIME1 is not supported anymore by developers and
409 has been replaced by QIIME2. This suggests that if data processed using QIIME1 would be reanalysed
410 with QIIME2 or another pipeline, it would yield different results. Furthermore, we observed that
411 QIIME1 yielded the highest number of unique taxa [6, 25]. MC-based results suggested that QIIME1
412 (and mothur) overrepresented bacterial richness. Thus, in agreement with Prodan et al. (2020), our
413 advice is that for biological conclusions based on alpha-diversity, QIIME1 users should switch to

414 another pipeline or at least confirm their results with another pipeline [6]. For users interested in low
415 frequency taxa, our study showed that QIIME1 and mothur are most appropriate, as they detected
416 more low abundant genera (abundance <0.01%) compared to QIIME2, NG-Tax1 and NG-Tax2 (with
417 NG-Tax being stricter than QIIME2); however, researchers should take into account that this comes at
418 the costs of having a higher number of spurious taxa.

419 There is dispute in the research community regarding the matter of keeping or removing singletons,
420 and on the best method to remove them. By default, mothur and QIIME2 keep the singletons (69.5%
421 of total OTU/ASVs compared to 0.14% in QIIME2). Both pipelines have different ways of dealing with
422 singletons [19, 21], where mothur yielded highest percentage of singletons. Many of these reads might
423 be noise [43]. Indeed, based on the MCs, we saw that singletons might explain a large number (65% in
424 case of MC3, 40% in case of MC4) of spurious genera (data not shown). However, effects on relative
425 abundance were limited, since singletons accounted for only 0.64% of total reads (for the NeuroIMAGE
426 dataset). Based on these results, we suggest to remove singletons even with the pipelines that suggest
427 keeping them. In addition to the effects on the structure (presence/absence of genera), very low
428 frequency values pose a great challenge for statistical analysis. This is especially relevant if data are
429 analysed at the OTU/ASV level.

430 This is the first time the output (relative abundance table) of the five pipelines is used together to
431 detect case-control differences and evaluate their consistency and stability in a common statistical
432 framework. Other researchers compared some of these pipelines, and findings partly overlap with
433 ours. For instance, Ducarmon et al. (2020) compared NG-Tax1 and QIIME2 and concluded that the
434 pipelines showed different results in terms of richness [13]. In concordance with our study, NG-Tax1
435 accurately retrieved richness at the genus level. However, QIIME2 overestimated genus-based
436 richness, whereas in our paper it approached the expected richness in MCs. Furthermore, we observed
437 that the choice of pipeline influenced the analyses of bacterial composition and structure, whereas in
438 the analysis reported by Ducarmon et al. (2020), diversity and compositional profiles were comparable.

439 With regard to the MCs, in Ducarmon et al. (2020), QIIME2 failed to classify *Salmonella*, and NG-Tax1
440 detected *Salmonella*, whereas in our study, none of the pipelines detected this genus. This could be
441 due to the difference in the expected RA. In our case, it was 1.2% for MC3 and 2.5% for MC4. For
442 Ducarmon et al. (2020), it was approximately 9%. When looking closer at QIIME2 performance,
443 Almeida et al. (2018) suggested QIIME2 as an optimal pipeline for 16S rRNA gene profiling based on
444 the lowest distance between the expected and observed sample compositions based on synthetic,
445 simulated datasets, and based on the best recall and precision [44]. We observed similar results, where
446 correlations between expected and observed MC composition were highest for QIIME2. In addition
447 to that, according to Prodan et al. (2020), DADA2 (we used QIIME2 with the DADA2 option as a
448 denoising algorithm) offered the best sensitivity, detecting all 22 true ASVs present in their MC [6].
449 Moreover, our results agree with those of Allali et al. (2017), where DADA2 resulted in lower numbers
450 of ASVs when compared to the number of OTUs of QIIME1 [25] and mothur (this paper); however, this
451 was not seen when comparing QIIME2 to NG-Tax1 and NG-Tax2, suggesting that NG-Tax is even more
452 strict than QIIME2 in terms of quality control settings (e.g., abundance threshold). Altogether, based
453 on our results and existing comparisons, QIIME2 (or DADA2) is a highly recommended pipeline for
454 microbiome research.

455 Studies investigating differences between bioinformatics pipelines have so far focused on general
456 characteristics of the OTUs/ASVs/reads such as richness, diversity and microbial compositional profiles
457 rather than the biological conclusions to be drawn from comparing these characteristics e.g., between
458 clinically relevant groups [6, 13, 14]. One study investigating if the same biological conclusions could
459 be reached using different pipelines was Allali et al. (2017), based on data from chicken cecum
460 microbiome (vaccinated, prebiotic treated, control group). They tested different settings of QIIME1,
461 UPARSE and DADA2 and concluded that they could discriminate samples by treatment, despite
462 differences in diversity and abundance, leading to similar biological conclusions [25]. Allali et al. (2017)
463 based their conclusion on beta-diversity rather than a comparative analysis of individual genera (as
464 presented in the current paper). However, they reported differences in RA of specific genera between

465 pipelines, suggesting that also in their data different pipelines resulted in different lists of genera
466 discriminating between treatments. In our study, MC analysis helped to interpret clinical data. The
467 results (e.g., beta-diversity) showed that the MC-based analysis does not necessarily reflect the real
468 dataset as the complexity of a real microbiota sample is much larger. This underlines the importance
469 of deciding which pipeline best serves the analysis of your dataset based on how this pipeline performs
470 on real data as well as MCs.

471 *Limitations and open questions*

472 Our results should be viewed in the context of some limitations. Our study was limited by a small
473 sample size (N=90), but taking into consideration that this is a crossover study the sample size should
474 be sufficient to detect the differences in the output produced by each of the pipelines and how these
475 differences affect the downstream statistical analysis. Nevertheless, since microbiome data is
476 notoriously diverse and sensitive to protocol and technical variations [45, 46], the effect of datasets
477 with different designs should be investigated. Another limitation of this study was the use of nominal
478 (and standard) statistical significance cut-off ($p<0.05$) as a measure of statistical difference.
479 Considering the number of tested genera, several false positives could be expected. Although a
480 corrected p-value is considered a better measure of success, the case-control study may not contain
481 large enough differences or enough statistical power to properly classify the differences between
482 groups as statistically significant. Given the aim of this paper, establishing the true (biological)
483 difference between groups is not evaluated and comes second to the difference in observed effects
484 brought in by the choice of the bioinformatic pipeline, which is why nominal significance was sufficient
485 to select multiple taxa (showing different RA and p-values across pipelines) and evaluate the effect on
486 analysis. Lastly, the number of ASVs/OTUs varied considerably between pipelines, which can result in
487 differences in FC magnitude, as seen for example in case-control ratio differences between QIIME1
488 and QIIME2 on the *Clostridiales vadin* genus group. A different direction of FC could be driven by a
489 differential effect of filtering/denoising steps per group, potentially driven by a larger number of

490 sequencing artefacts in either of them. Future research should focus on more technical aspects of
491 bioinformatics pipelines comparisons, to identify what exactly drives such differences.

492 **5. Conclusion**

493 Our results indicate that a choice of bioinformatic pipeline has not only an impact on the analysis of
494 16S rRNA gene sequencing data but also the case-control comparison results. This means that the
495 choice of pipeline can influence the list of significantly different genera between study groups. Thus,
496 we underscore a significant limiting factor in current microbiome research: the lack of consistency
497 between study results and how this limits their comparability and the validity of conclusions to be
498 drawn from them.

499 Based on our results we recommend i) using QIIME1 and Mothur to researchers that are interested in
500 rare and/or low-abundant taxa, ii) using NG-Tax1 or NG-Tax2 when favouring strict artefact filtering,
501 iii) using QIIME2 when looking for a balance between the two abovementioned points, and iv) using at
502 least two pipelines to assess the stability of results.

503 We would like to point out that the field still needs to develop “best practice” for microbiome analysis
504 and apply it consensually across studies, before we can have a deeper understanding of the gut
505 microbiota’s contribution to human health and disease. With our current work, we hope to contribute
506 to the gut microbiota research field and make other researchers aware of the strengths and limitations
507 of their choice of bioinformatic pipeline in terms of influencing the results of case-control studies with
508 16S rRNA marker gene sequencing data.

509

510 **Declarations**

511 **Ethics approval and consent to participate**

512 The study was approved by the Institutional Review Board at Radboud University Medical Centre,
513 Nijmegen, The Netherlands (registration number 2012/542; NL nr.: 41950.091.12). An informed

514 written consent was obtained from all participants and/or their parents prior to the sample and data
515 collections.

516 **Data Availability**

517 The data underlying this article will be shared on reasonable request to the corresponding author.

518 **Competing interests**

519 The authors declare that they have no competing interests.

520 **Funding**

521 This project was supported by the European Union's Horizon 2020 research and innovation
522 programme under the Marie Skłodowska-Curie [grant agreement number 643051]. The funding body
523 had no role in the preparation of the manuscript. The research leading to these results also received
524 funding from the European Community's Horizon 2020 research and innovation programme under
525 project no. 728018 (Eat2BeNICE) and 667302 (CoCA).

526 **Acknowledgements**

527 The authors would like to thank Prokopis Konstanti, Ward de Witte, Fini de Gruyter, QIIME2 forum
528 (<https://forum.qiime2.org/>) and mothur forum (<https://forum.mothur.org/>) for helpful discussions.

529

530 **Supplementary Figures:**

531 **Figure S1.** Results of Jaccard similarity index. *A*) Principal Coordinates Analysis (PCoA) plots with the
532 percentage explained variance by the principal coordinates. *B*) Canonical Analysis of Principal
533 coordinates (CAP) ordination plot reveals structure in microbial communities associated with
534 bioinformatics pipelines. *C*) TukeyHSD, a pairwise comparison of group mean dispersions, revealed
535 that the intra-sample variation is quite similar across pipelines, except for QIIME1.

536 **Figure S2.** Scatter plot representing a correlation of the PCS with the relative abundance (A) and
537 prevalence (B) based on the 10 genera showing nominally significant differences ($p < 0.05$) between
538 cases and controls in at least one pipeline (Table 2).

539 **Figure S3.** Correlation matrix of Spearman correlation coefficient values between observed mock
540 community (OBS MC) composition as a result of five different bioinformatics pipelines (NG-Tax1, NG-
541 Tax2, QIIME1, QIIME2 and mothur) and corresponding expected mock composition (EXP MC), Mock_3
542 (A) and Mock_4 (B). The results are based on genera only present in EXP MCs. The observed values
543 represent statistically significant correlations ($P < 0.05$).

544 **Figure S4.** Interactive heatmap of the expected (EXP) and observed (OBS) MC3 based on all genera
545 (N=36) present in EXP MC. The rows of the matrix are ordered to highlight patterns by using default
546 settings.

547 **Figure S5.** Interactive heatmap of the expected (EXP) and observed (OBS) MC4 based on all genera
548 present (N=36) in EXP MC. The rows of the matrix are ordered to highlight patterns by using default
549 settings.

550

551 **Supplementary Tables:**

552 **Table S1.** Results of pairwise PERMANOVA.

553 **Table S2.** Descriptive statistics of 16 most abundant genera.

554 **Table S3.** Descriptive statistics of 10 genera shown in Table 2.

555 **Table S4.** Number of genera based on observed and expected (EXP) MC.

556

557 **References**

558 1. Cryan JF, O'Riordan KJ, Cowan CSM, Sandhu K V, Bastiaanssen TFS, Boehme M, et al. The
559 Microbiota-Gut-Brain Axis. *Physiol Rev.* 2019;99:1877–2013. doi:10.1152/physrev.00018.2018.

560 2. Saffouri GB, Shields-Cutler RR, Chen J, Yang Y, Lekatz HR, Hale VL, et al. Small intestinal microbial
561 dysbiosis underlies symptoms associated with functional gastrointestinal disorders. *Nat Commun.*
562 2019;10:2012. doi:10.1038/s41467-019-09964-7.

563 3. Kang DW, Park JG, Ilhan ZE, Wallstrom G, Labaer J, Adams JB, et al. Reduced incidence of
564 Prevotella and other fermenters in intestinal microflora of autistic children. *PLoS One.*
565 2013;8:e68322. doi:10.1371/journal.pone.0068322.

566 4. Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, et al. The neuroactive potential
567 of the human gut microbiota in quality of life and depression. *Nat Microbiol.* 2019;4:623–32.
568 doi:10.1038/s41564-018-0337-x.

569 5. Dam SA, Mostert JC, Szopinska-Tokov JW, Bloemendaal M, Amato M, Arias-Vasquez A. The Role of
570 the Gut-Brain Axis in Attention-Deficit/Hyperactivity Disorder. *Gastroenterol Clin North Am.*
571 2019;48:407–31. doi:10.1016/j.gtc.2019.05.001.

572 6. Prodan A, Tremaroli V, Brolin H, Zwinderman AH, Nieuwdorp M, Levin E. Comparing bioinformatics
573 pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One.* 2020;15:e0227434.
574 doi:10.1371/journal.pone.0227434.

575 7. Knight R, Vrbanac A, Taylor BC, Aksakov A, Callewaert C, Debelius J, et al. Best practices for
576 analysing microbiomes. *Nat Rev Microbiol.* 2018;16:410–22.

577 8. Xia Y, Sun J, Chen D-G. Bioinformatic Analysis of Microbiome Data. In: *Statistical Analysis of*
578 *Microbiome Data with R.* 2018. p. 5–11.

579 9. Rintala A, Pietila S, Munukka E, Eerola E, Pursiheimo JP, Laiho A, et al. Gut Microbiota Analysis
580 Results Are Highly Dependent on the 16S rRNA Gene Target Region, Whereas the Impact of DNA
581 Extraction Is Minor. *J Biomol Tech.* 2017;28:19–30. doi:10.7171/jbt.17-2801-003.

582 10. Szopinska JW, Gresse R, van der Marel S, Boekhorst J, Lukovac S, van Swam I, et al. Reliability of a
583 participant-friendly fecal collection method for microbiome analyses: a step towards large sample
584 size investigation. *BMC Microbiol.* 2018;18:110. doi:10.1186/s12866-018-1249-x.

585 11. Balvociute M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies
586 compare? *BMC Genomics.* 2017;18 Suppl 2:114. doi:10.1186/s12864-017-3501-4.

587 12. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, et al. Microbiome
588 differential abundance methods produce different results across 38 datasets. *Nat Commun.*
589 2022;13:342. doi:10.1038/s41467-022-28034-z.

590 13. Ducarmon QR, Hornung BVH, Geelen AR, Kuijper EJ, Zwittink RD. Toward Standards in Clinical
591 Microbiota Studies: Comparison of Three DNA Extraction Methods and Two Bioinformatics pipelines.
592 *mSystems.* 2020;5. doi:10.1128/mSystems.00547-19.

593 14. Lopez-Garcia A, Pineda-Quiroga C, Atxaerandio R, Perez A, Hernandez I, Garcia-Rodriguez A, et al.
594 Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S
595 rRNA Amplicon Sequences. *Front Microbiol.* 2018;9:3010. doi:10.3389/fmicb.2018.03010.

596 15. Ramiro-Garcia J, Hermes GDA, Giatsis C, Sipkema D, Zoetendal EG, Schaap PJ, et al. NG-Tax, a
597 highly accurate and validated pipeline for analysis of 16S rRNA amplicons from complex biomes.
598 *F1000Res.* 2016;5:1791. doi:10.12688/f1000research.9227.2.

599 16. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible,
600 interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.*
601 2019;37:852–7. doi:10.1038/s41587-019-0209-9.

602 17. Poncheewin W, Hermes GDA, van Dam JCI, Koehorst JJ, Smidt H, Schaap PJ. NG-Tax 2.0: A
603 Semantic Framework for High-Throughput Amplicon Analysis. *Front Genet.* 2019;10:1366.
604 doi:10.3389/fgene.2019.01366.

605 18. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows

606 analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7:335–6.

607 doi:10.1038/nmeth.f.303.

608 19. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: 609 open-source, platform-independent, community-supported software for describing and comparing 610 microbial communities. *Appl Env Microbiol*. 2009;75:7537–41. doi:10.1128/AEM.01541-09.

611 20. Edgar RC. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nat Methods*. 612 2013;10:996–8. doi:10.1038/nmeth.2604.

613 21. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution 614 sample inference from Illumina amplicon data. *Nat Methods*. 2016;13:581–3.

615 doi:10.1038/nmeth.3869.

616 22. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly 617 Resolves Single-Nucleotide Community Sequence Patterns. *mSystems*. 2017;2.

618 doi:10.1128/mSystems.00191-16.

619 23. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 620 2010;26:2460–1. doi:10.1093/bioinformatics/btq461.

621 24. Pittayanon R, Lau JT, Leontiadis GI, Tse F, Yuan Y, Surette M, et al. Differences in Gut Microbiota 622 in Patients With vs Without Inflammatory Bowel Diseases: A Systematic Review. *Gastroenterology*. 623 2020;158:930-946 e1. doi:10.1053/j.gastro.2019.11.294.

624 25. Allali I, Arnold JW, Roach J, Cadenas MB, Butz N, Hassan HM, et al. A comparison of sequencing 625 platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC* 626 *Microbiol*. 2017;17:194. doi:10.1186/s12866-017-1101-8.

627 26. Szopinska-Tokov J, Dam S, Naaijen J, Konstanti P, Rommelse N, Belzer C, et al. Correction: 628 Szopinska-Tokov et al. Investigating the Gut Microbiota Composition of Individuals with Attention- 629 Deficit/Hyperactivity Disorder and Association with Symptoms. *Microorganisms* 2020, 8, 406.

630 Microorganisms . 2021;9.

631 27. von Rhein D, Mennes M, van Ewijk H, Groenman AP, Zwiers MP, Oosterlaan J, et al. The
632 NeuroIMAGE study: a prospective phenotypic, cognitive, genetic and MRI study in children with
633 attention-deficit/hyperactivity disorder. Design and descriptives. Eur Child Adolesc Psychiatry.
634 2015;24:265–81. doi:10.1007/s00787-014-0573-4.

635 28. QIIME2docs. “Moving Pictures” tutorial. 2019. <https://docs.qiime2.org/2019.4/tutorials/moving-pictures/>.

637 29. Earl JP, Adappa ND, Krol J, Bhat AS, Balashov S, Ehrlich RL, et al. Species-level bacterial
638 community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-
639 length 16S rRNA genes 06 Biological Sciences 0604 Genetics 06 Biological Sciences 0605
640 Microbiology. Microbiome. 2018;6:1–26. doi:10.1186/s40168-018-0569-2.

641 30. Xia Y, Sun J, Chen D-G. Community Diversity Measures and Calculations. In: Statistical Analysis of
642 Microbiome Data with R. 2018. p. 180–9.

643 31. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol.
644 2001;26:32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x.

645 32. Xia Y, Sun J, Chen D-G. Exploratory Analysis of Microbiome Data and Beyond. In: Statistical
646 Analysis of Microbiome Data with R. 2018. p. 208–48.

647 33. Anderson MJ, Ellingsen KE, McArdle BH. Multivariate dispersion as a measure of beta diversity.
648 Ecol Lett. 2006;9:683–93.

649 34. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community
650 Ecology Package. 2019. doi:10.4135/9781412971874.n145.

651 35. Warnes G, Bolker B, Lumley T. gtools: Various R Programming Tools. 2018. <https://cran.r-project.org/package=gtools>.

653 36. RStudio Team. RStudio: integrated development environment for R. RStudio, Inc, Boston, MA.

654 2019.

655 37. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics

656 of microbiome census data. *PLoS One*. 2013;8:e61217. doi:10.1371/journal.pone.0061217.

657 38. Lahti L, Shetty S. microbiome R package.

658 39. R Core Team (2020). R A language and environment for statistical computing. Vienna, Austria.

659 <http://www.r-project.org/index.html>. Accessed 14 Jul 2020.

660 40. Chen H, Boutros PC. VennDiagram: A package for the generation of highly-customizable Venn and

661 Euler diagrams in R. *BMC Bioinformatics*. 2011;12:35. doi:10.1186/1471-2105-12-35.

662 41. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots. R package version 0.2.4. 2019.

663 <https://cran.r-project.org/package=ggpubr>.

664 42. Galili T, O’Callaghan A, Sidi J, Sievert C. heatmaply: an R package for creating interactive cluster

665 heatmaps for online publishing. *Bioinformatics*. 2017;34:1600–2. doi:10.1093/bioinformatics/btx657.

666 43. Edgar RC. Singletons. <http://drive5.com/usearch/manual/singletons.html>. Accessed 14 Jul 2020.

667 44. Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on

668 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience*.

669 2018;7. doi:10.1093/gigascience/giy054.

670 45. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, et al. Meta-

671 analyses of studies of the human microbiota. *Genome Res*. 2013;23:1704–14.

672 doi:10.1101/gr.151803.112.

673 46. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results.

674 *Genome Biol*. 2010;11:220. doi:10.1186/gb-2010-11-12-220.

675