

 Open access • Posted Content • DOI:10.1101/592154

## **Library preparation and sequencing platform introduce bias in metagenomics characterisation of microbial communities — Source link**

Casper Sahl Poulsen, Sünje Johanna Pamp, Ekstrøm Ct, Frank MÅ,ller Aarestrup

**Institutions:** Technical University of Denmark, University of Copenhagen

**Published on:** 28 Mar 2019 - bioRxiv (Cold Spring Harbor Laboratory)

**Topics:** Metagenomics

Related papers:

- [Sample storage conditions significantly influence faecal microbiome profiles.](#)
- [Two-Target Quantitative PCR To Predict Library Composition for Shallow Shotgun Sequencing.](#)
- [The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies](#)
- [Direct PCR Offers a Fast and Reliable Alternative to Conventional DNA Isolation Methods for Gut Microbiomes.](#)
- [Environmental Sequencing of Microbes in Water Samples](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/library-preparation-and-sequencing-platform-introduce-bias-hz6inhafeb>

1

2 **Library preparation and sequencing platform introduce bias in**  
3 **metagenomics characterisation of microbial communities**

4

5

6 Casper S. Poulsen<sup>1\*</sup>, Sünje J. Pamp<sup>1</sup>, Claus T. Ekstrøm<sup>2</sup> and Frank M. Aarestrup<sup>1</sup>

7 <sup>1</sup>Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark,  
8 Kongens Lyngby, Denmark.

9 <sup>2</sup>Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark.

10

11 \* Corresponding author

12 cspoulsen@hotmail.com (CP)

13

## 14 **Abstract**

15 Next generation sequencing technologies have become increasingly used to describe microbial communities.  
16 Metagenomics characterization of microbiomes is associated with minimal manipulation during sample  
17 processing, which includes sampling, storage, DNA isolation, library preparation and sequencing, before the  
18 raw data are obtained. Here we assess the effect of library preparation using a kit with a polymerase chain  
19 reaction (PCR) step (Nextera) and two PCR-free kits (NEXTflex and KAPA), and the effect of sequencing  
20 platform (HiSeq and NextSeq) on the description of microbial communities in pig feces and sewage. Two pig  
21 fecal samples were obtained from different farms and two sewage samples were collected as inlet water at  
22 a local wastewater treatment facility. Samples were processed to both perform DNA-isolation immediately  
23 upon arrival in the lab and after storage for 64 hours at -80°C, DNA isolation was performed in duplicate.

24 We find that both library preparation and sequencing platform had systematic effects on the microbial  
25 community description. The effects were at a level that made differentiating between the two pig fecal  
26 samples difficult. The sewage samples represented two very different communities and were at all times  
27 distinguishable from each other. We find that library preparation and sequencing platform introduced more  
28 variation than freezing the samples. The community changes did not seem associated with contamination  
29 during processing and distinct patterns connected specific types of organisms with a processing method, but  
30 it was difficult to generalize between samples. This highlights the need for continuous validation of the effect  
31 of sample processing in different types of samples and that all processing steps need to be considered when  
32 comparing between studies. We believe standardization of sample processing is key to generate comparable  
33 data within a study and that comparisons of differently generated data, e.g. in a meta-analysis, should be  
34 performed cautiously.

35

## 36 **Introduction**

37 Microbes are omnipresent and inhabit even the most extreme environments on earth. Metagenomics has  
38 provided unprecedented detail into these microbial communities, but the application is extending beyond  
39 environmental ecology. Metagenomics is applied heavily to human microbiomes and is being implemented  
40 to understand disease state (1–4) for diagnostic purposes (5) and surveillance (6–9). Data are a growing  
41 resource that can be utilized in meta-analysis and data-mining, revolutionizing the epidemiology of microbial  
42 diseases (6,9–12).

43 Findings from research related to human health and disease can be difficult to replicate as observed in  
44 different meta-analyses of 16S rRNA gene amplicon studies (13–16). Considering the large number of  
45 features (functional or taxonomic) under investigation in metagenomics, it is not surprising that studies never  
46 seem to lack significant results (17). Data dredging is a real concern in metagenomics, which brings to mind  
47 the “replication crisis” that has been highlighted in the field of psychology (18,19). Due to the challenge of  
48 replicating results, one must not over-emphasize the results from exploratory research and keep in mind  
49 with the maturation of metagenomics, that there is a need to continually validate the robustness and ability  
50 to replicate results. (20,21). With the improvement of reference databases and bioinformatics tools, the  
51 validation is an ongoing process (22–25).

52 Technical variation due to sample processing is an important factor that researchers have to minimize to  
53 make proper inferences in metagenomics studies. The effect of DNA isolation has been investigated in papers  
54 emphasizing the importance of this parameter (26–28). The effect of library preparation and sequencing  
55 platform has been investigated in metagenomics, primarily on human fecal samples. Library preparation  
56 affects taxonomic and functional characterization of human fecal samples and *in silico* constructed mock  
57 communities (21,29). However, in a study by Costea et al. (26), the effect of library preparation was lower  
58 compared with DNA isolation and intra- and inter-sample variation in general. The possibility that the  
59 sequencing platform could also have an effect on the characterization of microbiomes is highlighted in a  
60 study utilizing both metagenomics and 16S rRNA gene amplicon sequencing (30).

61 The aim of the present study is to assess the effect of library preparation (KAPA PCR-free, NEXTflex PCR-free  
62 and Nextera) and sequencing platform (Illumina HiSeq and NextSeq) on the metagenomics based description  
63 of two different microbiomes that includes two different sewage and pig fecal samples. We show that library  
64 preparation and sequencing infer systematic bias to the microbial characterization and that this effect is  
65 important when comparing similar samples, highlighting the need for consistent sample processing and  
66 demonstration of cautiousness when comparing data from different studies.

67

## 68 **Methods**

### 69 **Sample processing**

70 A subset of DNA samples was selected from an ongoing investigation of the effect of different aspects of  
71 sample processing. The DNA samples were from two pig fecal samples (P1 and P2) and two sewage samples  
72 (S1 and S2). The two pig fecal samples were collected on different occasions from different conventional pig  
73 production farms near the laboratory. The pig fecal samples were collected immediately after observed  
74 defecation, transferred to a cooling box and delivered to the laboratory for further processing within 3 hours.  
75 The two sewage samples were collected at a local wastewater treatment facility on different occasions. The  
76 sewage samples were 20 L inlet water, transported in cooling boxes and delivered for further processing  
77 within 20 mins. The sewage samples were centrifuged immediately upon arrival in the laboratory. Each pig  
78 fecal sample and sedimented sewage sample was processed in the same way by first homogenizing the  
79 samples then performing DNA isolation immediately and after 64 hours of storage at -80°C. DNA isolation  
80 was performed in duplicate with a modified QIAamp Fast DNA Stool Mini Kit (Qiagen) protocol including an  
81 initial bead beating step (MoBio garnet beads) (27) (S1 Fig). A negative DNA isolation (blank) control was  
82 included at each time of DNA isolation. The concentration of DNA samples was measured with the Qubit  
83 dsDNA High Sensitivity (HS) assay kit on a Qubit 2.0 fluorometer (Invitrogen, Carlsbad, CA) before storing the  
84 DNA at -20 °C.

### 85 **Library preparation and sequencing**

86 Library preparation and sequencing were performed in the order described below and the DNA was frozen  
87 between the sequencing runs:

88 **NEXTflex PCR-free on the HiSeq (NFHI).** Sequencing was performed at an external provider (Oklahoma  
89 Medical Research Foundation, Oklahoma, USA). The DNA (500 ng) was fragmented mechanically (Covaris

90 E220 evolution, aimed insert size=350bp, additional information was not possible to obtain from the  
91 provider) using ultrasonication. The NEXTflex library preparation was run PCR-free according to the  
92 manufacturer's recommendations. Sequencing was performed on the HiSeq 4000 (2x150 cycles, paired end).

93 **KAPA PCR-free on the HiSeq (KAHI).** Sequencing was performed at an external provider (Admera Health,  
94 New Jersey, USA). The DNA (500 ng) was fragmented mechanically (Covaris E220 evolution, aimed insert  
95 size=350bp, additional information was not possible to obtain from the provider) using ultrasonication. The  
96 KAPA library preparation was run PCR-free according to the manufacturer's recommendations. Sequencing  
97 was performed on the HiSeq 4000 (2x150 cycles, paired end).

98 **NEXTflex PCR-free on the NextSeq (NFNS).** The DNA (500 ng) was fragmented with mechanical  
99 fragmentation (Covaris E210, aimed insert size=350bp, Duty cvd=10 %, Intensity=5, Cycle burst = 200,  
100 Treatment time=240 sek) using ultrasonication. The NEXTflex library preparation was run PCR-free with  
101 Nextflex barcodes (NEXTflex-96 DNA barcodes) and sequenced in-house. The NEXTflex protocol was run  
102 according to the manufacturer's recommendations. Sequencing was performed on the NextSeq 500 (Mid  
103 output v2, 2x150 cycles, paired end).

104 **Nextera 1 and 2 on the NextSeq (NX1NS, NX2NS).** The Nextera XT library preparation was performed twice  
105 and sequenced in-house. The Nextera XT protocol was carried out according to the manufacturer's  
106 recommendations. This included a tagmentation step that fragments the DNA (1 ng) and ligates adaptors,  
107 and a PCR step amplifying DNA and adding indexing primers. Library cleanup was performed with AMPure  
108 XP beads and normalized before sequencing on the NextSeq 500 (Mid output v2, 2x150 cycles, paired end).  
109 The bioanalyzer results revealed that the aimed insert size of 350 bp was larger than expected (S1 File).

## 110 **Bioinformatics and statistical analysis**

111 Pre-processing of raw reads included trimming (Phred quality score = 20) and removal of reads shorter than  
112 50bp (BBduk2) (31). Mapping was performed with a Burrows-Wheeler aligner (BWA-mem) as implemented  
113 in MGmapper (22). Mapping was performed in the default "best mode" to 11 databases, first filtering against  
114 the human database then extracting the number of raw reads mapping to the genomes of bacteria, fungi,  
115 archaea, viruses and *Cryptosporidium*. A read count correction was implemented to adjust large hit counts  
116 to specific contigs as implemented in Hendriksen et al. (9). All counts in the count table were divided by two  
117 to account for reads were mapping as proper pairs and then aggregating to genus level. The processed count  
118 table, metadata and feature data are available as S2 (File) and the raw reads are deposited at the European  
119 Nucleotide Archive (ENA) (Project acc.: PRJEB31650).

120 All statistical analyses adhered to the compositional data analysis framework and were performed in R  
121 version 3.5.2 (32–34). Initial filtering of the count matrix was performed in all analyses, removing all genera  
122 below an average count of 5. The estimation of zeroes was performed using simple multiplicative  
123 replacement (35). Isometric log-ratio transformation (ILR) was used in: The principal component analysis  
124 (PCA), heatmaps to perform complete-linkage clustering analysis of the samples, boxplots to calculate  
125 pairwise Euclidean distance between samples and in multivariate analysis of variances testing which  
126 parameters that significantly influence the multivariate outcome the most using permutation tests  
127 (32,34,36,37). Centered log-ratio transformation (CLR) was used in: Sparse partial-least-squares discriminant  
128 analysis (sPLS-DA) and constrained ordination with redundancy analysis (rda), where it was important to keep

129 genera information after the transformation (32,34,38). Analyses performed are included in the publication  
130 as S3 (File) and the code is available from  
131 <https://github.com/csapou/LibraryPreparationandSequencingPlatform>.

132

## 133 Results

### 134 Quality control of sequencing output

135 The number of raw reads from the different library preparations and sequencing platforms were similar with  
136 about a factor 2 difference when comparing the medians. The highest number of reads were obtained from  
137 the NEXTflex HiSeq run (median: 12.1, range: 6.3 – 30.8 million reads) and the lowest from the NEXTflex  
138 NextSeq run (median: 7.6, range: 2.7 – 9.4 million reads). The outputs from the KAPA HiSeq run (median: 9.4,  
139 range: 7.8 – 17.4 million reads) and the Nextera NextSeq runs (median: 10.2, range: 6.5 – 16.5 million reads)  
140 were about the same. More reads were obtained from the pig fecal samples compared with the sewage, but  
141 a larger proportion of the sewage reads mapped to the reference databases. The microbial community of the  
142 sewage samples exhibited a higher  $\alpha$ -diversity (Simpson) than the pig feces (Table S1). However, the number  
143 of mapped reads were higher for the sewage samples, and many of the samples had reached a plateau as  
144 observed when creating a rarefaction curve (S2 Fig). Similar results were obtained when comparing percent  
145 of unmapped reads across the different library preparation and sequencing platform runs (S1 Table).

### 146 Sample processing impact on microbial characterization

147 The pairwise Euclidean distance was calculated between all of the samples and visualized using PCA (S3A Fig).  
148 The sample type explained the most variance and pig feces and sewage samples were clearly separated on  
149 the first axis. Separation of the two sewage samples was observed on the second axis. However, the two pig  
150 fecal samples formed a single group. Ordination of the pig feces and sewage samples separately revealed  
151 that the two pig fecal samples seemed to belong to two separate groups (S3B Fig), and a clear separation of  
152 the two sewage samples was still observed (S3C Fig). Creating boxplots of the pairwise distances revealed  
153 that both library preparation, sequencing platform and storage did not hamper the ability to differentiate  
154 between the two sewage samples as observed in the PCA (Fig 1). However, a large degree of overlap was  
155 observed between pig feces 1 and 2 comparisons relative to comparing within the two samples representing  
156 the effect of the different sample processing parameters. In general, larger distances were calculated for the  
157 comparisons of sample processing parameters in pig fecal samples compared with sewage. The shortest  
158 distances were observed when comparing the DNA isolation replicates and the replicates of the Nextera  
159 NextSeq runs. The distances between samples that only differed in library preparation and sequencing  
160 platform were greater compared with samples that differed in whether they were processed directly or after  
161 freezing at  $-80^{\circ}\text{C}$  for 64 hours. The sequencing platform seemed to be the major contributor of variation  
162 when comparing the samples that were prepared with NEXTflex and sequenced on the HiSeq and NextSeq  
163 (Fig 1). To investigate the effect of sample processing further, PCAs were created for the individual samples  
164 (P1, P2, S1 and S2). Similar patterns were observed in all samples indicating that there was a systematic effect  
165 from storage, library preparation and sequencing platform. In general, the DNA isolation replicates were  
166 similar as well as the two Nextera NextSeq runs (Fig 2). Investigating how large an effect the different  
167 parameters had by partitioning sums of squares of the Euclidean distance matrix revealed that all of the

168 parameters had a significant effect when assessing uncorrected p-values except for storage when comparing  
169 all of the samples and in pig feces 2. Comparing the percent variation in pig feces attributed to sample (P1  
170 and P2) (21.1-%) library preparation (32.7-%) and sequencing platform (19.1-%) were at a similar level, further  
171 emphasizing the importance of sample processing when comparing communities that are more similar in  
172 general (Table 1).

173 **Fig 1. Boxplots of pairwise distances between different groupings of samples.** Within the different groups,  
174 dots representing the distances were colored according to which sample the comparison was made in. Black  
175 dots represent a distance between two different samples.

176 **Fig 2. Principal component analysis (PCA) subset to the different sample matrices.** Variance explained by  
177 the two first axes are included in their labels. The unique DNA samples processed differently are connected  
178 with dotted lines.

179 **Table 1: Comparing the effect of sample (P1, P2, S1 and S2) and different parameters in sample processing.**  
180 Statistical test were performed by multiple permutations partitioning sum of squares. The *P-value* as well as  
181 the percent of variation explained by the parameters is reported testing different inclusions of samples (All,  
182 pig feces, sewage, P1, P2, S1 and S2).

Samples included	Sample <i>P-value</i> (%)	Storage <i>P-value</i> (%)	Library preparation <i>P-value</i> (%)	Sequencing platform <i>P-value</i> (%)
All	<10 <sup>-5</sup> (81.9)	6.6×10 <sup>-2</sup> (0.5)	7.2×10 <sup>-4</sup> (2.6)	3.6×10 <sup>-4</sup> (1.8)
Pig feces	<10 <sup>-5</sup> (21.1)	3.8×10 <sup>-3</sup> (3.3)	<10 <sup>-5</sup> (32.7)	<10 <sup>-5</sup> (19.1)
Sewage	<10 <sup>-5</sup> (61.7)	2.5×10 <sup>-2</sup> (2.9)	1.1×10 <sup>-2</sup> (5.1)	4.3×10 <sup>-3</sup> (4.5)
Pig feces 1	Na*	3.1×10 <sup>-3</sup> (9.7)	<10 <sup>-5</sup> (40.6)	<10 <sup>-5</sup> (26.2)
Pig feces 2	Na	0.17 (2.7)	2×10 <sup>-5</sup> (50.3)	2×10 <sup>-5</sup> (25.3)
Sewage 1	Na	<10 <sup>-5</sup> (15.1)	4×10 <sup>-5</sup> (16.9)	<10 <sup>-5</sup> (12.8)
Sewage 2	Na	<10 <sup>-5</sup> (14.0)	2×10 <sup>-5</sup> (20.6)	<10 <sup>-5</sup> (19.6)

183 \* No statistics were obtained when subsetting to a single sample (P1, P2, S1 and S2).

## 184 Sample processing impact on indicator organisms

185 To investigate the effect of library preparation and sequencing platform on specific organisms, an initial  
186 overview was obtained for the 30 most abundant genera in heatmaps of pig feces and sewage separately. As  
187 highlighted above, the importance of sequencing platform in differentiating pig feces was observed by being  
188 the first branching of the samples (P1 and P2) in the dendrogram (Fig 3A). Clustering was also observed for  
189 the storage condition and library preparation. The pig feces contained both Gram-negative and -positive  
190 bacteria, and the third cluster exclusively consisted of Gram-negatives. There were a few Gram-negatives in  
191 the other clusters, indicating that sample processing shifts the abundance profiles for specific types of  
192 organisms, in this case, it seemed associated with cell wall structure (Fig 3A). A similar pattern was observed  
193 in sewage that mainly consisted of Gram-negatives, but the majority of Gram-positives were part of cluster  
194 four including *Clostridium*, *Faecalibacterium* and *Roseburia*. However, this cluster also contained Gram-  
195 negative genera (Figure 3B).

196 **Fig 3. Heatmaps of pig feces and sewage samples separately with the 30 most abundant genera.** Complete-  
197 linkage clustering was performed to create dendrograms for both genera and samples. Pearson correlation  
198 was used to cluster the genera and Euclidean distances were calculated on the isometric log-ratio

199 transformed count matrix were used to cluster the samples. (A) Heat map of all pig feces samples, where the  
200 first branching was according to sequencing platform. The third cluster of genera exclusively contained  
201 Gram-negatives. (B) Heat map of all sewage samples. The fourth cluster mainly consisted of Gram-positives.  
202 A few Gram-positives were also present in the other clusters.

203 One explanation for the community differences observed by sample processing could be a possible  
204 contamination during the library preparation and sequencing steps. To elucidate this, sPLS-DA was  
205 performed, assessing which genera best characterize the library preparation and sequencing platform  
206 processing methods. Component 1, 2 and 3 were included in the model containing 5, 50 and 20 different  
207 genera, respectively (S4 Fig). The majority of microorganisms were the highly abundant organisms observed  
208 across all of the sample processing methods. However, a few were clear indicators of contamination during  
209 library preparation and sequencing and were mainly present in a single processing method. This included  
210 *Methylobacterium* in the KAPA HiSeq run that has previously been associated with kit contamination and  
211 *Cutibacterium* in the second Nextera NextSeq run, a typical bacterium inhabiting the skin (39). A heat map of  
212 the 30 most abundant genera in the blank controls additionally revealed a high abundance of *Ralstonia*  
213 in the Nextera NextSeq runs that were performed with the same kit reagents (S5 Fig). The separation of the  
214 samples according to the different processing parameters therefore seemed to be real changes to the relative  
215 abundances between organisms inherently present in the communities and not due to contamination. A  
216 constrained ordination, also subsetted according to if samples were processed directly or after freezing, was  
217 performed to assess if groups of organisms at a taxonomic higher level were associated with a specific library  
218 preparation and sequencing method. In the pig feces, Proteobacteria seemed associated with the HiSeq runs.  
219 However, this was not observed in sewage. In sewage, Archaea were associated with the HiSeq runs, but also  
220 Eukaryotes consisting of Fungi and *Cryptosporidium* seemed associated with the HiSeq runs in sewage 1 (S6  
221 Fig). Overall, it was difficult to observe a pattern when assessing this grouping of genera, highlighting that it  
222 might be difficult to generalize the effect of sample processing in different sample types and different  
223 samples of the same type.

224

## 225 Discussion

226 With the increasing amount of metagenomics data in public repositories; meta-analysis and cross-study  
227 analysis based on data from different studies are exciting new opportunities to gain further insight into the  
228 microbial world (10–12,24,40). Data generation is usually not performed with a standard procedure across  
229 studies, and sample processing is an important factor to be aware of when trying to make inferences in these  
230 cross study investigations (21,26). In the present study, both library preparation and sequencing platform  
231 had a significant effect on explaining the variance in the data (Table 1). That these parameters infer changes  
232 to the community description has also been observed previously (21,29,30). In the study by Costea et al. (26),  
233 DNA isolation had the largest effect compared with other technical variations. In the first phase of the study  
234 by Costea et al. (26) samples were sent out for DNA isolation and sequenced centrally. In the present study,  
235 DNA isolation was performed centrally by the same person and library preparation and sequencing in-house  
236 or at external providers, but not in any of the cases by the same person, possibly increasing variation due to  
237 DNA shipping and handling in this specific step. When performing a validation study assessing the technical  
238 variation of sample processing, the large number of methodologies and variations thereof make it impossible



239 to test all parameters. It is likely that selecting methods that are based on different principles and for specific  
240 purposes yield results that highlight the importance of this specific step. Bowers et al. (29) investigated  
241 community changes using different amounts of input DNA, and observed that this modification had a  
242 significant effect on community description. In the present study, investigation of sequencing platforms were  
243 limited to the NextSeq and HiSeq, which are both Illumina platforms resembling each other in technology,  
244 and which were selected due to their popularity in metagenomics with low cost relative to output (41).  
245 Nonetheless, a very large effect was attributed to the sequencing platform and that was also observed when  
246 using the same library preparation kit (NEXTflex PCR-Free) (Fig 1). The library preparation included two  
247 methods that required pre-fragmented DNA that was prepared PCR-free (KAPA and NEXTflex). It was decided  
248 to include the Illumina Nextera library preparation as well to compare with a technique that does not  
249 resemble the others in having enzymatic fragmentation and which involved a PCR step that is commonly  
250 applied when too little DNA is available to prepare DNA for sequencing PCR-free. However, the two Nextera  
251 runs were relatively similar compared with the NEXTflex run when sequenced on the NextSeq (Fig 2). The  
252 present study was not a full factorial experiment and this should be emphasized when comparing the effect  
253 sizes of specific processing parameters.

254 One explanation for the differences observed between the processing runs can be contamination bias. When  
255 designing a metagenomics study, it is to some extent possible to remove kit contaminations or carry-over  
256 between sequencing runs from the data *in-silico*, if for instance, blank controls are included or by rotating  
257 indexing primers between adjacent runs, respectively (42). In the present study, comparing the sPLS-DA  
258 results with the blank controls rarely identified the same genera, indicating that the genera reported to  
259 explain the specific sample processing the most were not due to contamination during DNA extraction. The  
260 general variation associated with redoing the library preparation and sequencing was low when comparing  
261 the two Nextera sequencing runs (Figs 1 and 2). The differences observed are therefore most likely true  
262 technical variation associated with the sample processing. Furthermore, it was possible to detect that these  
263 patterns were systematic in the different samples (Fig 2), and that this could partly be explained with some  
264 crude features such as distinguishing between Gram-negative and -positive bacteria or at a higher taxonomic  
265 classification (Fig 3 and S6 Fig). The grouping of genera were selected before analysis to be investigated, but  
266 they might be confounders of the underlying explanation that could be associated with DNA characteristics  
267 such as guanine-cytosine percent (GC%) or other specific DNA patterns. Another possibility is that DNA  
268 fragmentation during sampling, storage and DNA isolation provide DNA of different quality for specific  
269 organism groups. A shift in community structure is then reflected in the selection of different fragment sizes  
270 during the library preparation and sequencing.

271 The Euclidean distances obtained from comparing within the two pig fecal samples separately relative to the  
272 two sewage samples also revealed that storage, library preparation and sequencing platform has a larger  
273 effect in pig feces (Fig 1). Since, the distances between the two pig fecal samples were smaller relative to the  
274 distances between the two sewage samples, it was difficult to discern the two pig fecal samples when  
275 samples were processed differently (Fig 3). It is concerning that the variation due to sample processing  
276 hampers the ability to differentiate between two different pig fecal samples, and this might hamper the  
277 ability to draw meaningful conclusions when technical variations cannot be distinguished from “true”  
278 changes. These results should on the other hand not be overstated; the two pig fecal samples were obtained  
279 from an in-bred race raised under very similar conditions including feeding, even though they were obtained  
280 from two different healthy pigs at two different farms, the two communities are relatively similar. The finding

281 highlights that the importance of technical variation depends on the differences that one is trying to detect  
282 (16). If sewage samples were the only sample matrix investigated, the technical variation did not hamper the  
283 ability to differentiate between the two sewage samples. These findings suggest that library preparation and  
284 sequencing are important parameters to keep constant when a study is trying to detect small changes in  
285 community structure.

286

## 287 **Acknowledgments**

288 The authors wish to thank Ms. Marie Jensen, Ms. Berith Knudsen and Mr. Carsten Bidstrup for their help with  
289 the sampling. Mr. Jacob Jensen and Ms. Marlene Dalggaard for technical assistance during in-house library  
290 preparation and sequencing. Mr. Rolf Kaas for help with the upload of raw sequencing reads to ENA. Mr.  
291 Jeffrey Skiby for language editing. This study has received funding from the European Union's Horizon 2020  
292 research and innovation programme under grant agreement no. 643476 (COMPARE) and The Novo Nordisk  
293 Foundation (NNF16OC0021856: Global Surveillance of Antimicrobial Resistance).

294

## 295 **References**

- 296 1. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in  
297 type 2 diabetes. *Nature* [Internet]. Nature Publishing Group; 2012;490(7418):55–60. Available from:  
298 <http://dx.doi.org/10.1038/nature11450>
- 299 2. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-  
300 naive microbiome in new-onset Crohn's disease. *Cell Host Microbe*. 2014;15(3):382–92.
- 301 3. Yu J, Feng Q, Wong SH, Zhang D, Yi Liang Q, Qin Y, et al. Metagenomic analysis of faecal microbiome  
302 as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut*. 2017;66(1):70–8.
- 303 4. Zeller G, Tap J, Sobhani I, Amiot A, Tap J, Tran Van Nhieu J, et al. Potential of fecal microbiota for  
304 early-stage detection of colorectal cancer. *Mol Syst Biol*. 2014;10(11):766–766.
- 305 5. Dekker JP. Metagenomics for Clinical Infectious Disease Diagnostics Steps Closer to Reality. *J Clin*  
306 *Microbiol*. 2018;1–7.
- 307 6. Petersen TN, Rasmussen S, Hasman H, Carøe C, Baelum J, Schultz AC, et al. Meta-genomic analysis of  
308 toilet waste from long distance flights; a step towards global surveillance of infectious diseases and  
309 antimicrobial resistance. *Nat Publ Gr*. 2015;
- 310 7. Nieuwenhuijse DF, Koopmans MPG. Metagenomic Sequencing for Surveillance of Food- and  
311 Waterborne Viral Diseases. *Front Microbiol*. 2017;8(February):1–11.
- 312 8. Hjelmsø MH, Møllerup S, Jensen RH, Pietroni C, Lukjancenko O, Schultz AC, et al. Metagenomic  
313 analysis of viruses in toilet waste from long distance flights—A new procedure for global infectious  
314 disease surveillance. *PLoS One*. 2019;14(1):e0210368.
- 315 9. Hendriksen RS, Munk P, van Bunnik B, McNally L, Lukjancenko O, Röder T, et al. Global monitoring of  
316 antimicrobial resistance based on human sewage. *Nat Commun*. 2018;

- 317 10. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large  
318 Metagenomic Datasets : Tools and Biological Insights. *PLoS Comput Biol*. 2016;1–26.
- 319 11. Armour C, Nayfach S, Pollard K, Sharpton T. A Metagenomic Meta-Analysis Reveals Functional  
320 Signatures of Health and Disease in the Human Gut Microbiome. *bioRxiv*. 2019;
- 321 12. Sze MA, Schloss PD. Leveraging Existing 16S rRNA Gene Surveys To Identify Reproducible Biomarkers  
322 in Individuals with Colorectal Tumors. *MBio*. 2018;9(3):1–16.
- 323 13. Finucane MM, Sharpton TJ, Laurent TJ, Pollard KS. A Taxonomic Signature of Obesity in the  
324 Microbiome ? Getting to the Guts of the Matter. *PLoS One*. 2014;9(1):1–5.
- 325 14. Sze MA, Schloss PD. Looking for a Signal in the Noise : Revisiting Obesity and the Microbiome. *MBio*.  
326 2016;7(4):1–10.
- 327 15. Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with obesity and IBD.  
328 *FEBS Lett*. 2016;588(22):4223–33.
- 329 16. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Jansson JK, Gordon JI, et al. Meta-analyses of  
330 studies of the human microbiota. *Genome Res*. 2013;1704–14.
- 331 17. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies  
332 identifies disease-specific and shared responses. *Nat Commun*. 2017;
- 333 18. Pashler H, Wagenmakers EJ. Editors ' Introduction to the Special Section on Replicability in  
334 Psychological Science : A Crisis of Confidence? *Assoc Psychol Sci*. 2012;2011–3.
- 335 19. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* (80- ).  
336 2015;349(6251).
- 337 20. Vogtmann E, Chen J, Kibriya MG, Chen Y, Islam T, Eunes M, et al. Comparison of Fecal Collection  
338 Methods for Microbiota Studies in Bangladesh. Elkins CA, editor. *Appl Environ Microbiol* [Internet].  
339 2017 May 15 [cited 2018 Jan 2];83(10):e00361-17. Available from:  
340 <http://aem.asm.org/lookup/doi/10.1128/AEM.00361-17>
- 341 21. Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, et al. Library preparation  
342 methodology can influence genomic and functional predictions in human microbiome research. *Proc*  
343 *Natl Acad Sci U S A* [Internet]. 2015 Nov 10 [cited 2018 Jan 2];112(45):14024–9. Available from:  
344 <http://www.pnas.org/lookup/doi/10.1073/pnas.1519288112>
- 345 22. Petersen TN, Lukjancenko O, Thomsen MCF, Maddalena Sperotto M, Lund O, Møller Aarestrup F, et  
346 al. MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence  
347 reads. An L, editor. *PLoS One* [Internet]. 2017 May 3 [cited 2018 Jan 2];12(5):e0176469. Available  
348 from: <http://www.ncbi.nlm.nih.gov/pubmed/28467460>
- 349 23. Visconti A, Martin TC, Falchi M. YAMP : a containerized workflow enabling reproducibility in  
350 metagenomics research. *Gigascience*. Oxford University Press; 2019;(June 2018):1–9.
- 351 24. Li X, Naser SA, Khaled A, Hu H, Li X. When old metagenomic data meet newly sequenced genomes ,  
352 a case study. *PLoS One*. 2018;1–16.
- 353 25. Kirstahler P, Bjerrum SS, Friis-Møller A, La Cour M, Aarestrup FM, Westh H, et al. Genomics-Based  
354 Identification of Microorganisms in Human Ocular Body Fluid. *Sci Rep*. 2018;8(1):1–14.

- 355 26. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human  
356 fecal sample processing in metagenomic studies. *Nat Biotechnol* [Internet]. Nature Publishing  
357 Group; 2017;35(11):1069–76. Available from: <http://dx.doi.org/10.1038/nbt.3960>
- 358 27. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, et al. Impact of Sample  
359 Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. Jansson JK,  
360 editor. *mSystems* [Internet]. 2016 Oct 25 [cited 2018 Jan 2];1(5):e00095-16. Available from:  
361 <http://msystems.asm.org/lookup/doi/10.1128/mSystems.00095-16>
- 362 28. Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, et al.  
363 Choice of bacterial DNA extraction method from fecal material influences community structure as  
364 evaluated by metagenomic analysis. *Microbiome*. 2014;2:1–11.
- 365 29. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation protocols  
366 and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC*  
367 *Genomics* [Internet]. 2015 Oct 24 [cited 2018 Jan 2];16(1):856. Available from:  
368 <http://www.biomedcentral.com/1471-2164/16/856>
- 369 30. Clooney AG, Fouhy F, Sleator RD, O’ Driscoll A, Stanton C, Cotter PD, et al. Comparing Apples and  
370 Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. White BA, editor.  
371 *PLoS One* [Internet]. 2016 Feb 5 [cited 2018 Jan 2];11(2):e0148028. Available from:  
372 <http://dx.plos.org/10.1371/journal.pone.0148028>
- 373 31. JGI. BBDuk Guide [Internet]. 2019. Available from: [https://jgi.doe.gov/data-and-tools/bbtools/bb-  
374 tools-user-guide/bbduk-guide/](https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/)
- 375 32. Cao KAL, Costello ME, Lakis VA, Bartolo F, Chua XY, Brazeilles R, et al. MixMC: A multivariate  
376 statistical framework to gain insight into microbial communities. *PLoS One*. 2016;
- 377 33. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput  
378 sequencing data. *Can J Microbiol*. 2016;62(8):692–703.
- 379 34. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional:  
380 And this is not optional. *Front Microbiol*. 2017;8(NOV):1–6.
- 381 35. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with Zeros and Missing Values in  
382 Compositional Data Sets Using Nonparametric Imputation. *Math Geol*. 2003;35(3):253–78.
- 383 36. Egozcue J, Pawlowsky Glahn V, Mateu-Figueras G, Barceló Vidal C. Isometric logratio for  
384 compositional data analysis. *Math Geol*. 2003;35(3):279–300.
- 385 37. Filzmoser P, Hron K, Reimann C. The bivariate statistical analysis of environmental (compositional)  
386 data. *Sci Total Environ* [Internet]. Elsevier B.V.; 2010;408(19):4230–8. Available from:  
387 <http://dx.doi.org/10.1016/j.scitotenv.2010.05.011>
- 388 38. Aitchison J. *The Statistical Analysis of Compositional Data*. London: Chapman and Hall; 1986.
- 389 39. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory  
390 contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;1–12.
- 391 40. Segata N. On the Road to Strain-Resolved Comparative Metagenomics. *mSystems*. 2018;3(2):1–6.
- 392 41. Escobar-Zepeda A, Vera-Ponce De León A, Sanchez-Flores A. The Road to Metagenomics: From

393 Microbiology to DNA Sequencing Technologies and Bioinformatics BRIEF HISTORY OF MICROBIAL  
394 COMMUNITIES STUDY. Front Microbiol. 2015;6.

395 42. Gruber K. Here, there, and everywhere. EMBO Rep. 2015;16(8):898–901.

396

## 397 **Supporting information**

398 **S1 Fig. Study design.** Two pig feces samples and two sewage samples were processed directly or after storage  
399 at -80°C for 64 hours. The DNA isolation were performed in duplicates. Library preparation and sequencing  
400 was performed in four different combinations: KAPA PCR-free on the HiSeq, NEXTflex PCR free both on the  
401 HiSeq and NextSeq, and the Nextera protocol were run twice on the NextSeq. The setup resulted in a total  
402 of 80 metagenomes plus five negative controls.

403 **S2 Fig. Rarefaction curves.**

404 **S3 Fig. Principal component analysis (PCA) for all samples and subsetted to pig feces and sewage.** Variance  
405 explained by the two first axes are included in their labels. (A) PCA were generated for all samples forming  
406 three clusters with pig feces together and sewage 1 and sewage 2 seperately, (B) PCA were generated for all  
407 pig feces samples separation were now observed between pig feces 1 and pig feces 2, and (C) PCA were  
408 generated for all sewage samples that were still easy to discriminate.

409 **S4 Fig. Sparse partial least square discriminant analysis (sPLS-DA).** The sPLS-DA were run with a unique  
410 identifier for a specific DNA sample that were then processed differently in the generation of libraries and  
411 sequencing to select the most discriminative genera in explaining this aspect of sample processing.

412 **S5 Fig. Heatmap of negative controls with the thirty most abundant genera.** Complete-linkage clustering  
413 was performed to create dendograms for both genera and samples. Pearson correlation were used to cluster  
414 the genera and Euclidean distances calculated on the isometric log-ratio transformed count matrix were used  
415 to cluster the samples.

416 **S6 Fig. Redundancy analysis (rda) subsetted to sample matrix and if samples were frozen or processed  
417 directly.** Taxonomic patterns were investigated by plotting genera coloured according to different taxonomic  
418 groups. DNA isolation replicates were connected with lines.

419 **S1 Table. Sequencing quality control and alpha-diversity overview.**

420 **S1 File. Bioanalyzer electropherograms.**

421 **S2 File. Count table, metadata and feature data used in the analysis.**

422 **S3 File. R markdown file containing the R analysis.**

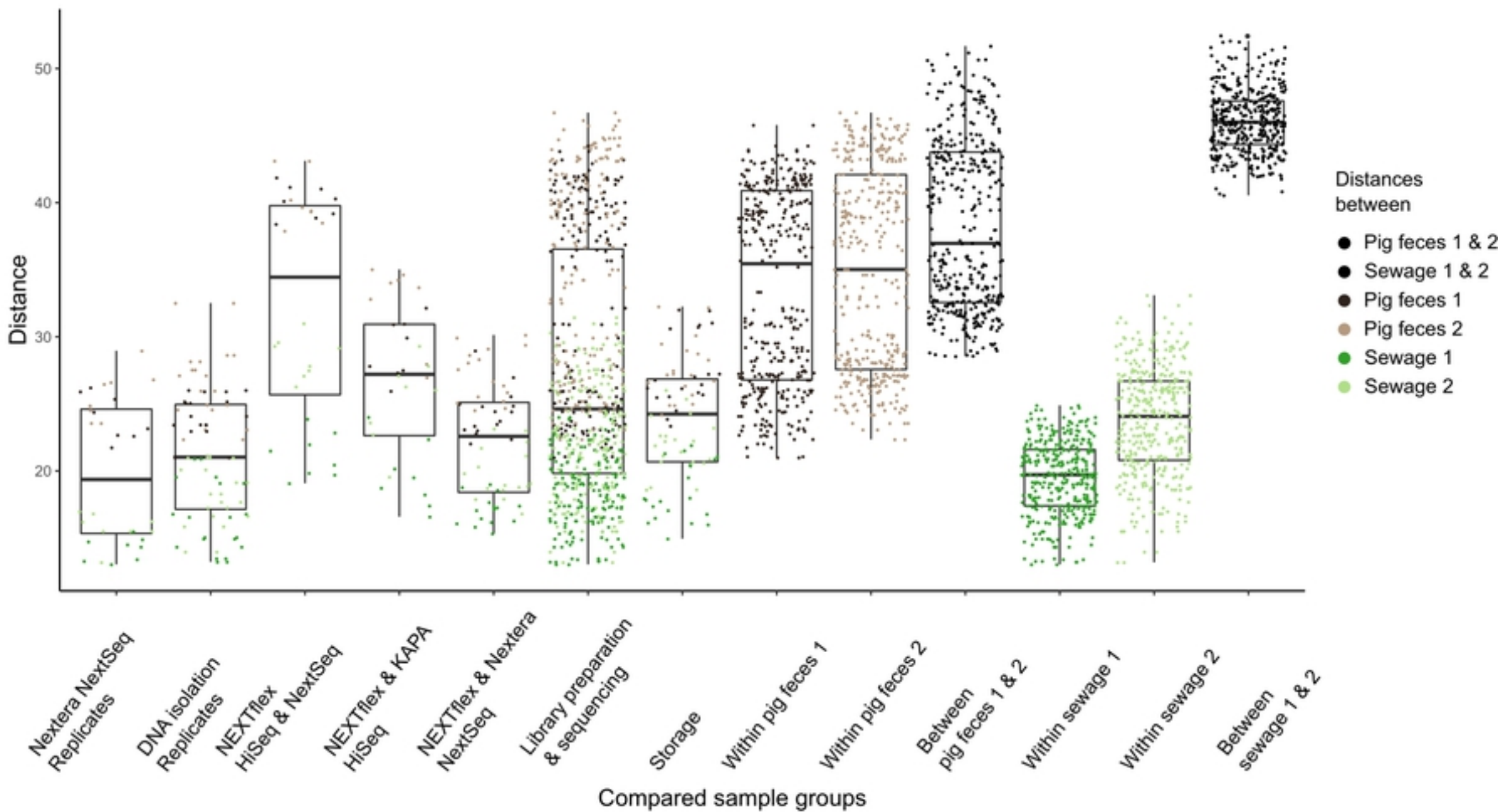


Figure 1

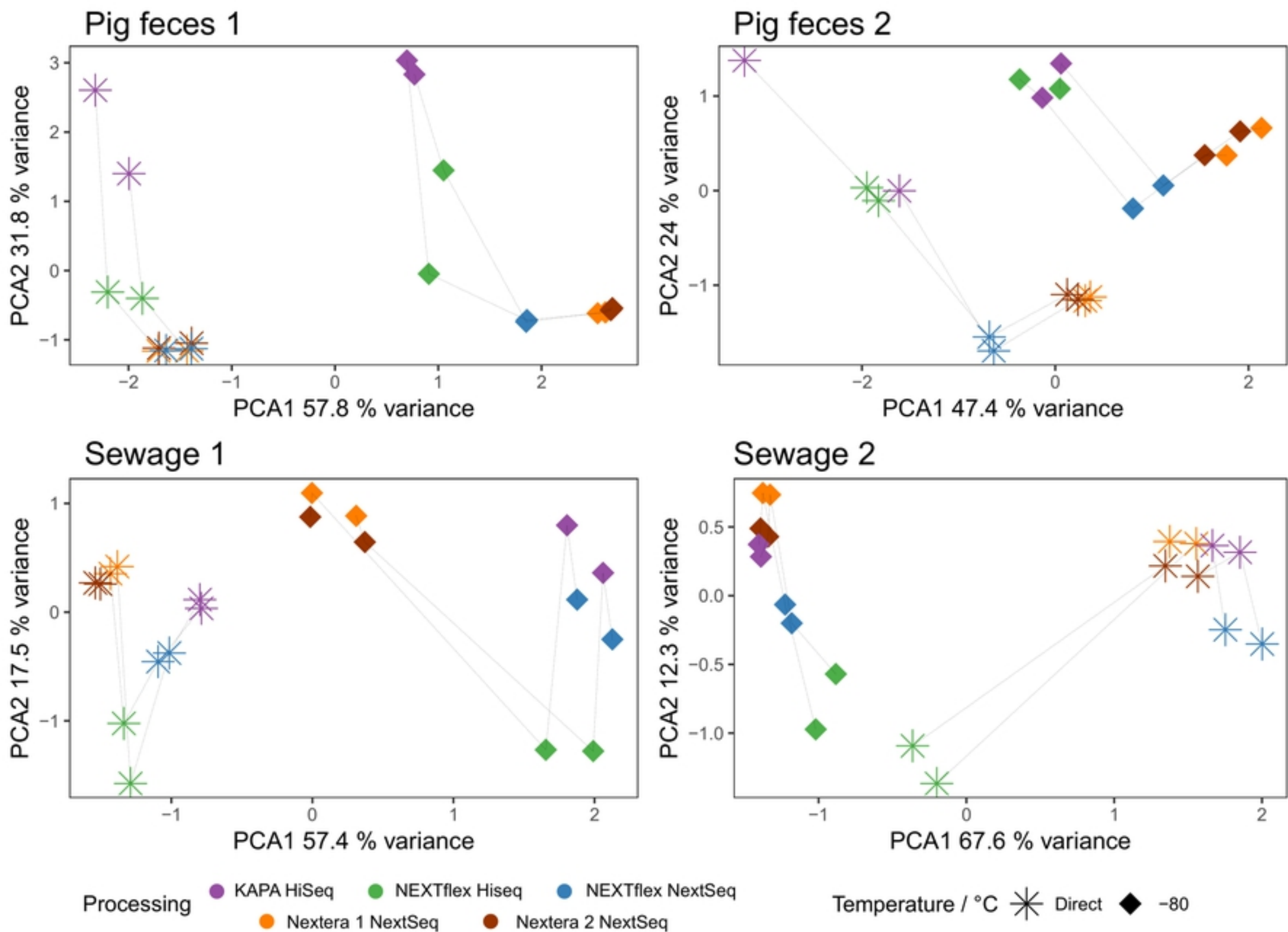


Figure 2

