bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 # Toward standards in clinical microbiome studies: comparison of three DNA extraction methods and two bioinformatic pipelines
2

3 **Q.R. Ducarmon[a, b], B.V.H. Hornung[a,b], A.R. Geelen[a,b], E.J. Kuijper[a,b], R.D. Zwittink[a,b]**

4 [a]Center for Microbiome Analyses and Therapeutics, Leiden University Medical Center,
5 Leiden, The Netherlands

6 [b]Department of Medical Microbiology, Leiden University Medical Center, Leiden, The
7 Netherlands

8 Running Head: DNA isolation and informatics for microbiome studies

9 #Address correspondence to Quinten R. Ducarmon, q.r.ducarmon@lumc.nl

10

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11 **ABSTRACT**

12 When studying the microbiome using next generation sequencing, DNA extraction method,
13 sequencing procedures and bioinformatic processing are crucial to obtain reliable data.
14 Method choice has been demonstrated to strongly affect the final biological interpretation.
15 We assessed the performance of three DNA extraction methods and two bioinformatic
16 pipelines for bacterial microbiota profiling through 16S rRNA gene amplicon sequencing,
17 using positive and negative controls for DNA extraction and sequencing, and eight different
18 types of high- or low-biomass samples. Performance was evaluated based on quality control
19 passing, DNA yield, richness, diversity and compositional profiles. All DNA extraction
20 methods retrieved the theoretical relative bacterial abundance with maximum three-fold
21 change, although differences were seen between methods, and library preparation and
22 sequencing induced little variation. Bioinformatic pipelines showed different results for
23 estimating richness, but diversity and compositional profiles were comparable. DNA
24 extraction methods were successful for feces and oral swabs and variation induced by DNA
25 extraction methods was lower than inter-subject (biological) variation. For low-biomass
26 samples, a mixture of genera present in negative controls and sample-specific genera,
27 possibly representing biological signal, were observed. We conclude that the tested
28 bioinformatic pipelines perform equally with pipeline-specific advantages and disadvantages.
29 Two out of three extraction methods performed equally well, while one method was less
30 accurate regarding retrieval of compositional profiles. Lastly, we demonstrate the importance
31 of including negative controls when analyzing low bacterial biomass samples.

32 **IMPORTANCE**

33 Method choice throughout the workflow of a microbiome study, from sample collection to
34 DNA extraction and sequencing procedures, can greatly affect results. This study evaluated

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35 three different DNA extraction methods and two bioinformatic pipelines by including
36 positive and negative controls, and various biological specimens. By identifying an optimal
37 combination of DNA extraction method and bioinformatic pipeline use, we hope to
38 contribute to increased methodological consistency in microbiome studies. Our methods were
39 not only applied to commonly studied samples for microbiota analysis, e.g. feces, but also for
40 more rarely studied, low-biomass samples. Microbiota composition profiles of low-biomass
41 samples (e.g. urine and tumor biopsies) were not always distinguishable from negative
42 controls, or showed partial overlap, confirming the importance of including negative controls
43 in microbiome studies, especially when low bacterial biomass is expected.

44 **KEYWORDS**: microbiome, DNA extraction, positive controls, negative controls,
45 bioinformatics, 16S rRNA gene amplicon sequencing

46 **INTRODUCTION**

47 Humans constantly interact with microbes that are present in the environment and reside on
48 or within the human body. Recently, the attention for microbes has shifted from an exclusive
49 interest in the pathogenicity of specific microbes toward the potential beneficial role of the
50 microbiota in human health (1). The gastrointestinal tract contains the highest number of
51 microbes and has been the most extensively studied body site of all human microbial
52 communities (2). However, many other body sites are inhabited by various microbes
53 composing a specific microbiome, such as the oral region, skin and urogenital system.
54 Microbial complexity varies between these niches, e.g. the healthy vaginal microbiota is
55 mainly composed of a few *Lactobacillus* strains (3), while gut and skin microbiota are more
56 diverse (3).

57 A limiting factor in current microbiome research is that comparison of various study results is
58 often difficult due to the application of different methodologies and lack of appropriate

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

59 controls. These differences can affect data outcomes and lead to variation as large as
60 biological differences (4). Variation can be introduced throughout the entire workflow, from
61 sample collection, storage and processing to data analysis (5-8). Recently, more attention has
62 been devoted to standardizing the workflow of microbiome research. For instance, it was
63 observed that DNA extraction has a large impact on obtained data (4, 9) and consensus has
64 been achieved regarding the application of bead-beating to increase efficiency of cell wall
65 lysis and thereby improve the yield of Gram-positive bacterial DNA (10). Nevertheless,
66 various kits and in-house extraction methods are used across different laboratories. Recently,
67 Costea *et al.* evaluated 21 DNA extraction methods across three continents and suggested one
68 protocol, named protocol Q, as ‘golden standard’ for human fecal samples. (9). They stated
69 that it was unknown whether this method is optimal for other samples than fecal material, e.g.
70 for low-biomass samples. To evaluate performance of DNA extraction for low-biomass
71 samples, it is crucial to include multiple negative controls to allow for identification of
72 bacterial DNA introduced during the entire workflow, from sample collection to sequencing
73 (11).

74 As part of optimizing the procedures for 16S rRNA gene amplicon sequencing-based
75 microbiome studies in our facility, we evaluated three DNA extraction methods and two
76 bioinformatic pipelines using various positive controls and negative controls. In addition, we
77 applied these DNA extraction methods to various biological specimens.

78 **MATERIALS AND METHODS**

79 **Sample collection and pre-processing**

80 Eight different biological specimens were included in this study, namely feces, urine, saliva,
81 oral swabs, colorectal cancer tissue, colorectal cancer supernatant, vulvar squamous cell
82 carcinoma tissue and formalin-fixed vulvar squamous cell carcinoma. Of each biological

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

83 specimen, three unique samples were included. Only for oral swabs, six unique samples were
84 included (Table S1). These samples were anonymized and treated according to the medical
85 ethical guidelines described in the Code of Conduct for Proper Secondary Use of Human
86 Tissue of the Dutch Federation of Biomedical Scientific Societies. A detailed overview of
87 sample types, sample processing and storage conditions can be found in Table S1.

88 **Mock communities and DNA standard**

89 Two mock communities (ZymoBiomics Microbial Community Standard, Zymo Research,
90 Irvine, California, USA and 20 Strain Even Mix Whole Cell Material ATCC® MSA2002™,
91 ATCC, Wesel, Germany) were included as positive controls for DNA extraction. Exact
92 composition and relative abundances of 16S copies was provided on the product sheet for
93 ZymoBiomics Microbial Community standard (hereafter referred to as Zymo mock), while
94 for ATCC® MSA2002™ (hereafter referred to as ATCC mock) we calculated expected 16S
95 profiles based on genomic information (Table S2). ZymoBiomics Microbial Community
96 DNA Standard (hereafter referred to as DNA standard) was taken along as a positive
97 sequencing control.

98 **DNA extraction**

99 **Procedures**

100 Cancer samples were pre-processed for DNA extraction comparably to a recent study on
101 pancreatic cancer microbiota (12), urine samples according to a recent publication on how to
102 study urinary microbiota (13) and other samples according to in-house methods for sample
103 processing (Table S1). For solid cancer samples, the beating steps during pre-processing were
104 performed using a Qiagen TissueLyser LT (Qiagen Benelux, Venlo, the Netherlands) at 50Hz
105 for one minute (Table S1). As single saliva samples did not contain sufficient volume for

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

106 multiple extractions, several samples of the same individual were pooled to obtain the
107 appropriate volume. DNA was extracted in duplicate from three unique samples for each
108 biological material, only for oral swabs from six unique samples, and from the two mock
109 communities. DNA was extracted using three different extraction protocols (see Protocols
110 section), and for each protocol a negative extraction (no sample) was included in duplicate.
111 The DNA standard was taken along in duplicate. DNA was quantified using a Qubit 3.0
112 Fluorometer (Invitrogen, Breda, the Netherlands) and the Qubit™ dsDNA HS Assay Kit
113 (Thermo Fisher, Landsmeer, the Netherlands). A schematic overview of the study setup is
114 shown in Figure 1.

115 **DNA extraction protocols**

116 Detailed protocols, including all minor adaptations, are present in Supplementary Methods.
117 DNA extraction was performed using three methods: 1) the Quick-DNA Fecal/Soil Microbe
118 kit (hereafter referred to as Zymo) (Zymo Research) according to manufacturer instructions
119 with minor adaptations, 2) protocol Q (hereafter referred to as Q) (9) and 3) automated DNA
120 extraction with MagNA Pure 96 ™ (hereafter referred to as Magna) (Roche Diagnostics,
121 Almere, the Netherlands) using the MagNA Pure 96 DNA and viral NA small volume kit
122 (Roche Diagnostics), according to standard operating procedures with minor adaptations.
123 Mock communities were diluted to 10^4-10^5 cells per sample for extraction using Magna. For
124 Q, several buffers and other materials were not provided in the kit and therefore purchased
125 elsewhere, namely BeadBug™ prefilled tubes with 2.0 mL capacity and 0.1 mm Zirconium
126 beads (Sigma-Aldrich, Zwijndrecht, the Netherlands), RNase A, DNase and protease-free
127 water (10 mg/mL) (Thermo Fisher, the Netherlands) and TE buffer (Thermo Fisher).

128 **MALDI-TOF Mass Spectrometry (Biotyper)**

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

129 To verify whether all bacteria of the ATCC mock were lysed after the first mechanical lysis
130 step of both Zymo and Q, the lysate was plated on Blood Agar Plate, 5% Sheep Blood in
131 Tryptic Soy Agar (VWR International, Amsterdam, the Netherlands) and aerobically and
132 anaerobically incubated at 37°C for five days. The MALDI Biotyper system was used
133 (Bruker Daltonics, Germany) to identify the bacterial species. Samples were prepared in the
134 following way: A bacterial colony was taken from the culturing plate and spread in duplicate
135 on single spots on a Bruker polished steel targetplate. Subsequently, one µl of 70% formic
136 acid was added on each single spot and when dried, one µl prepared Bruker Matrix HCCA
137 according to clinical laboratory protocols was added per spot. The Bruker polished steel
138 targetplate was then used for MALDI-TOF MS Biotyper analysis.

139 **Library preparation and 16S rRNA gene amplicon sequencing**

140 Of each duplicate DNA extraction from biological specimens, the duplicate with highest
141 genomic DNA concentration was used for sequencing. Duplicate samples from controls were
142 both sequenced. Quality control, library preparation and sequencing were performed by
143 GenomeScan B.V. (Leiden, The Netherlands) using the NEXTflex™ 16S V4 Amplicon-Seq
144 Kit (BiooScientific, TX, USA) and Illumina NextSeq 500 (paired-end, 150bp) according to
145 their standard operating procedures. QC passing was based on intact genomic DNA and DNA
146 concentrations measured by GenomeScan B.V. Therefore, those DNA concentrations were
147 used for downstream analysis. Several samples were sequenced on multiple lanes, which is
148 indicated in all relevant figures and tables.

149 **Sequencing data analysis**

150 Read filtering, operational taxonomic unit (OTU)-picking and taxonomic assignment were
151 performed using two different bioinformatic pipelines, QIIME 2 and NG-Tax 0.4 (14, 15),
152 both using the Silva_132_SSU Ref database for taxonomic classification (16). The following

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

153 settings were applied for QIIME 2: forward and reverse read length of 120, quality control
154 using Deblur, identity level of 100%. A read length of 120 was chosen due to low quality
155 sequence regions at the end of the reads. The following settings were applied for NG-Tax:
156 forward and reverse read length of 120, ratio OTU abundance of 2.0, classify ratio of 0.9,
157 minimum threshold of 1*10-7, identity level of 100%, error correction of 98.5. Prior to the
158 NG-Tax run, potential left over primers were removed with cutadapt v. 1.9.1 (17), in paired-
159 end mode, with additional setting -e 0.2 (increased error tolerance, 20%). This setting was
160 required since database truncating based on the applied primers is part of the pipeline and, as
161 such, primer sequences need to be removed to avoid mismatching with the database.
162 Furthermore, all sequences with any deviating barcode in the fastq header were changed to
163 the original barcode to allow inclusion into the NG-Tax pipeline.

164 The obtained OTU-tables were filtered for OTUs with a number of sequences less than
165 0.005% of the total number of sequences (18). Downstream analysis was performed in R
166 (v3.5.1), mainly using the phyloseq (v.1.24.2) microbiome (v.1.2.1) and ggplot2 (v.3.0.0)
167 packages (19-21).

168 **Data accessibility**

169 All raw sequencing data used in the current study are deposited in the European Nucleotide
170 Archive with accession number PRJEB34118.

171 **RESULTS AND DISCUSSION**

172 **Mock communities pass quality control**

173 We evaluated three different DNA extraction methods and two bioinformatic pipelines for
174 microbiota profiling through 16S rRNA gene amplicon sequencing (Fig 1) using several
175 positive and negative controls. Included positive controls were two bacterial mock
176 communities and one DNA standard. Included negative controls were DNA extraction

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

177 controls and sequencing controls. Quality control (QC) passing (DNA concentration and
178 intact genomic fragment) were evaluated to determine extraction method performance. It was
179 expected that positive controls would pass QC, while negative controls would not. Regarding
180 mock communities, all extractions using Zymo and Q passed QC, while for Magna one
181 extraction did not pass QC for both the ATCC mock community and Zymo mock community
182 (Table S3). This was not unexpected, as mock communities were diluted for extraction using
183 Magna and, therefore, DNA concentrations were lower. Negative extraction controls did not
184 pass QC for Q and Magna, but they did for Zymo. This likely represents a higher
185 contamination load during the extraction process for Zymo, which was also reflected by
186 higher DNA concentrations (Table S3). A full overview of all samples included in this study,
187 their QC passing and DNA concentrations can be found in Table S4.
188
189 **Positive controls: Classification, richness, diversity and relative species abundance**
190 *Primer choice may limit correct classification of all bacterial species in mock communities*
191 Performance of the three extraction methods in combination with two bioinformatic
192 pipelines, NG-Tax and QIIME 2, was evaluated on correctly identifying richness, diversity
193 and relative abundances from bacterial mock communities and a DNA standard. Richness
194 and diversity were computed at OTU level and at genus level. Analysis of compositional
195 profiles was performed at genus level. Both pipelines failed to classify one organism from
196 either mock community; NG-Tax did not detect *Cutibacterium* from the ATCC mock, while
197 QIIME 2 did not detect *Salmonella* from the Zymo mock. The inability to detect
198 *Cutibacterium* is most likely a primer choice issue, since the universal 515F and 806R
199 primers are known to poorly amplify *Cutibacterium acnes* (22). This could be solved by
200 choosing primers targeting different 16S regions, or by using adapted V4 region primers
201 which do allow for accurate amplification of *Cutibacterium* (22, 23). Regarding QIIME 2 and

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

202 the inability to detect *Salmonella*, there was an *Enterobacteriaceae* family with
203 approximately expected relative abundance for *Salmonella*, and we were therefore confident
204 this represented *Salmonella*. This *Enterobacteriaceae* family was subsequently included as
205 *Salmonella*, and designated as *Enterobacteriaceae (Salmonella)*. This classification error
206 likely resulted from the fact that *Enterobacteriaceae* members cannot always be
207 discriminated based on the 16S rRNA V4 region (24).
208
209 *DNA standard and Zymo mock community data can be recovered independent of extraction*
210 *protocol or pipeline*
211 The Zymo mock and DNA standard consist of respectively cell material or DNA of eight
212 bacterial species and two fungal species. As the 16S rRNA gene was targeted, fungi should
213 not be detected. Therefore, theoretical richness is eight and theoretical Shannon diversity was
214 calculated to be 2.01.
215 Regarding the DNA standard, NG-Tax overestimated OTU-based estimated richness for both
216 duplicates, DNA 1 and DNA 2 (Fig 2A, table S3). Richness was however accurately retrieved
217 at genus level (Fig 2C). The same was observed regarding diversity, which was
218 overestimated at OTU level (Fig 2B), but accurate at genus level (Fig 2D). QIIME 2
219 approached theoretical richness and diversity values at OTU level (Fig 2A+B, table S3).
220 Richness estimates slightly improved at genus level (Fig 2C), while diversity did not differ
221 from OTU-based diversity (Fig 2D). Thus, QIIME 2 better estimated richness and diversity at
222 OTU level, while NG-Tax performed better at genus level (Table S3).
223 Compositional profiles of DNA 1 and DNA 2 are highly similar to theoretical abundance (Fig
224 3). To quantify differences in compositional profiles, Bray-Curtis dissimilarity and Kullback-
225 Leibler divergence (Fig 4) (25) and fold errors for each taxon (Fig 5) were determined. For
226 the dissimilarity and divergence values, a value of zero represents an identical microbiota

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

227 composition to the theoretical expectation. NG-Tax obtained values closer to zero than
228 QIIME 2 for both DNA 1 and DNA 2, although the difference is minimal (Fig 4 and Table
229 S2) and the performance of both pipelines can therefore be regarded as equal. A similar
230 conclusion can be drawn from the fold errors (Fig 5), since both pipelines accurately
231 retrieved expected relative abundance, with all genera having a fold error between -1.5 and
232 1.5 (Table S3).
233 Similar analyses were performed for the Zymo mock to evaluate performance of DNA
234 extraction methods in combination with the bioinformatic pipelines. All DNA extraction
235 methods, independent of pipeline, resulted in OTU-based richness above 20 for most
236 samples, far higher than theoretical expectance (Fig 2A). This is especially noteworthy for
237 QIIME 2, as it was highly accurate in retrieving correct richness for the DNA standard, in
238 contrast to NG-Tax. Zymo and Q protocols in combination with NG-Tax retrieved accurate
239 genus level-based richness, while a slightly inflated richness was observed for Magna (Fig
240 2C). No extraction method was consistent in retrieving correct genus level-based richness in
241 combination with QIIME 2. Regarding diversity, all DNA extractions, independent of
242 pipeline, retrieved highly accurate values at genus level (Table S3). At OTU-level, however,
243 the NG-Tax pipeline resulted in overestimation of diversity independent of DNA extraction
244 method, and can therefore be considered a result of bioinformatic processing. Magna
245 extraction resulted in Bray-Curtis and Kullback-Leibler values closer to zero than Zymo and
246 Q, independent of pipeline (Fig 4 and Table S3). A similar conclusion can be drawn from the
247 fold errors, which are lowest for Magna and pipeline-independent (Fig 5 and Table S3).
248 Taken together, results obtained from the DNA standard indicate that QIIME 2 and NG-Tax
249 perform equally well in general, except for overestimation of OTU-level richness and
250 diversity when using NG-Tax. Results obtained from the Zymo mock, which is a better
251 representation of the full procedure for a microbiome study, indicate that richness is most

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

252 accurate at genus level using protocol Zymo or Q in combination with the NG-Tax pipeline.
253 In addition, bacterial microbiota composition profiles are best retrieved using Magna,
254 followed by Zymo, and are pipeline-independent.
255 In concordance with current literature (9) and independent of extraction method, a general
256 underestimation of Gram-positive bacteria was observed, with *Enterococcus* being the sole
257 exception (Fig 5). This is most likely due to incomplete cell wall lysis of Gram-positive
258 bacteria. Based on the DNA standard and the Zymo mock, we conclude that Zymo and
259 Magna in combination with either pipeline are the best performing combinations (Table S3).
260 However, when high-throughput DNA extraction is required (e.g. for large cohort studies),
261 Magna may be preferred from a practical point of view, although it overestimates richness
262 independent of pipeline.
263 In general, overestimation of OTUs may stem from the 100% identity setting for clustering,
264 combined with the natural divergence of the 16S gene (26, 27). There is no current consensus
265 on OTU identity setting, and cut-offs between 97% and 100% are used. An advantage of the
266 100% cut-off is that unique taxa differing a single nucleotide are clustered into different
267 OTUs. A disadvantage is that, as intragenomic diversity in the 16S rRNA gene is common
268 within bacterial genomes, a 100% cut-off can lead to multiple OTUs stemming from a single
269 bacterium and thereby inflate richness (27). Apart from this biological explanation, the
270 different algorithms and internal filtering steps used in QIIME 2 and NG-Tax can affect the
271 outcome for richness.
272
273 *ATCC mock is recovered incorrectly, independent of extraction protocol or pipeline*
274 The ATCC mock consists of 20 unique bacterial species, with four of them belonging to two
275 genera (*Staphylococcus* and *Streptococcus*). Therefore, theoretical richness at OTU level

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

276 would be 20, but eighteen at genus level. In addition, these 20 unique bacterial species come
277 from different environments, including gut, oral and skin microbiome.

278 No values close to the theoretical profiles for the ATCC mock for any extraction
279 method/bioinformatic pipeline were observed, and one sample from Q consisted almost
280 entirely of non-classifiable reads (Fig 6), indicating sample-related issues. *Bacillus* was
281 highly overrepresented in all other samples, with a relative abundance over 30% in Zymo and
282 Magna extracted samples, while 6.13% is expected. Curiously, after the first mechanical lysis
283 step in Q, we could culture *Bacillus cereus* and *Cutibacterium acnes* (identification scores of
284 1.90 and 2.00, respectively), and *Bacillus cereus* (identification score 2.05) after mechanical
285 lysis in Zymo. This is clinically important, as it means that infectious materials cannot be
286 considered safe or non-infectious after mechanical lysis. As culturing of *B. cereus* indicates
287 that cell wall lysis was incomplete, it would be expected that its relative abundance was
288 underestimated, contrarily to what was observed. Another research group recently reported a
289 similar overrepresentation of *Bacillus* in the ATCC community (28). ATCC itself was also
290 unable to retrieve abundances close to theoretical expectation, neither with 16S amplicon
291 sequencing nor with shotgun sequencing (29). Several reasons could explain this discrepancy
292 between theoretical profiles and obtained profiles. For example, physical cell-to-cell
293 interactions or presence of different metabolites may interfere with DNA extraction (26, 30).
294 Therefore, based on this synthetic community, no conclusions on the optimal extraction-
295 pipeline combination could be made. This proposed positive control prompts the question
296 whether mock communities are always reliable for assessing performance of DNA extraction
297 methods. As can be observed from the Zymo mock, DNA extraction kits do not necessarily
298 inflict observed deviations, but may rather be a result of mock community-specific properties.
299 Outcomes may depend on extraction kit / community type combination, indicating the

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

300 potential necessity to use a positive control that strongly resembles the investigated
301 microbiome.

302 **Negative controls: Contaminating sequences are not always consistent**

303 Negative controls were taken along for each extraction method to check for kit-specific
304 contaminants, which is especially relevant for deciding whether low-biomass samples contain
305 real microbiota. Regarding Zymo, clear kit-contaminants were *Pseudomonas* and *Delftia* (Fig
306 S2A+C), consistent across the different pipelines at genus level, and with previous findings
307 (11, 31). For Magna and Q, specific contaminants were less obvious, although *Pseudomonas*
308 was present. Generally, negative controls mostly consisted of genera commonly found in gut
309 and oral microbiota, most of them also previously described as contaminants (11). In
310 addition, negative sequencing controls were taken along, and here no consistent contaminants
311 could be observed (Fig S2B+D). Potential contamination sources are multifold, such as kit
312 contamination, index hopping, or well-to-well contamination (32, 33). Index-hopping is
313 however not a likely source of contamination, as the negative control for Magna was
314 sequenced in different lanes, and profiles look highly similar (Fig S2A+C). Additionally, we
315 did not observe index-hopping in our positive controls.

316 One of the contaminants we identified has not been previously described as a contaminant,
317 namely *Clostridioides*. This likely represents *C. difficile*, and contamination by this bacterium
318 can be explained by the fact that DNA extractions were performed in our National Reference
319 Laboratory for *C. difficile*, which probably contains minor amounts of *C. difficile* spores
320 during most time points. *C. difficile* contamination on laboratory surfaces has also recently
321 been described in another clinical microbiology laboratory (34).

322 By incorporating this information with the Zymo positive controls, it can be concluded that
323 Zymo and Magna are most optimal. Magna most accurately captured the expected

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

324 community profile, while kit-specific contaminants are clear and easy to discriminate from
325 biological signal using Zymo (Table S2). When investigating different biological sample
326 types it might be warranted to use a kit for which kit contaminants do not overlap with the
327 biological signal, e.g. *Pseudomonas* contamination when studying sputum samples from
328 cystic fibrosis patients who are frequently colonized with *Pseudomonas* spp.

329 **Automatic Magna extraction yields lowest DNA for biological samples**

330 Twenty-seven biological samples were available per extraction protocol (Table S1) and Q
331 was most successful in passing QC (22/27), followed by Zymo (20/27) and Magna (17/27)
332 (Table S3). DNA concentrations were on average lowest for Magna, while yields were
333 comparable between Q and Zymo (Figure S1). Processing of raw sequencing data from
334 biological samples was performed using the NG-Tax pipeline at genus level.
335

336 **The fecal microbiome can be sufficiently investigated independent of method**

337 DNA extracted from fecal samples using the three different protocols all passed QC. Magna,
338 Zymo and Q achieved an average concentration of approximately 29 ng/µl, 111 ng/µl and
339 212 ng/µl, respectively (Fig. S1). While DNA yield varied between extraction methods, all
340 were sufficient for sequencing. Microbiota profiles were comparable between extraction
341 methods for each sample (Figure S3A). In addition, differences in compositional profiles
342 were quantified using Kullback-Leibler divergence (Figure 7A). This heatmap shows that
343 technical variation induced by DNA extraction method is much lower than biological
344 variation between feces samples. Profiles of the feces donors contained many bacterial genera
345 commonly present in fecal microbiomes (35, 36). Healthy fecal microbiomes largely consist
346 of Bacteroidetes and Firmicutes phyla (~90%), while Actinobacteria and Proteobacteria are
347 present in smaller proportions. At genus level, *Bacteroides*, *Prevotella* and *Faecalibacterium*
348 are among the most prevalent genera, all of which were found in high abundance herein (3).

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

349

350 **Low DNA yield of oral swabs does not seem to impact the microbiota profile**

351 Out of eighteen DNA extractions, fifteen extractions passed QC for oral swabs. Only for
352 Zymo, all extractions passed QC. DNA yields were highly variable for all extraction
353 methods, ranging from 0.12 to 6.34 ng/µl. Half of the extractions (nine/eighteen) yielded a
354 concentration below one ng/µl. All compositional profiles were dominated by *Streptococcus*,
355 *Prevotella spp.*, *Haemophilus* and *Veillonella*, which was individual-independent. In addition,
356 technical variation induced by DNA extraction and subsequent steps was lower than
357 biological variation (Fig 7B). The oral microbiome, like the gut microbiome, is highly
358 diverse. Nevertheless, a certain core of genera (e.g. *Streptococcus spp.* and *Prevotella spp.)* is
359 present in most people, all of which were found in our study (3, 37, 38). Together, the good
360 QC passing rate, DNA concentrations and consistency of compositional profiles between
361 extraction methods lead us to conclude that all three methods work well for oral swabs.

362

363 **Applied methodology renders the urine microbiome unresolved.**

364 During the last decade, microbiome studies showed that urine contains a bacterial microbiota
365 (39, 40). Despite using 30-40 ml of urine and centrifugation prior to extraction (13), we were
366 not able to convincingly capture a urinary microbiota for all samples (Fig S3C). DNA
367 concentrations were high for an infected sample (between thirteen and 42 ng/µl), but
368 concentrations for the other samples were between 0.11 and 0.99 ng/µl. Six out of nine
369 samples passed QC. For the infected sample with a high bacterial load, we were able to
370 classify the cause of infection to *Enterobacteriaceae*, which is in agreement with the fact that
371 most UTIs are caused by members of *Enterobacteriaceae*. One urine sample showed high
372 similarity to negative controls for respective kits, with non-classifiable reads for Q and
373 Magna, and high abundance of *Pseudomonas* for Zymo (Fig S3C). Another urine sample

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

374 contained a high *Lactobacillus* abundance, which has previously been shown to be abundant
375 in urine samples (40). In addition, presence of *Atopobium*, *Gardnerella*, *Campylobacter*,
376 *Prevotella* and *Anaerococcus* point towards an existing urinary microbiota (41). However,
377 *Pseudomonas,* a common Zymo kit contaminant, was still found in this urine sample, and for
378 Magna more than 25% of reads could not be classified (Fig S3C). This could indicate that the
379 biological signal is not much stronger than contamination, and therefore a mixed profile is
380 observed. Further efforts and method optimization should be undertaken, although this can be
381 difficult to implement in routine work (42). In addition, culturing could be used as a follow-
382 up method to confirm that contaminants are not viable bacteria, but rather bacterial DNA.
383
384 **Applied procedures for saliva handling seem to be unsuitable for microbiome research**
385 DNA yield from saliva samples was lower as compared to literature (43, 44) (Fig S1). Only a
386 single DNA extraction had a concentration of slightly above one ng/µl (1.18; Table S4),
387 while all other extractions had concentrations between 0.04 and 0.68 ng/µl. This may be
388 associated with storage duration (~fifteen years) and the fact that samples were thawed and
389 refrozen several times. This also explains why only three out of nine DNA extracts passed
390 QC. The included saliva samples were chosen as investigators within our facility were
391 interested to see if microbiota studies could be performed using these samples.
392 Compositional profiles consisted of a mixture of genera present in the normal oral microbiota
393 (*Oribacterium*, two *Prevotella* genera, *Streptococcus*, *Veillonella*) (3), genera present in our
394 negative controls (*Pseudomonas*, *Delftia*) and non-classifiable reads (Fig S3D). In
395 combination with low DNA yields, it is likely that a mixture between biological signal and
396 contamination signal is present. Therefore, we consider the applied extraction methods
397 unsuitable for saliva samples with a long duration of storage time and multiple freezing-
398 thawing cycles.

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

399

400 **The colorectal cancer microbiome cannot be distinguished from negative controls or**
401 **fecal microbiome**

402 As colorectal cancer development has been associated with specific gut bacteria, we were
403 interested to see if colorectal cancer tissue itself also contained bacteria (45, 46). DNA
404 concentrations were sufficient for all samples to pass QC, but extracted DNA was likely
405 mostly human-derived. Two of three extraction methods were not successful, as samples
406 extracted using Zymo and Magna showed high similarity to their respective negative controls
407 (Fig S3E). Using Q, *Bacteroides*, *Fusobacterium* and *Gemella* were identified, all being
408 previously associated with colorectal cancer development (45, 47). Several gut commensals,
409 including *Faecalibacterium* and *Escherichia-Shigella* were present in both the negative
410 controls and these colorectal cancer samples. It is therefore difficult to discriminate whether
411 these are contaminant bacteria, or whether they represent biological signal.

412 We hypothesized that by spinning down the material, the supernatant would contain more
413 bacteria than the cancer tissue. DNA concentrations of supernatant were between 0.16 and
414 2.32 ng/µl, and seven out of nine DNA extractions passed QC (Table S4). For one sample, it
415 was clear that across all methods many genera were observed which were present in negative
416 controls (e.g. *Pseudomonas*), or reads could not be classified at all (Fig S3F). A second
417 sample seemed to contain a real microbiota. Profiles were consistent across extraction
418 methods, did not contain many contaminants and had specific bacteria previously linked to
419 colorectal cancer (e.g. *Fusobacterium*) (45). The third sample showed a profile reflecting a
420 mix between biological signal and technical contamination. Profiles were consistent across
421 methods and contained genera representative of a gut microbiome, but also contained non-
422 classifiable reads and contamination. Therefore, profiles are likely a mixture of biological
423 signal and technical contamination, and further optimization is necessary prior to using this

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

424 sample type for experimental studies. We have the same recommendation for colorectal
425 cancer sample types as for urine, as discussed above.
426
427 **Vulvar squamous cell carcinoma does probably not contain bacterial DNA**
428 Vulvar squamous cell carcinoma (VSCC) has different etiological pathways, of which one is
429 associated with human papilloma virus (HPV). The counterpart is non-virally related and is
430 frequently associated to lichen sclerosis, a benign chronic inflammatory lesion and *TP53*
431 mutations (48, 49). We extracted DNA from HPV-negative VSCC tissue as a pilot study to
432 determine if investigating the relationship between bacterial microbiota and HPV-negative
433 VSCC would be potentially feasible. DNA concentrations were high (Fig S1), only for three
434 extractions below one ng/µl, and eight out of nine extractions passed QC. However, DNA
435 was probably again largely human-derived. This was reflected in the obtained microbiota
436 profiles, as most reads were not classified or the profiles showed high similarity to negative
437 controls (e.g. high abundance of *Pseudomonas*) (Fig S3G). Therefore, it is unlikely that this
438 cancer tissue contains bacteria, or bacteria are so lowly abundant that they are overshadowed
439 by contamination load. In general, the vulvar microbiome has not been extensively studied. A
440 recent study on vulvar microbiome observed that *Lactobacillus*, *Corynebacterium*,
441 *Finegoldia*, *Staphylococcus and Anaerococcus* are most abundant on this body site, but the
442 use of negative controls was not reported (50). These genera are also part of the vaginal
443 microbiota, and might be sampling contamination or reflect high similarity between vulvar
444 and vaginal microbiota.
445 A large amount of formalin-fixed VSCC materials are stored in a biobank at our facility. To
446 investigate whether this sample collection could be used for microbiota profiling, DNA was
447 extracted from three formalin-fixed VSCC samples. DNA concentrations were all below 0.3
448 ng/µl, and only two out of nine extractions passed QC (Fig S4). One sample extracted with Q

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

449 was excluded from further analysis, as no reads were present after sequencing. Extraction and
450 sequencing of formalin-fixed material poses additional problems, as DNA molecules could be
451 highly fragmented and too short for amplicon sequencing of the V4 region (51). For Zymo,
452 samples resembled negative controls, with *Delftia* and *Pseudomonas* being highly abundant
453 (Fig S3H). The same samples had completely different microbiota profiles when using
454 protocol Q or Magna. Both extraction methods showed genera commonly found in the lower
455 urogenital tract, including *Staphylococcus*, *Streptococcus*, *Prevotella* and *Gordonia* (3, 36).
456 However, many of these genera were also detected in negative controls. In combination with
457 low DNA yield and inconsistent profiles across extraction methods, we conclude that no
458 reliable bacterial microbiota profile could be identified in these samples. For both VSCC
459 types, we suggest the same way forward as for urine samples.
460
461 **Sample groups with and without biological signal cluster apart**
462 Lastly, we performed t-distributed stochastic neighbor embedding (t-SNE) clustering using
463 Bray-Curtis measures on all samples used in the present study (Fig 8) (52). Based on
464 microbiota composition as measured by Bray-Curtis, t-SNE projects points in a two-
465 dimensional space, while maintaining local structures present in high-dimensional space.
466 Clear clusters could be identified for Zymo positive controls, feces, oral swabs and ATCC
467 mock (all but one sample) (Fig 8). Other biological samples and negative controls were more
468 dispersed throughout the plot, indicating that either more biological or technical variation was
469 present. This is in agreement with our detailed analysis, showing that their microbiota cannot
470 necessarily be distinguished from the negative controls. An example of the importance of
471 including negative controls comes from two studies aiming to unravel the placental
472 microbiota (53, 54). It is currently unclear whether a placental microbiota exists, but when

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

473 comparing placental samples of healthy deliveries to included negative controls, microbiota
474 compositions could not be distinguished (53, 54).
475

476 **Strengths and limitations**

477 The current study had several strengths and limitations. By using a positive control of cell
478 material with a corresponding DNA standard, we differentiated variation induced from
479 sequencing procedures and DNA extraction. We demonstrate the importance of using
480 positive and negative controls in microbiome studies, and show that negative controls are
481 crucial for interpretation of low-biomass samples. Another strength of the study was that for
482 several biological samples (feces and oral swabs), we showed that technical variation was
483 much smaller than biological variation. A shortcoming of the study is that we did not perform
484 any other quantification next to 16S sequencing (e.g. qPCR), which may be particularly
485 useful for quality control of the ATCC mock. Furthermore, the current study used only three
486 unique samples of most biological sample types. Especially for samples for which DNA
487 extraction was challenging (urine samples, colorectal cancer supernatant), a higher number of
488 unique samples would have allowed for a more thorough evaluation.
489

490 **CONCLUSION**

491 The current study evaluated three DNA extraction methods and two bioinformatic pipelines
492 for bacterial microbiota profiling using several positive and negative controls, and a range of
493 biological specimens. All three extraction methods quite accurately retrieved theoretical
494 abundance of the Zymo mock, but not of the ATCC mock. For DNA extraction, we
495 recommend using the Zymo and Magna protocol, since they showed good overall
496 performance for all samples. Sequencing procedure only induced minor variation, as shown

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

497 using a DNA standard. We furthermore showed that the NG-Tax and QIIME 2 pipelines
498 perform equally well overall, each having their specific flaws.
499 By including negative controls and comparing these with low-biomass samples, we evaluated
500 whether low-biomass samples consisted of technical noise, biological signal or a mixture. In
501 most cases, identification of a unique microbiome was not achieved, highlighting the
502 importance of negative controls and sufficiently sensitive methods. The results from this
503 study can help other microbiome study groups to select an appropriate DNA extraction
504 method and bioinformatic pipeline. We hope this study contributes to further standardization
505 in methodology in the microbiome field, and to increased awareness of the usage of controls,
506 especially when studying low-biomass samples.
507

508 **ACKNOWLEDGEMENTS**

509 We thank all collaborating partners who provided us with clinical biospecimens, namely Liz
510 Terveer, Eric Berssenbrugge, Erik Giltay, Noel de Miranda, Jitske van den Bulk, Natalja ter
511 Haar and Kim Kortekaas. We also thank Eric Claas for permission for using the MagNA Pure
512 96 ™ and the clinical laboratory for identification of *Bacillus cereus* and *Cutibacterium*
513 *acnes* using MALDI-TOF.
514

515 **FUNDING**

516 This research received no specific grant from any funding agency in the public, commercial,
517 or not-for-profit sectors. BH and EK are supported by an unrestricted grant from Vedanta
518 Biosciences Inc. EK has performed research for Cubist, Novartis and Qiagen, and has
519 participated in advisory forums of Astellas, Optimer, Actelion, Pfizer, Sanofi Pasteur and
520 Seres Therapeutics. The companies had no role in writing this manuscript.
521

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

522 **AUTHOR CONTRIBUTIONS**

523 QD, BH, AG, EK and RZ conceptualized and designed the study. QD and AG performed
524 practical laboratory work. BH and RZ processed raw sequencing data. QD analyzed data,
525 prepared figures and wrote the manuscript under supervision of BH and RZ. All authors
526 interpreted data, read and revised drafts of the manuscript, and approved the final version.

527

528 **REFERENCES**

529 1. Guarner F, Malagelada JR. 2003. Gut flora in health and disease. Lancet 361:512-9.
530 2. Rajilic-Stojanovic M, Smidt H, de Vos WM. 2007. Diversity of the human gastrointestinal
531 tract microbiota revisited. Environ Microbiol 9:2125-2136.
532 3. Human Microbiome Project C. 2012. Structure, function and diversity of the healthy human
533 microbiome. Nature 486:207-14.
534 4. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S,
535 Microbiome Quality Control Project C, Abnet CC, Knight R, White O, Huttenhower C. 2017.
536 Assessment of variation in microbial community amplicon sequencing by the Microbiome
537 Quality Control (MBQC) project consortium. Nature Biotechnology 35:1077-1086.
538 5. Fouhy F, Deane J, Rea MC, O'Sullivan O, Ross RP, O'Callaghan G, Plant BJ, Stanton C. 2015.
539 The effects of freezing on faecal microbiota as determined using MiSeq sequencing and
540 culture-based investigations. PLoS One 10:e0119355.
541 6. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, Thomson JM,
542 Consortium UIG, Satsangi J, Flint HJ, Parkhill J, Lees CW, Hold GL. 2014. The impact of
543 different DNA extraction kits and laboratories upon the assessment of human gut microbiota
544 composition by 16S rRNA gene sequencing. PLoS One 9:e88982.
545 7. Walker AW, Martin JC, Scott P, Parkhill J, Flint HJ, Scott KP. 2015. 16S rRNA gene-based
546 profiling of the human infant gut microbiota is strongly influenced by sample processing and
547 PCR primer choice. Microbiome 3:26.
548 8. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. 2015. Insight into biases and
549 sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids
550 Res 43:e37.
551 9. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M,
552 Hercog R, Jung FE, Kultima JR, Hayward MR, Coelho LP, Allen-Vercoe E, Bertrand L, Blaut M,
553 Brown JRM, Carton T, Cools-Portier S, Daigneault M, Derrien M, Druesne A, de Vos WM,
554 Finlay BB, Flint HJ, Guarner F, Hattori M, Heilig H, Luna RA, van Hylckama Vlieg J, Junick J,
555 Klymiuk I, Langella P, Le Chatelier E, Mai V, Manichanh C, Martin JC, Mery C, Morita H,
556 O'Toole PW, Orvain C, Patil KR, Penders J, Persson S, Pons N, Popova M, Salonen A, Saulnier
557 D, Scott KP, Singh B, et al. 2017. Towards standards for human fecal sample processing in
558 metagenomic studies. Nature Biotechnology 35:1069-1076.
559 10. Santiago A, Panda S, Mengels G, Martinez X, Azpiroz F, Dore J, Guarner F, Manichanh C.
560 2014. Processing faecal samples: a step forward for standards in microbial community
561 analysis. BMC Microbiol 14:112.
562 11. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. 2019. Contamination in
563 Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol
564 27:105-117.

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

565 12. Pushalkar S, Hundeyin M, Daley D, Zambirinis CP, Kurz E, Mishra A, Mohan N, Aykut B, Usyk
566 M, Torres LE, Werba G, Zhang K, Guo Y, Li Q, Akkad N, Lall S, Wadowski B, Gutierrez J,
567 Kochen Rossi JA, Herzog JW, Diskin B, Torres-Hernandez A, Leinwand J, Wang W, Taunk PS,
568 Savadkar S, Janal M, Saxena A, Li X, Cohen D, Sartor RB, Saxena D, Miller G. 2018. The
569 Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive
570 Immune Suppression. Cancer Discov 8:403-416.
571 13. Karstens L, Asquith M, Caruso V, Rosenbaum JT, Fair DA, Braun J, Gregory WT, Nardos R,
572 McWeeney SK. 2018. Community profiling of the urinary microbiota: considerations for low-
573 biomass samples. Nat Rev Urol 15:735-749.
574 14. Ramiro-Garcia J, Hermes GDA, Giatsis C, Sipkema D, Zoetendal EG, Schaap PJ, Smidt H. 2016.
575 NG-Tax, a highly accurate and validated pipeline for analysis of 16S rRNA amplicons from
576 complex biomes[version 1; referees: 2 approved with reservations, 1 not approved].
577 F1000Research 5.
578 15. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ,
579 Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT,
580 Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas
581 GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson
582 DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G,
583 Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D,
584 Koester I, Kosciolek T, et al. 2018. QIIME 2: Reproducible, interactive, scalable, and
585 extensible microbiome data science. PeerJ Preprints 6:e27295v2.
586 16. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. 2013. The
587 SILVA ribosomal RNA gene database project: improved data processing and web-based
588 tools. Nucleic Acids Res 41:D590-6.
589 17. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing
590 reads. 2011 17:3.
591 18. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG.
592 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon
593 sequencing. Nat Methods 10:57-9.
594 19. McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis
595 and graphics of microbiome census data. PLoS One 8:e61217.
596 20. Lahti L, Shetty S. 2017. Tools for microbiome analysis in R. Microbiome package version
597 1.2.1,
598 21. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business
599 Media.
600 22. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R. 2011.
601 Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 13:47-
602 58.
603 23. Zeeuwen P, Boekhorst J, Ederveen THA, Kleerebezem M, Schalkwijk J, van Hijum S,
604 Timmerman HM. 2017. Reply to Meisel et al. J Invest Dermatol 137:961-962.
605 24. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL,
606 Madsen KL, Wong GK. 2016. Characterization of the Gut Microbiome Using 16S or Shotgun
607 Metagenomics. Front Microbiol 7:459.
608 25. Kullback S, Leibler RA. 1951. On Information and Sufficiency. The Annals of Mathematical
609 Statistics 22:79-86.
610 26. Hornung BVH, Zwittink RD, Kuijper EJ. 2019. Issues and current standards of controls in
611 microbiome research. FEMS Microbiol Ecol 95.
612 27. Sun DL, Jiang X, Wu QL, Zhou NY. 2013. Intragenomic heterogeneity of 16S rRNA genes
613 causes overestimation of prokaryotic diversity. Appl Environ Microbiol 79:5962-9.
614 28. Scherz V, Aeby S, Bertelli C, Greub G. 2019. Microbiota profiling of saliva: unexpected results
615 of a preliminary study, abstr ECCMID, Amsterdam, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

616 29. ATCC. 2018. Microbiome standard & research solutions. https://www.lgcstandards-
617 atcc.org/~/media/A36E4E940E3F40B1A5E14D6851E9348F.ashx. Accessed April 10th 2019.
618 30. Angelakis E, Bachar D, Henrissat B, Armougom F, Audoly G, Lagier JC, Robert C, Raoult D.
619 2016. Glycans affect DNA extraction and induce substantial differences in gut metagenomic
620 studies. Sci Rep 6:26276.
621 31. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman
622 NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-
623 based microbiome analyses. BMC Biol 12:87.
624 32. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B, Green L, Howd T,
625 Mason T, Vicente G, Dasilva M, Brodeur W, DeSmet T, Dodge S, Lennon NJ, Gabriel S. 2018.
626 Characterization and remediation of sample index swaps by non-redundant dual indexing on
627 massively parallel sequencing platforms. BMC Genomics 19:332.
628 33. Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. 2019. Quantifying and
629 Understanding Well-to-Well Contamination in Microbiome Research. mSystems 4.
630 34. Reigadas E, Vazquez-Cuesta S, Onori R, Villar-Gomara L, Alcala L, Marin M, Martin A, Munoz
631 P, Bouza E. 2019. Clostridioides difficile contamination in the environment of a clinical
632 microbiology laboratory and laboratory workers. Clin Microbiol Infect
633 doi:10.1016/j.cmi.2019.06.027.
634 35. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F,
635 Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J,
636 Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E,
637 Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S,
638 Qin N, Yang H, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J,
639 Weissenbach J, et al. 2010. A human gut microbial gene catalogue established by
640 metagenomic sequencing. Nature 464:59-65.
641 36. Lloyd-Price J, Abu-Ali G, Huttenhower C. 2016. The healthy human microbiome. Genome
642 Med 8:51.
643 37. Lamont RJ, Koo H, Hajishengallis G. 2018. The oral microbiota: dynamic communities and
644 host interactions. Nat Rev Microbiol 16:745-759.
645 38. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the normal bacterial flora of
646 the oral cavity. J Clin Microbiol 43:5721-32.
647 39. Wolfe AJ, Toh E, Shibata N, Rong R, Kenton K, Fitzgerald M, Mueller ER, Schreckenberger P,
648 Dong Q, Nelson DE, Brubaker L. 2012. Evidence of uncultivated bacteria in the adult female
649 bladder. J Clin Microbiol 50:1376-83.
650 40. Hilt EE, McKinley K, Pearce MM, Rosenfeld AB, Zilliox MJ, Mueller ER, Brubaker L, Gai X,
651 Wolfe AJ, Schreckenberger PC. 2014. Urine is not sterile: use of enhanced urine culture
652 techniques to detect resident bacterial flora in the adult female bladder. J Clin Microbiol
653 52:871-6.
654 41. Bucevic Popovic V, Situm M, Chow CT, Chan LS, Roje B, Terzic J. 2018. The urinary
655 microbiome associated with bladder cancer. Sci Rep 8:12157.
656 42. Boers SA, Hays JP, Jansen R. 2017. Novel micelle PCR-based method for accurate, sensitive
657 and quantitative microbiota profiling. Sci Rep 7:45536.
658 43. Lim Y, Totsika M, Morrison M, Punyadeera C. 2017. The saliva microbiome profiles are
659 minimally affected by collection method or DNA extraction protocols. Sci Rep 7:8523.
660 44. Vesty A, Biswas K, Taylor MW, Gear K, Douglas RG. 2017. Evaluating the Impact of DNA
661 Extraction Method on the Representation of Human Oral Bacterial and Fungal Communities.
662 PLoS One 12:e0169877.
663 45. Kostic AD, Chun E, Robertson L, Glickman JN, Gallini CA, Michaud M, Clancy TE, Chung DC,
664 Lochhead P, Hold GL, El-Omar EM, Brenner D, Fuchs CS, Meyerson M, Garrett WS. 2013.
665 Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-
666 immune microenvironment. Cell Host Microbe 14:207-15.

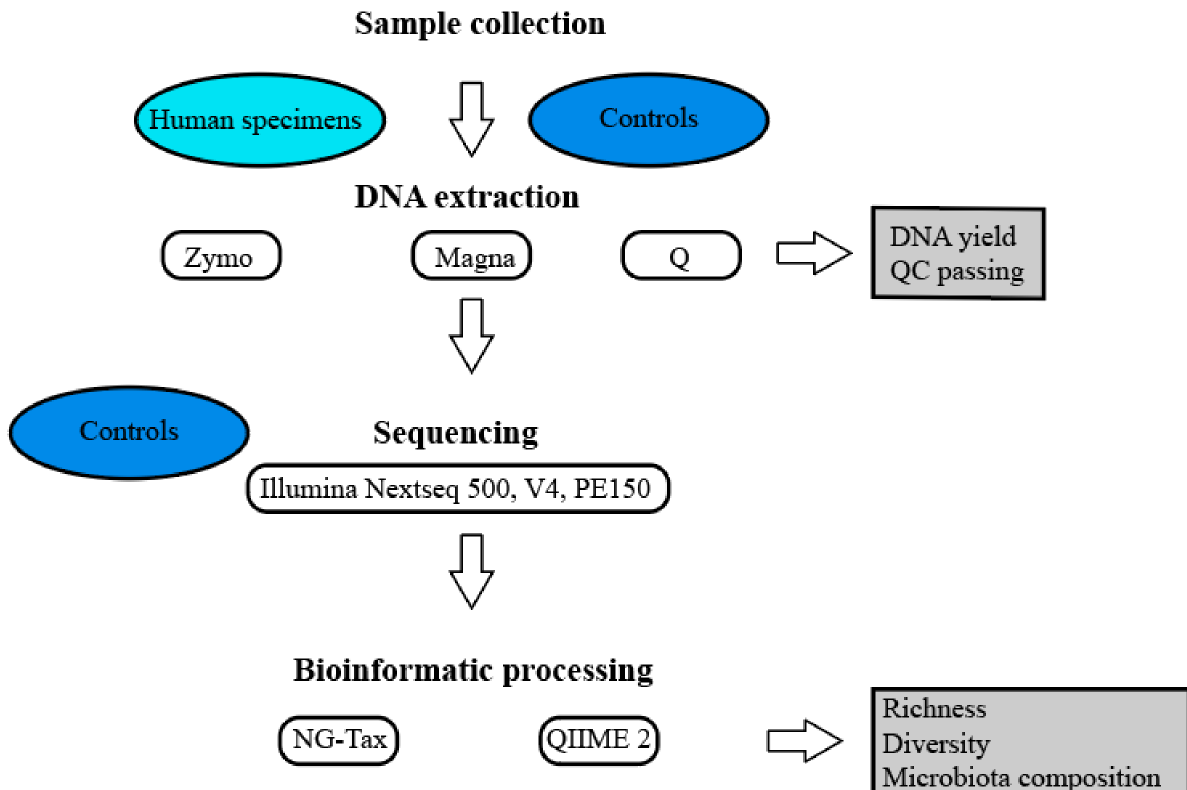
bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

667 46. He Z, Gharaibeh RZ, Newsome RC, Pope JL, Dougherty MW, Tomkovich S, Pons B, Mirey G,
668 Vignard J, Hendrixson DR, Jobin C. 2019. Campylobacter jejuni promotes colorectal
669 tumorigenesis through the action of cytolethal distending toxin. Gut 68:289-300.
670 47. Kwong TNY, Wang X, Nakatsu G, Chow TC, Tipoe T, Dai RZW, Tsoi KKK, Wong MCS, Tse G,
671 Chan MTV, Chan FKL, Ng SC, Wu JCY, Wu WKK, Yu J, Sung JJY, Wong SH. 2018. Association
672 Between Bacteremia From Specific Microbes and Subsequent Diagnosis of Colorectal Cancer.
673 Gastroenterology 155:383-390.e8.
674 48. Abdulrahman Z, Kortekaas KE, De Vos Van Steenwijk PJ, Van Der Burg SH, Van Poelgeest MI.
675 2018. The immune microenvironment in vulvar (pre)cancer: review of literature and
676 implications for immunotherapy. Expert Opin Biol Ther 18:1223-1233.
677 49. van de Nieuwenhof HP, van der Avoort IA, de Hullu JA. 2008. Review of squamous
678 premalignant vulvar lesions. Crit Rev Oncol Hematol 68:131-56.
679 50. Vongsa R, Hoffman D, Shepard K, Koenig D. 2019. Comparative study of vulva and abdominal
680 skin microbiota of healthy females with high and average BMI. BMC Microbiol 19:16.
681 51. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, Shental N. 2018. Combining 16S
682 rRNA gene variable regions enables high-resolution microbial community profiling.
683 Microbiome 6:17.
684 52. van der Maaten L, Hinton G. 2008. Visualizing Data using t-SNE. Journal of Machine Learning
685 Research 9:2579-2605.
686 53. Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, Leite R, Elovitz MA,
687 Parry S, Bushman FD. 2016. Comparison of placenta samples with contamination controls
688 does not provide evidence for a distinct placenta microbiota. Microbiome 4:29.
689 54. de Goffau MC, Lager S, Sovio U, Gaccioli F, Cook E, Peacock SJ, Parkhill J, Charnock-Jones DS,
690 Smith GCS. 2019. Human placenta has no microbiome but can contain potential pathogens.
691 Nature doi:10.1038/s41586-019-1451-5.

692

693

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

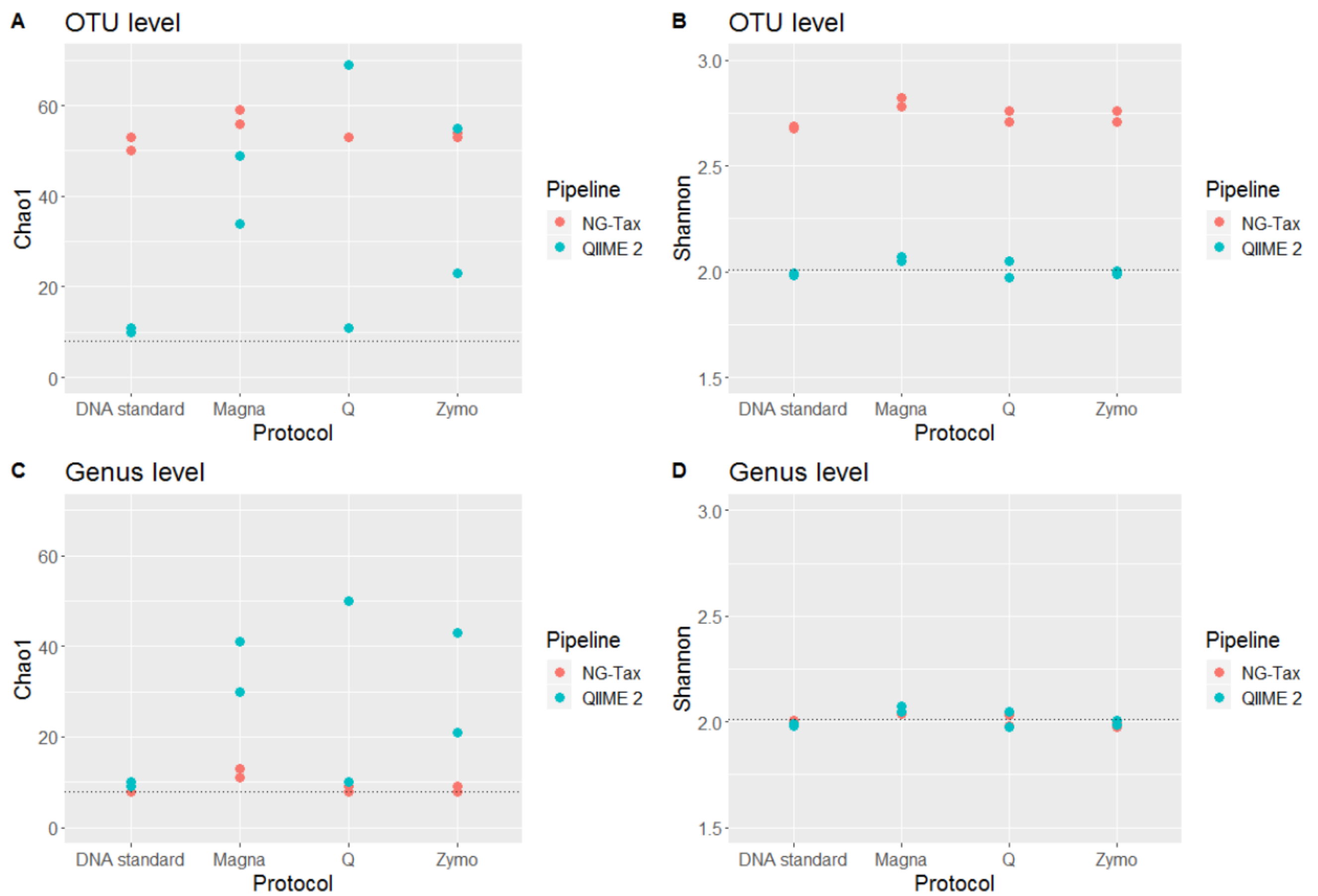


694

695

696 **Figure 1:** Study design workflow. DNA was extracted from human specimens and positive
697 and negative controls using three different DNA extraction methods. DNA extraction
698 performance was assessed on DNA yield and QC passing. Extracted DNA, and positive and
699 negative sequencing controls were sequenced. Raw sequencing data was processed using two
700 bioinformatic pipelines. Performance was assessed on microbiota composition, richness and
701 diversity.

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



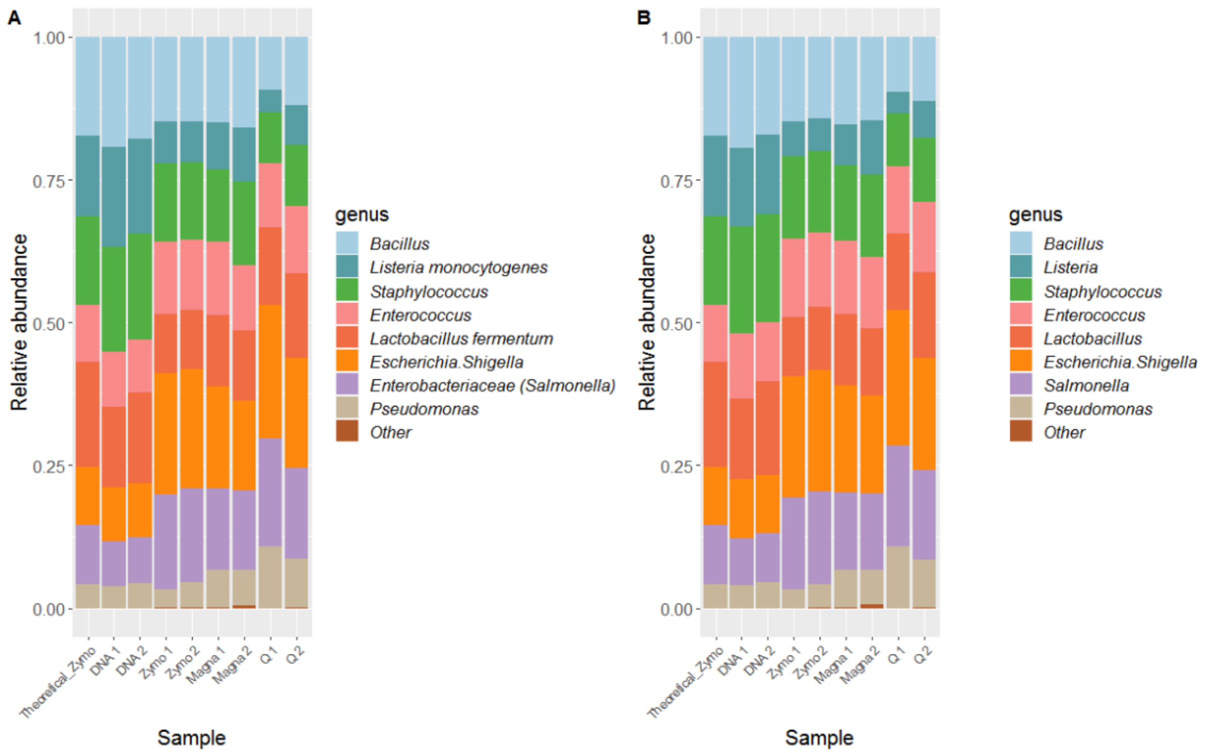
702

703 **Figure 2:** Richness (Chao1) and diversity (Shannon) computed for Zymo DNA and Zymo

704 mock at OTU level (A+B) and at genus level (C+D) for each combination of bioinformatic

705 pipeline and DNA extraction method. Dashed lines indicate theoretical values.

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



706

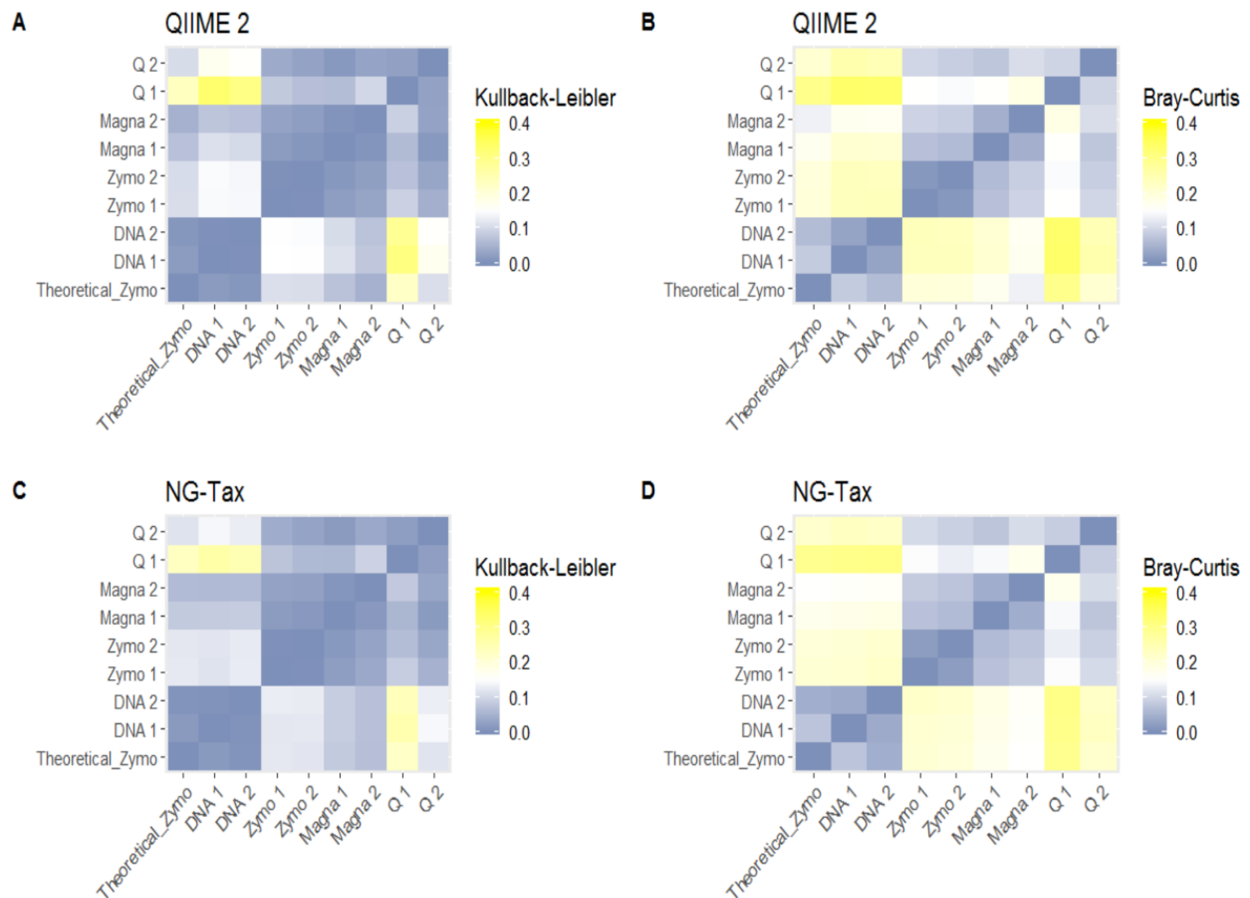
707 **Figure 3**: Compositional profiles at genus level for QIIME 2 (A) and NG-Tax (B) for Zymo

708 mock, theoretical composition is indicated in the first bar graph.

709

710

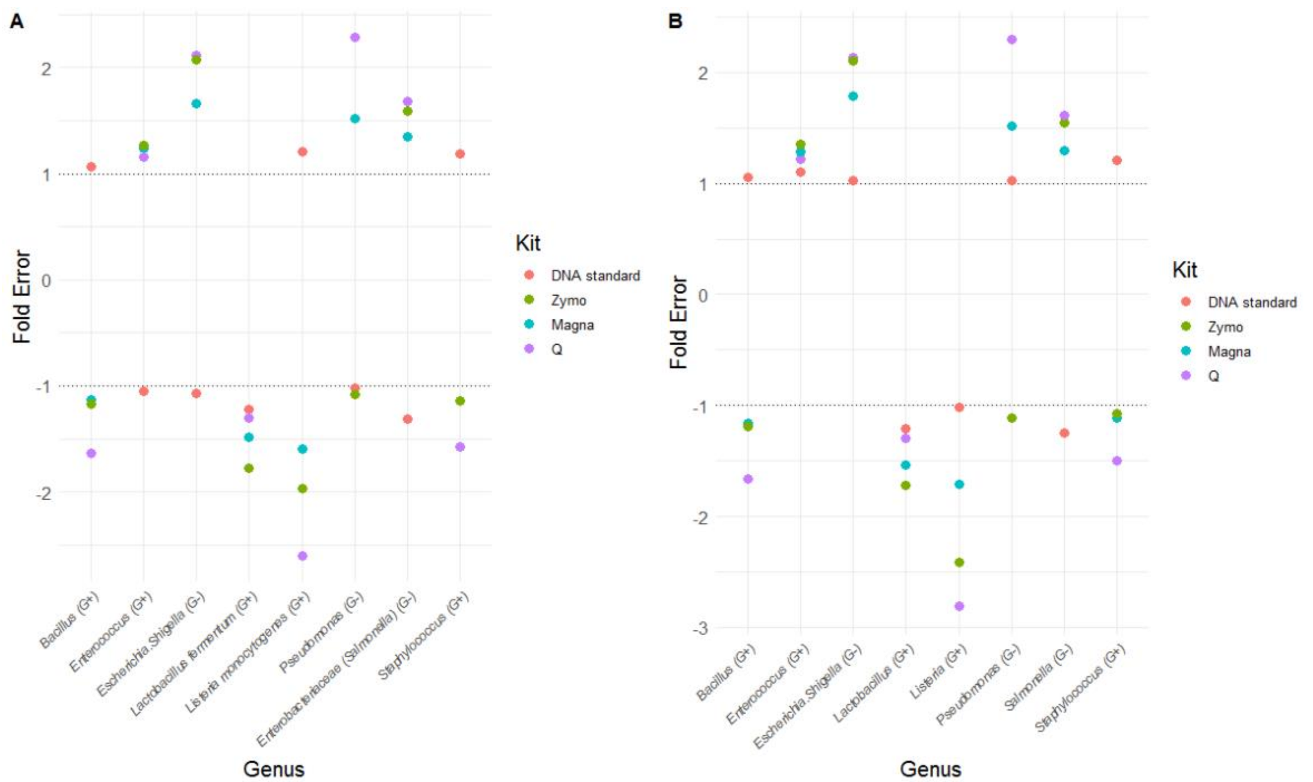
bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



711

712 **Figure 4:** Comparison of compositional profiles expressed by Kullback-Leibler divergence
713 (A+C) and Bray-Curtis dissimilarity (B+D) per pipeline. QIIME 2 results are shown in figure
714 A+B, NG-Tax results are shown in figure C+D. For both Kullback-Leibler and Bray-Curtis, 0
715 indicates an identical compositional profile, while higher numbers indicate more dissimilar
716 profiles.

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



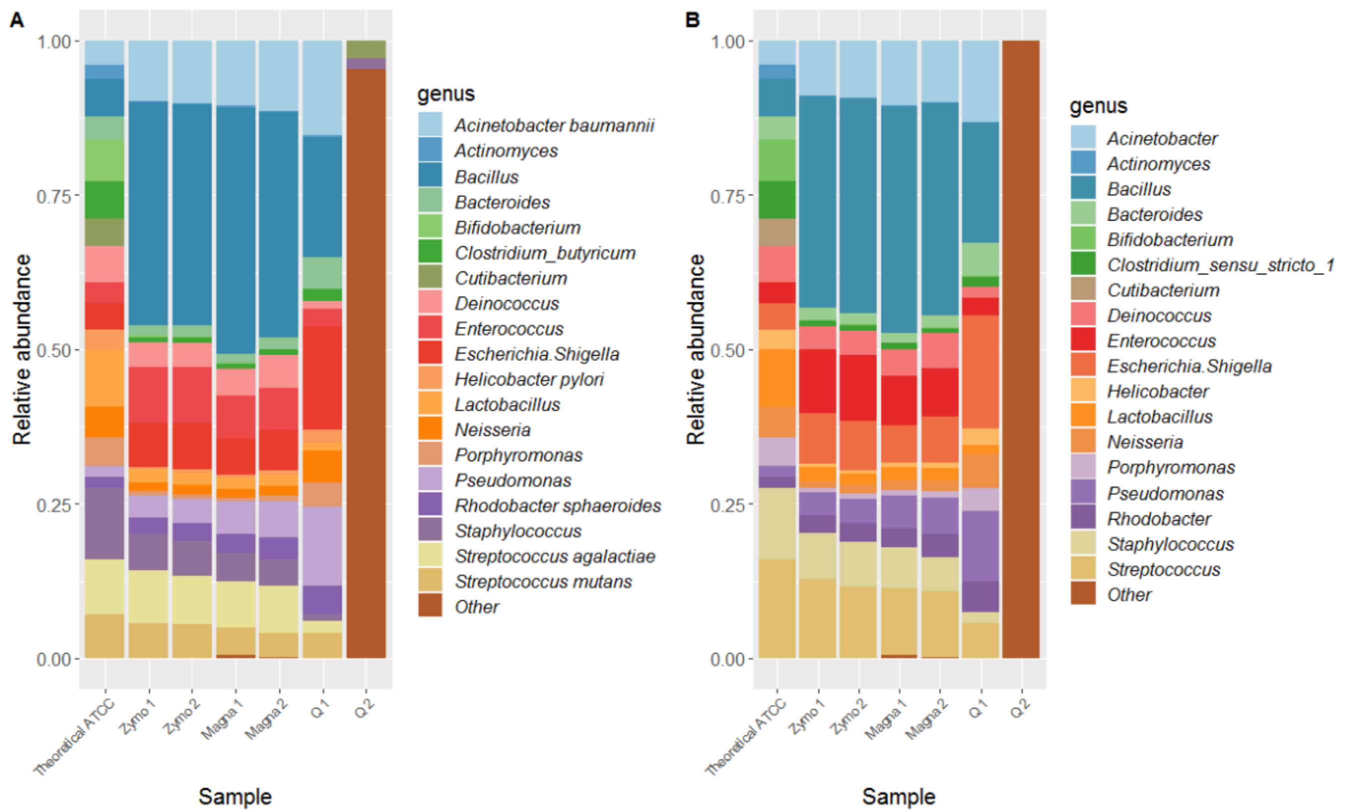
717

718 **Figure 5:** Fold error per bacterium as compared to theoretical values for QIIME 2 (A) and

719 NG-Tax (B). A value above 1 represents overestimation, and a value below -1 represents

720 underestimation.

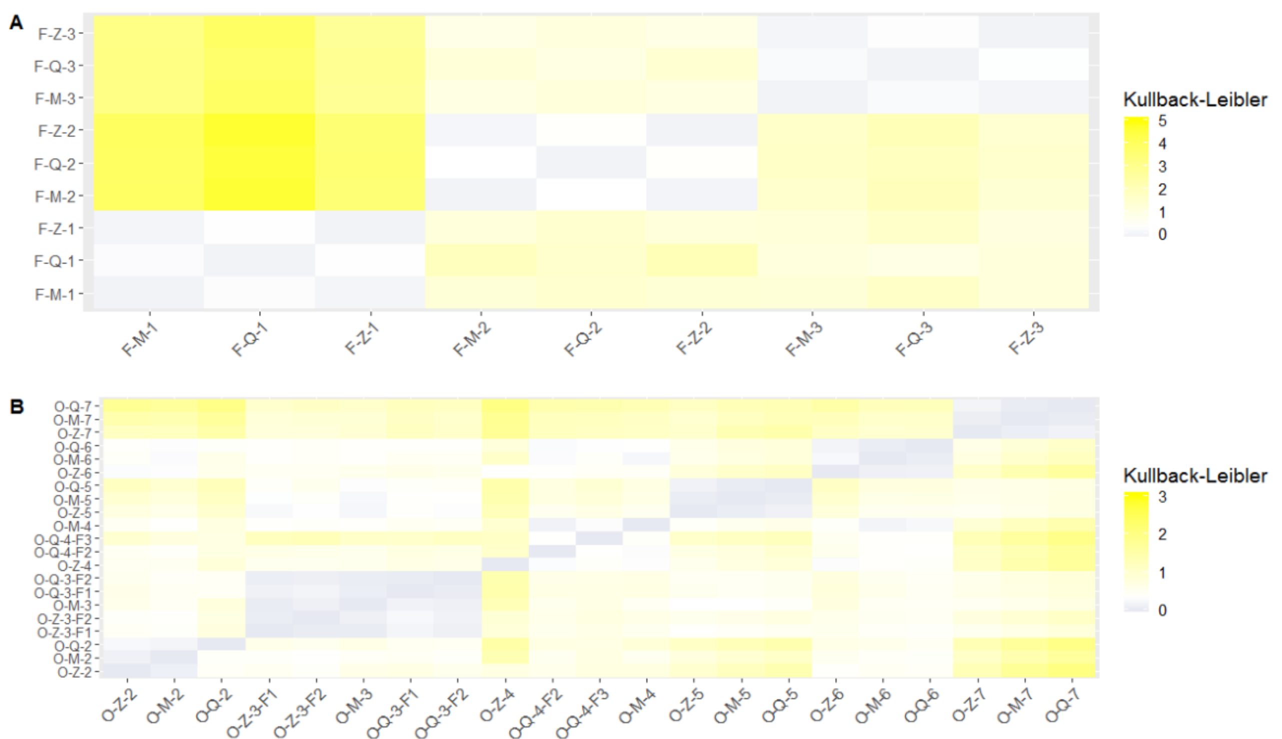
bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



721

722 **Figure 6:** Compositional profiles at genus level for QIIME 2 (A) and NG-Tax (B) for the

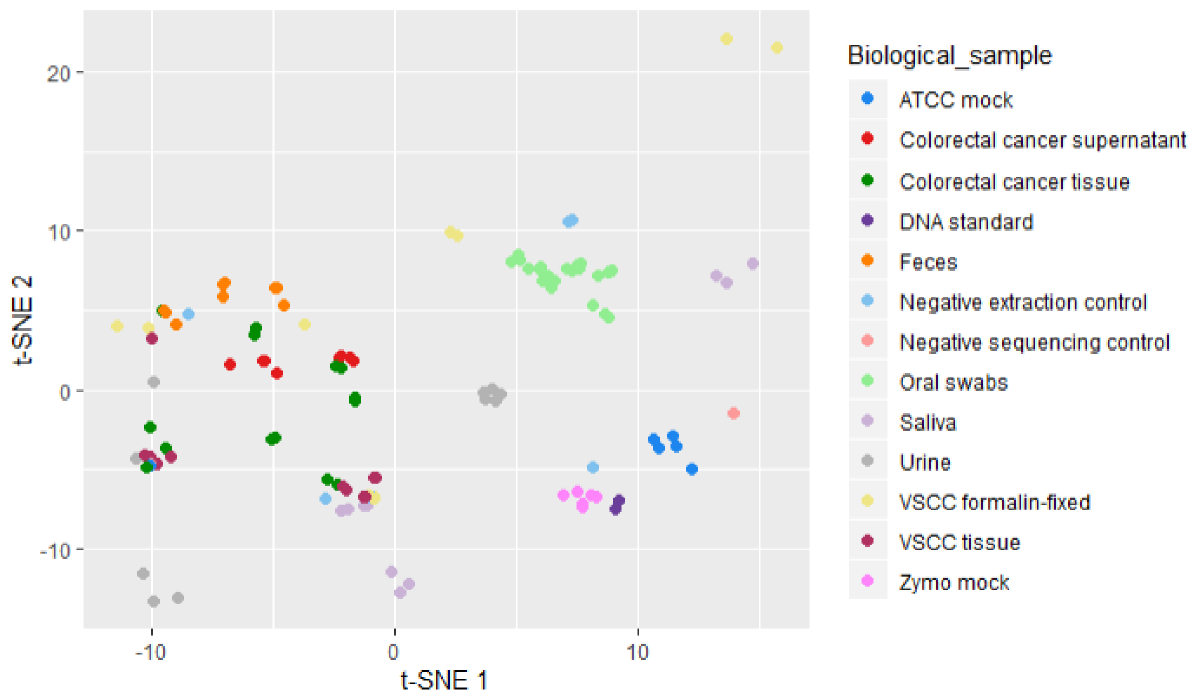
723 ATCC mock.



724

bioRxiv preprint doi: https://doi.org/10.1101/751123; this version posted August 30, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

725 **Figure 7:** Kullback-Leibler divergence heatmap of feces (A) and oral swabs (B). Blue
726 indicates highly similar composition, while yellow indicates divergence in composition. F1-
727 F2-F3 represent samples which have been sequenced in duplicate, but on different flow cells.



728

729 **Figure 8:** Bray-Curtis distance measures visualized by t-distributed stochastic neighbour
730 embedding (t-SNE) for all samples. Each dot in the plot represents a single sample, and short
731 distances between samples indicate high similarity.

732

733