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

Casper S. Poulsen[1]*, Sünje J. Pamp[1], Claus T. Ekstrøm[2] and Frank M. Aarestrup[1]

[1]Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark.

[2]Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark.

* Corresponding author

cspoulsen@hotmail.com (CP)

Abstract

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

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

Introduction

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

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

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

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

Methods

Sample processing

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

Library preparation and sequencing

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

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

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

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

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

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

Bioinformatics and statistical analysis

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

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

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

Results

Quality control of sequencing output

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

Sample processing impact on microbial characterization

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

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

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

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

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

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

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

Sample processing impact on indicator organisms

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

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

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

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

Discussion

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

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

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

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

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

Acknowledgments

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

References

1. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature [Internet]. Nature Publishing Group; 2012;490(7418):55–60. Available from: http://dx.doi.org/10.1038/nature11450

2. Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15(3):382–92.

3. Yu J, Feng Q, Wong SH, Zhang D, Yi Liang Q, Qin Y, et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. Gut. 2017;66(1):70–8.

4. Zeller G, Tap J, Sobhani I, Amiot A, Tap J, Tran Van Nhieu J, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol. 2014;10(11):766–766.

5. Dekker JP. Metagenomics for Clinical Infectious Disease Diagnostics Steps Closer to Reality. J Clin Microbiol. 2018;1–7.

6. Petersen TN, Rasmussen S, Hasman H, Carøe C, Baelum J, Schultz AC, et al. Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance. Nat Publ Gr. 2015;

7. Nieuwenhuijse DF, Koopmans MPG. Metagenomic Sequencing for Surveillance of Food- and Waterborne Viral Diseases. Front Microbiol. 2017;8(February):1–11.

8. Hjelmsø MH, Mollerup S, Jensen RH, Pietroni C, Lukjancenko O, Schultz AC, et al. Metagenomic analysis of viruses in toilet waste from long distance flights—A new procedure for global infectious disease surveillance. PLoS One. 2019;14(1):e0210368.

9. Hendriksen RS, Munk P, van Bunnik B, McNally L, Lukjancenko O, Röder T, et al. Global monitoring of 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.

26. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, et al. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol [Internet]. Nature Publishing Group; 2017;35(11):1069–76. Available from: http://dx.doi.org/10.1038/nbt.3960

27. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, et al. Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. Jansson JK, editor. mSystems [Internet]. 2016 Oct 25 [cited 2018 Jan 2];1(5):e00095-16. Available from: http://msystems.asm.org/lookup/doi/10.1128/mSystems.00095-16

28. Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, et al. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome. 2014;2:1–11.

29. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics [Internet]. 2015 Oct 24 [cited 2018 Jan 2];16(1):856. Available from: http://www.biomedcentral.com/1471-2164/16/856

30. Clooney AG, Fouhy F, Sleator RD, O' Driscoll A, Stanton C, Cotter PD, et al. Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. White BA, editor. PLoS One [Internet]. 2016 Feb 5 [cited 2018 Jan 2];11(2):e0148028. Available from: http://dx.plos.org/10.1371/journal.pone.0148028

31. JGI. BBDuk Guide [Internet]. 2019. Available from: https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/

32. Cao KAL, Costello ME, Lakis VA, Bartolo F, Chua XY, Brazeilles R, et al. MixMC: A multivariate statistical framework to gain insight into microbial communities. PLoS One. 2016;

33. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol. 2016;62(8):692–703.

34. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: And this is not optional. Front Microbiol. 2017;8(NOV):1–6.

35. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Math Geol. 2003;35(3):253–78.

36. Egozcue J, Pawlowsky Glahn V, Mateu-Figueras G, Barceló Vidal C. Isometric logratio for compositional data analysis. Math Geol. 2003;35(3):279–300.

37. Filzmoser P, Hron K, Reimann C. The bivariate statistical analysis of environmental (compositional) data. Sci Total Environ [Internet]. Elsevier B.V.; 2010;408(19):4230–8. Available from: http://dx.doi.org/10.1016/j.scitotenv.2010.05.011

38. Aitchison J. The Statistical Analysis of Compositional Data. London: Chapman and Hall; 1986.

39. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;1–12.

40. Segata N. On the Road to Strain-Resolved Comparative Metagenomics. mSystems. 2018;3(2):1–6.

41. Escobar-Zepeda A, Vera-Ponce De León A, Sanchez-Flores A. The Road to Metagenomics: From

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

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

Supporting information

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

S2 Fig. Rarefaction curves.

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

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

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

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

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

S1 File. Bioanalyzer electropherograms.

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

S3 File. R markdown file containing the R analysis.

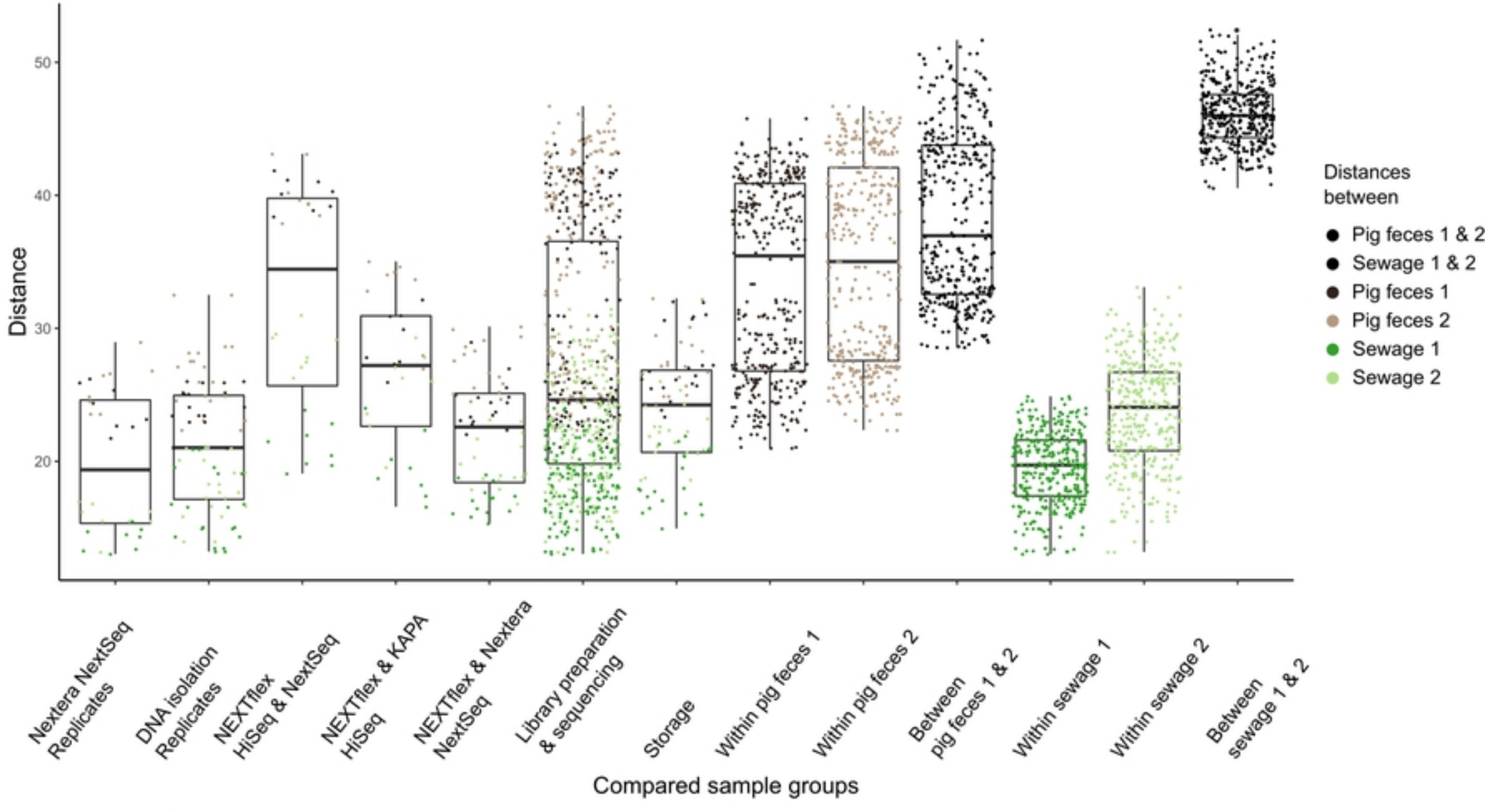
Distance
50
40
30
20
Distances between
Pig feces 1 & 2
Sewage 1 & 2
Pig feces 1
Pig feces 2
Sewage 1
Sewage 2
Nextera NextSeq Replicates
DNA isolation Replicates
NEXTflex HiSeq & NextSeq
NEXTflex & KAPA HiSeq
NEXTflex & Nextera NextSeq
Library preparation & sequencing
Storage
Within pig feces 1
Within pig feces 2
Between pig feces 1 & 2
Within sewage 1
Within sewage 2
Between sewage 1 & 2
Compared sample groups

Figure 1

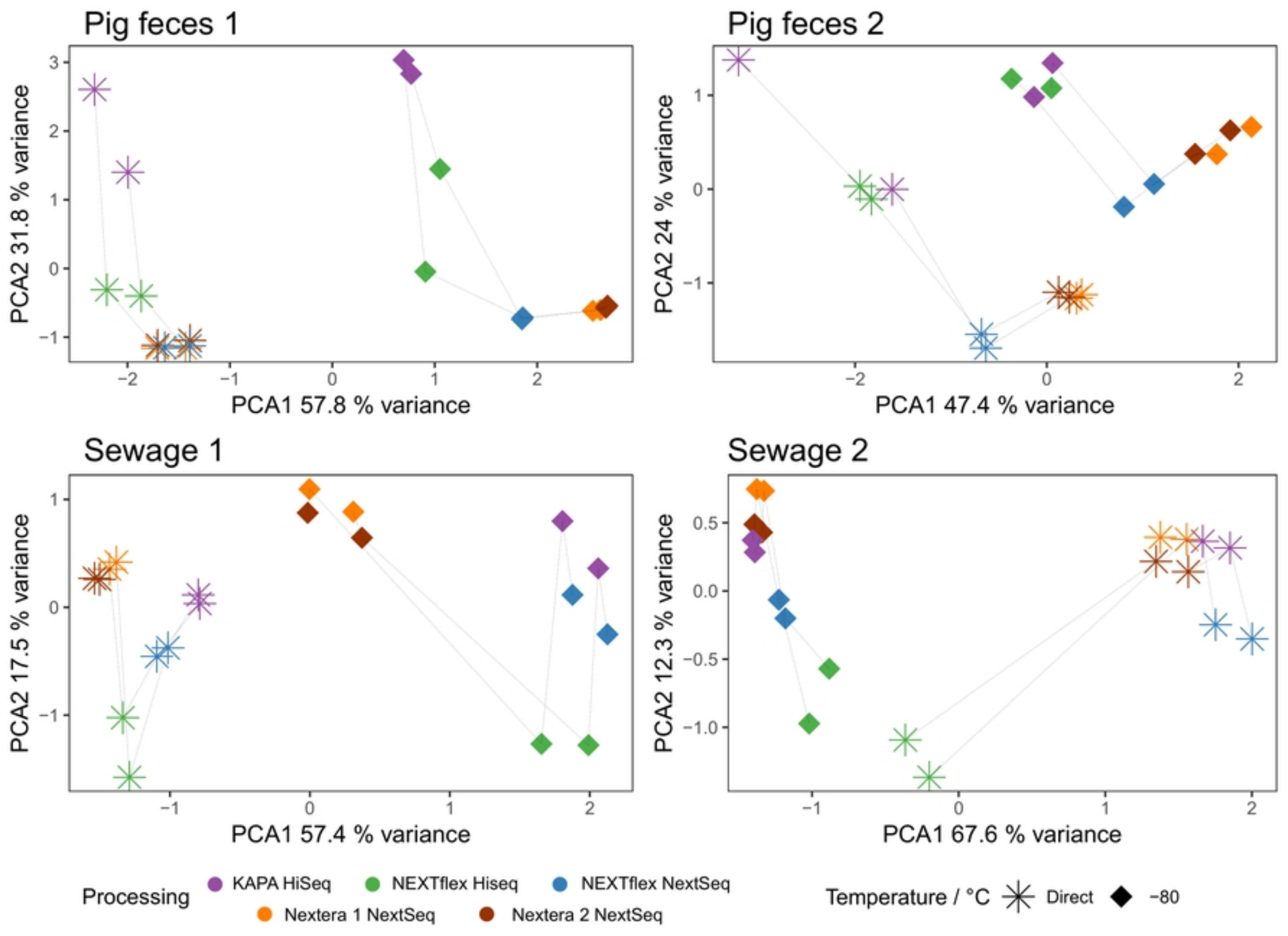
Pig feces 1
PCA2 31.8 % variance
PCA1 57.8 % variance
Pig feces 2
PCA2 24 % variance
PCA1 47.4 % variance
Sewage 1
PCA2 17.5 % variance
PCA1 57.4 % variance
Sewage 2
PCA2 12.3 % variance
PCA1 67.6 % variance
Processing
KAPA HiSeq
NEXTflex Hiseq
NEXTflex NextSeq
Nextera 1 NextSeq
Nextera 2 NextSeq
Temperature / °C
Direct
−80

Figure 2

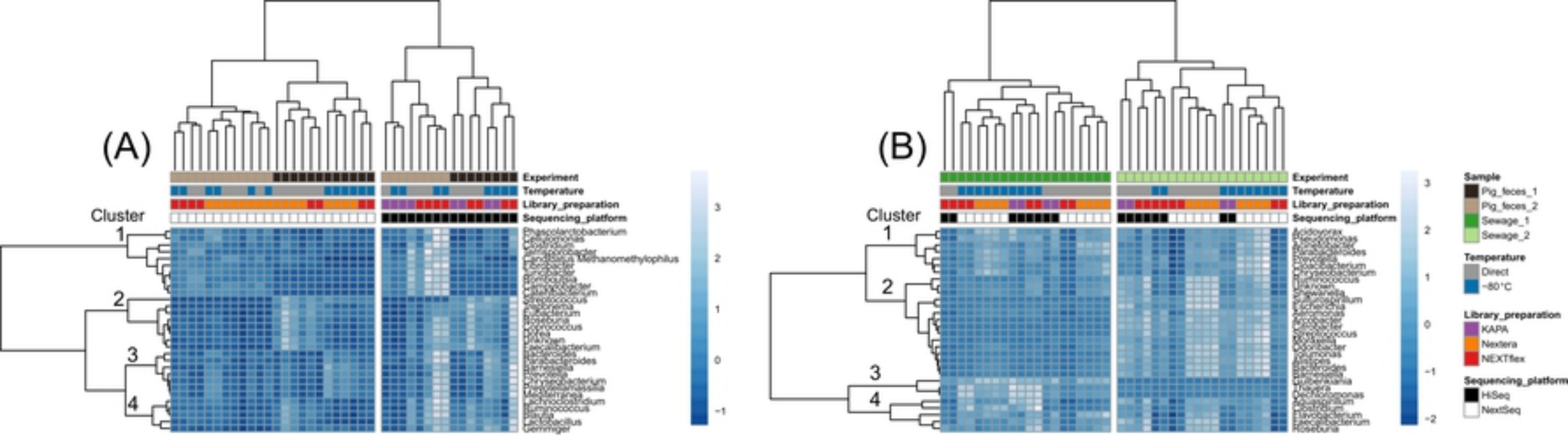
(A)
Cluster
1
2
3
4
Experiment
Temperature
Library_preparation
Sequencing_platform
Phascolarctobacterium
Methanomethylophilus
3
2
1
0
-1
(B)
Cluster
1
2
3
4
Experiment
Temperature
Library_preparation
Sequencing_platform
3
2
1
0
-1
-2
Sample
Pig_feces_1
Pig_feces_2
Sewage_1
Sewage_2
Temperature
Direct
-80°C
Library_preparation
KAPA
Nextera
NEXTflex
Sequencing_platform
HiSeq
NextSeq

Figure 3