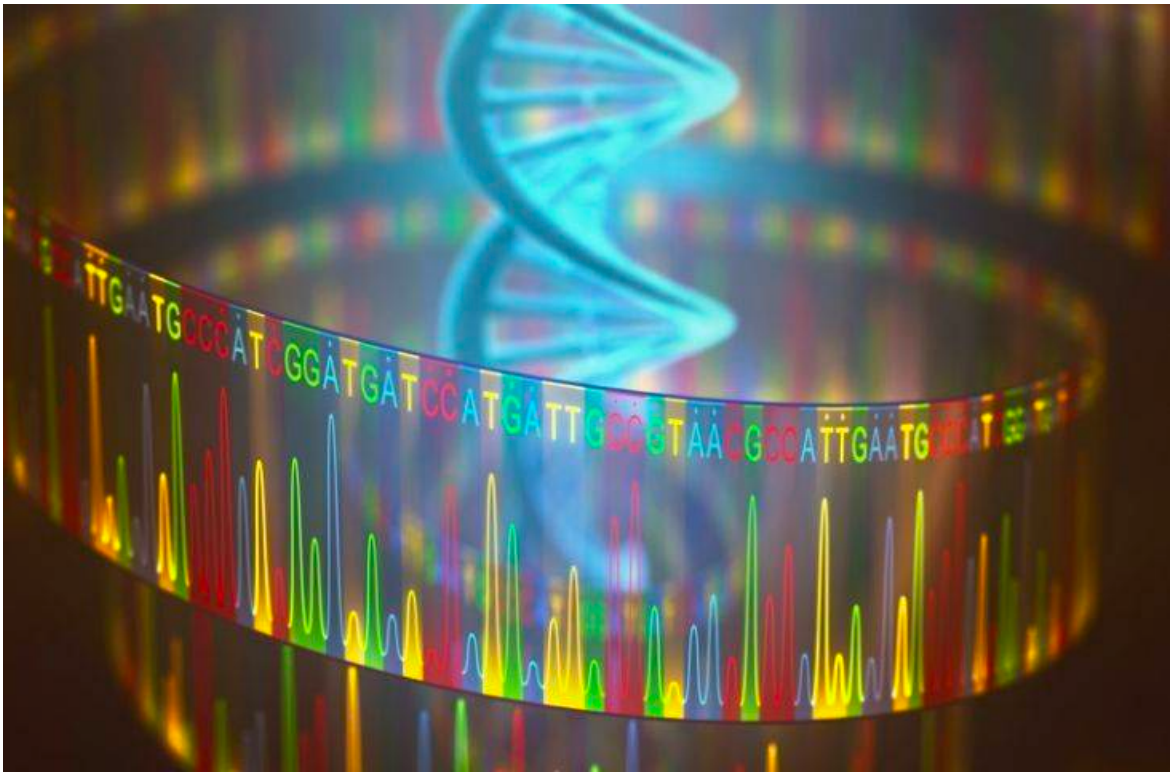
Towards tailored mock communities for quality-controlled microbiome profiling in water technology



Stijn Teunissen (000024753)

Supervisor Van Hall Larenstein: Wouter Suring & Rik veldhuis

Supervisor Wetsus: Pieter van Veelen

Date & place: 5/28/2024, Leeuwarden

Towards tailored mock communities for quality-controlled microbiome profiling in water technology

Student name: Stijn Teunissen (000024753)

Supervisor Van Hall Larenstein: Wouter Suring & Rik veldhuis

Supervisor Wetsus: Pieter van Veelen

Education and school: Biotechnology on Van Hall Larenstein, Leeuwarden

Company: Wetsus, Leeuwarden

Date & place: 5/28/2024, Leeuwarden

Source image: (Bron van Plaatje op voorpagina)

Preface

In this study, I investigated whether there is a relationship between DNA concentration and the level of cross-contamination among samples and mock communities using published sequencing data. This research was conducted as a graduation internship for my Biotechnology program with a Major in Biomedical Research at Van Hall Larenstein University of Applied Sciences in Leeuwarden. The study was conducted at Wetsus in Leeuwarden from February 2024 to July 2024.

During my education, I gained substantial theoretical and practical knowledge. I followed elective courses in biological data science, which I could utilise and expand upon during my internship at Wetsus. Wetsus provided an excellent learning environment. I was able to further develop personally by gaining additional bioinformatics/data science experience and learning a lot about report writing and using the English language.

For my graduation internship, I would like to thank Pieter van Veelen, who supervised me during my internship and from whom I learned a great deal. I would also like to thank Jippe Silvius, the bioinformatician at Wetsus, who also assisted me during my assignment. It was a great collaboration among the three of us, and I could always come to them with questions or clarifications. I also thank the authors who responded to my emails requesting additional data. Lastly, I want to thank Wouter Suring for the guidance from the school during my internship.

Enjoy reading my research report.

Stijn Teunissen

Leeuwarden, 28 May 2024

Samenvatting

Abstract

High-throughput sequencing of 16S rRNA gene amplicons (16S-seq) has become a widely deployed method for profiling complex microbial communities. However, technical pitfalls related to data reliability and quantification remain to be fully addressed. In this work, we have developed and implemented a set of synthetic 16S rRNA genes to serve as universal spike-in standards for 16S-seq experiments.

Table of Contents

1. Introduction

Water is essential in our daily lives, both for household use and industrial processes. Therefore, ensuring water quality and safety is of utmost importance for public health. Wetsus is a research institute that develops modern water treatment technologies. These innovations contribute to solving global water problems. Next-generation sequencing is a method that creates new opportunities for assessing water quality by analysing microbial ecosystems in water.

1.1. Next-generation sequencing (NGS)

Sequencing is the process of determining the order of nucleotides (sequence) in the entire genome or targeted regions of DNA. With next-generation sequencing (NGS), massively parallel sequencing can be performed, resulting in high throughput processing. This high throughput capability allows for the efficient and rapid analysis of complex biological samples of a wide range of ecosystems, including those relevant to water technology. NGS can be applied to small, targeted regions or the entire genome through various methods. (Tan et al., 2015).

In this study, amplicon sequencing is utilised. Amplicon sequencing is a targeted approach that enables the analysis of genetic variation in specific genomic regions. The deep sequencing of PCR products (amplicons) allows for efficient identification and characterisation of amplicon variants. A typical application of amplicon sequencing is sequencing the bacterial 16S ribosomal RNA gene (16S) for profiling bacterial and archaeal communities. An advantage of amplicon sequencing is, for example, that its reduced cost and fast turnaround times of sequencing enable larger experiments and sample sets to be analysed, compared to approaches such as whole-genome sequencing (Ranjan et al., 2016) (*Amplicon Sequencing Solutions*, n.d.).

1.2. 16s ribosomal RNA

The prokaryotic 16S ribosomal RNA (16S) gene is approximately 1500 base pairs long, with nine variable regions (V1-V9) interspersed between conserved regions. The 16S gene is present in nearly all bacteria, allowing the variable regions of the 16S gene to identify microbial diversity (Caporaso et al., 2011) (*16S RRNA Sequencing*, n.d.).

1.3. NGS-workflow

NGS is performed in several steps, referred to as the NGS workflow. This NGS workflow consists of Polymerase chain reaction (PCR) amplification, library preparation, cluster generation, sequencing on various platforms (such as Illumina, Ion Torrent, PacBio, and Oxford Nanopore), and data analysis. In this study, Illumina sequencing technology is utilised for its high-throughput capability and accuracy in analysing complex DNA libraries, particularly for profiling microbial communities. Illumina sequencing enables the sequencing of small reads of 50-300 base pairs (bp) with low error rates, making it suitable for applications such as 16S gene analysis. Before the NGS workflow begins, DNA must be extracted from the sample. Once the DNA is extracted, the NGS workflow starts with amplifying a specific gene, the 16S gene, using PCR with specific primers for the 16S gene. This primer set may differ for each analysis where the primers bind to the nine regions of the 16S gene. The 16S gene contains sufficient variation among bacteria to distinguish up to the genus level. After the PCR is performed, the amplified PCR products are used in library preparation. Library preparation is crucial for NGS, preparing DNA samples for sequencing analysis. During the library preparation, adapters and index (barcode) are ligated to the DNA fragments at the 5' and 3' ends (figure 1A) (Illumina, n.d.).

The next step is cluster generation, where the library is loaded onto a flow cell, allowing DNA fragments to bind complementary to oligos on the surface of the flow cell. The flow cell ensures that DNA fragments are presented in a structured and controlled manner for sequencing analysis. Each DNA fragment is then amplified into several clonal clusters through bridge amplification, resulting in millions of copies of single-stranded DNA (figure 1B). Once cluster generation is complete, it is ready for sequencing. Bridge amplification is an amplification reaction that occurs on the surface of an

Illumina flow cell. The surface is coated with a lawn of oligonucleotides during flow cell manufacturing. In the first step of bridge amplification, a single-stranded sequencing library (with complementary adapter ends) is loaded into the flow cell. Individual molecules in the library bind to complementary oligos as they 'flow' across the oligo lawn. Priming occurs as the opposite end of a ligated fragment bends over and 'bridges' to another complementary oligo on the surface. Repeated denaturation and extension cycles (like PCR) result in localised amplification of single molecules into millions of unique clonal clusters across the flow cell. This process is called 'clustering' (Illumina, n.d.).

Illumina sequencing uses the sequencing by synthesis (SBS) process; chemically modified nucleotides

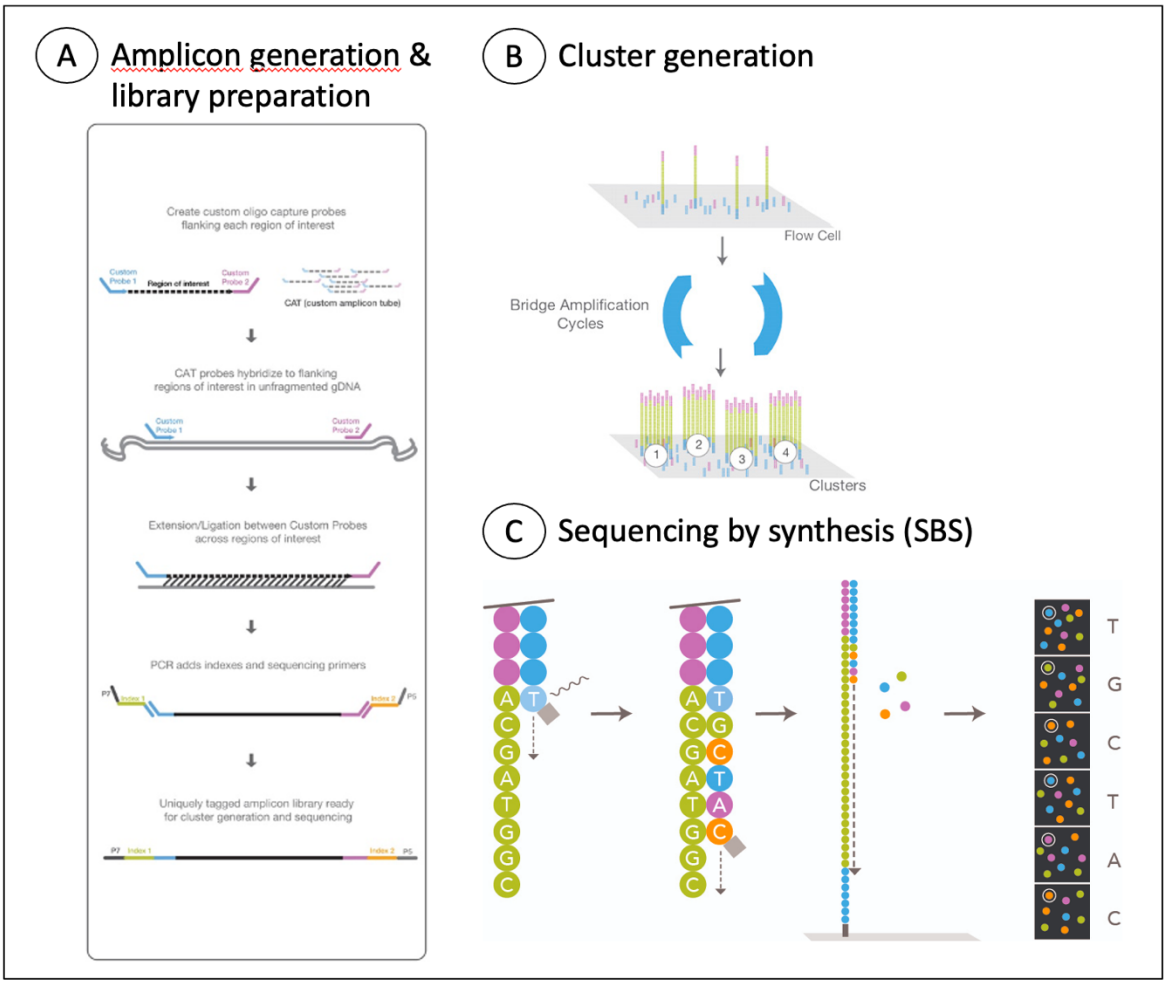


Figure 1 Amplicon sequencing workflow overview - Illumina sequencing includes three steps. (A) Amplicon generation & library preparation, (B) Cluser generation, and (C) Sequencing by synthesis (SBS). (Illumina, n.d.)

bind to the DNA fragments. Each nucleotide contains a fluorescent label and a reversible terminator that blocks the incorporation of the next base. When a nucleotide binds, the fluorescent signal indicates which nucleotide has been added, and the terminator is cleaved, allowing the next base to bind (figure 1C). After the forward DNA fragment is read, the reads are washed away, and the process is repeated for the reverse DNA fragment. This is also known as paired-end sequencing (Illumina, n.d.).

Once sequencing is complete, data analysis can be performed. The unique identified sequences are compared to a database with reference sequences with known taxonomic affiliation during data analysis. The sequence count data, combined with taxonomic information, differential abundance, bar plots/heatmaps, diversity analysis including alpha and beta diversity, and statistical testing can be performed. This can be done using tools such as QIIME2 and R.

1.3.1. Multiplexing

In a single run, multiple samples can be sequenced simultaneously, known as multiplexing. Multiplexing enables the pooling and sequencing of multiple libraries (samples) simultaneously during a single sequencing run, offering benefits such as increased efficiency, reduced cost per sample, and optimised resource utilisation.

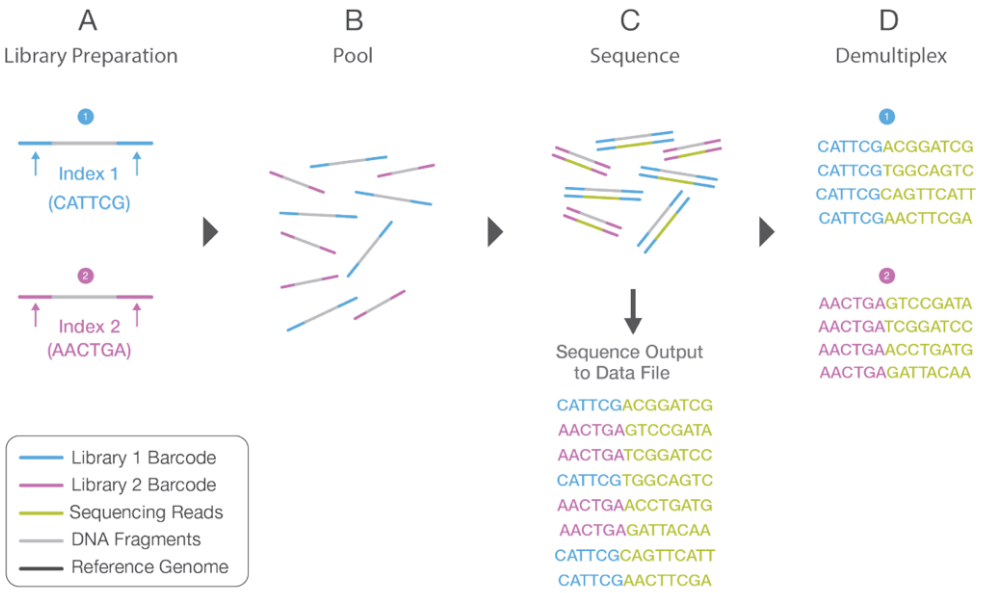


Figure 2 Multiplexing - Multiplexing of multiple libraries. (A) Unique index sequences are assinged to each DNA fragment during library preparation. (B) Multiple libraries are pooled together. (C) Libraries are sequenced. (D) The sequence data is demultiplexed. (Illumina, n.d.)

Unique 'barcodes' (index) sequences are assigned to each DNA fragment during library preparation (figure 2A). Each 'Barcode' has a known sequence. Multiple libraries are pooled together (figure 2B), and the pooled library is then sequenced (figure 2C). However, 'index hopping', a phenomenon where sequencing reads are incorrectly assigned to the wrong sample, can occur and poses a challenge in multiplexed sequencing experiments. The sequence data needs to be identified and sorted, known as demultiplexing (figure 2D). Challenges in accurately assigning reads to their respective samples require bioinformatics tools for accurate data analysis (Illumina, n.d.).

1.4. Mock Community

Mock communities are artificial microbial communities constructed from bacterial strains or whole cells mixed in known quantities and proportions. These properties make mock communities suitable to serve as positive controls to microbiome sequencing experiments, providing insights into the precision of estimated taxa proportions. Mock communities also enable estimating the proportion of false positive detections from cross-contamination among samples. Mock communities are commercially available for microbiome analyses from providers such as ZymoBIOMICS, the NITE Biological Resource Center (NBRC), the American Type Culture Collection (ATCC), and others. However, a commercial mock community is not always ideal for every study for several reasons. First, they may not represent the correct taxonomy relevant to the environmental setting in which they are used. Second, these mock communities can be expensive. Therefore, it is also possible to create your own mock community in-house (Colovas et al., 2022).

Various mock communities are used in research, each designed to meet specific experimental needs. Equal mock communities, where the proportions of bacteria are evenly distributed. Each microbial

strain in these communities is present in roughly equal quantities, which allows for straightforward comparative analysis across different microbial types (Tourlousse et al., 2022).

In contrast, Log10 mock communities distribute bacteria on a logarithmic scale based on powers of ten. This arrangement is particularly useful for testing and calibration across a broad range of bacterial concentrations, from very high to very low, simulating more complex and variable natural environments (Tourlousse et al., 2022).

Similarly, Log2 mock communities distribute bacteria on a logarithmic scale using powers of two. This provides a finer gradation in concentration levels than what is offered by the Log10 scale, offering a different perspective on microbial dynamics (Tourlousse et al., 2022) (Colovas et al., 2022)

1.5. Contamination during Sequencing

During microbiome profiling, two main types of contamination can arise. Contaminant DNA can originate from many sources and cross-contamination among multiplexed samples. The study by Salter et al., (2014) revealed that bacterial DNA contamination is present in DNA extraction kits and laboratory reagents, which can significantly impact the results of microbiota studies, especially in samples with low microbial biomass. Contamination during DNA extraction is another common source of DNA contamination. Cross-contamination between samples can occur during sample processing and

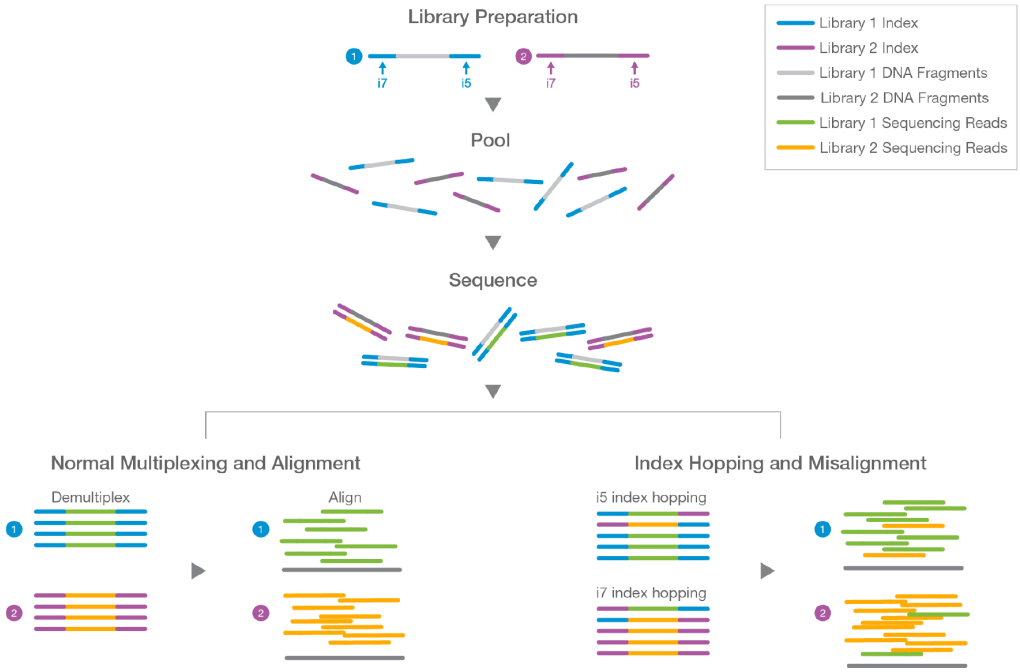


*Figure 3 Index hopping - Shows the library preparaion process, sample pooling, sequencing, and indexing for two individual samples. The difference between index hopping and normal multiplexing is shown. (*What Is Index Hopping?*, n.d.)*

DNA extraction. Cross-contamination during library preparation, often outsourced at sequencing facilities, is another challenge. Barcode cross-contamination, for example, can occur when barcodes end up in incorrect wells or tubes, causing barcodes to bind to the wrong samples, also known as 'tag-switching'.

Cross-contamination can also occur during the sequencing process if a barcode is assigned to the wrong sample, referred to as 'index hopping' (figure 3). Contamination during sequencing is a problem in 16S sequencing, where PCR is performed, but also in shotgun metagenomics, where no PCR is performed. DNA contamination and cross-contamination can critically impact sequence-based microbiome analyses, potentially leading to erroneous conclusions and misleading interpretations of microbial community compositions and diversity (Salter et al., 2014) (Eisenhofer et al., 2019).

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7849411/ contamination

1.6. Project aims

This project aims to quantify the degree of cross-contamination as a function of DNA concentration difference between samples by leveraging the known information in mock communities in published sequencing data. The goal is to establish if a tailored mock community in terms of DNA concentration (difference in DNA concentration towards zero, by normalising to the DNA concentration of a sample set) outperforms the current standard use of mock community DNA.

The approach to achieve these project aims is to collect raw sequence data and sample DNA concentrations from published studies in water technology, environmental technology, biotechnology, and other relevant fields. These studies should have raw sequence data available, include mock communities, and provide metadata regarding the DNA concentration of samples. If necessary, data requests are sent to authors to complete missing data. Additionally, unpublished data from Wetsus is used. Bioinformatic data analysis, including quality control and filtering, is conducted once the data is collected.

In this study, the final data set consists of seventeen mock samples from four different mock communities (in-house Wetsus 2022, in-house Wetsus 2023, Zymo Log10, and Zymo Equil) from thirteen different studies, of which eleven studies are from Wetsus, and two are from published articles. The final data set is then used to determine the amount of contamination, examine whether the amount of contamination in mock communities is influenced by the abundance and prevalence of sample genera, and estimate the cross-contamination effects of DNA concentration differences between samples and mock communities.

2. Methods

This study uses the software QIIME2 (version 2022.11) (Bolyen et al., 2019) and R (version 4.3.2) (*R: The R Project for Statistical Computing*, n.d.)to analyse the raw and processed published sequencing data. The data analysis workflow is shown in Figure 4. QIIME 2 is a powerful bioinformatics tool explicitly designed for microbiome analysis, offering a wide range of features for processing, analysing, and visualising NGS data.

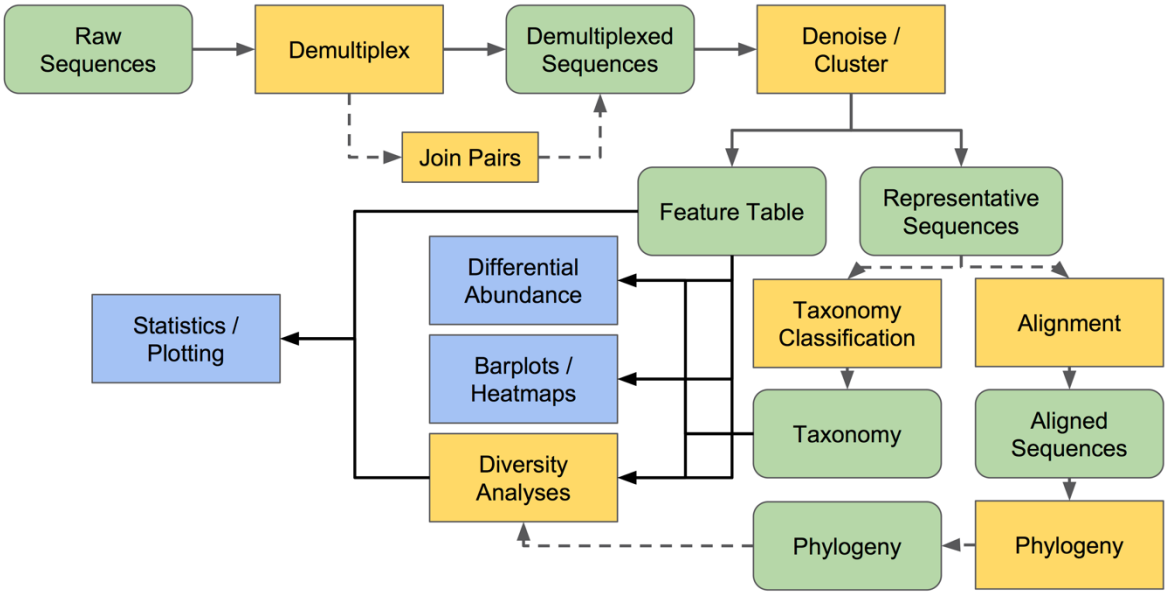


*Figure 4 Data analyse workflow - The workflow to analse raw sequence data using qiime and R. (*QIIME 2*, n.d.)*

In this study, QIIME2 is used for demultiplexing, denoising/clustering, and taxonomy classification. In Figure 4, these are represented by the yellow blocks, with the green block as the outcome. R is a versatile programming language and environment commonly used for statistical computing and graphics, providing flexibility and customisation for data analysis. R is used for statistical analysis, generating bar plots/heatmaps, and hypothesis testing, as depicted by the blue blocks in Figure 4 (Bolyen et al., 2019b).

2.1. Sequencing data

The sequencing data consisted of studies conducted within Wetsus and published data from articles. The published sequencing data was collected by searching for articles sequenced at the 16S rRNA gene level and utilising a mock community. The authors of these articles were contacted and asked to provide the missing DNA concentration data if it was not already available. The sequencing data was downloaded from databases such as NCBI SRA and subsequently analysed using QIIME2.

2.2. Qiime2 analyse

2.2.1. Denoising & clustering

The published sequencing data is already demultiplexed, so after downloading and uploading the sequencing data from NCBI SRA to the cluster. The process can move on to denoising/clustering. This is done using the package DADA2 in QIIME2. During denoising/clustering, filtering is performed on noisy sequences which contain errors due to sequencing, such as incorrect nucleotide assignment. This filtering is done using quality scores, also known as Phred scores. Each read receives its own score from

0 to 40, with 40 being the best. For example, a score of 10 means there is a 90% chance that the sequencer correctly identified the base; at a score of 40, this increases to 99.99%. Therefore, a median of 30 is used as the lower threshold.

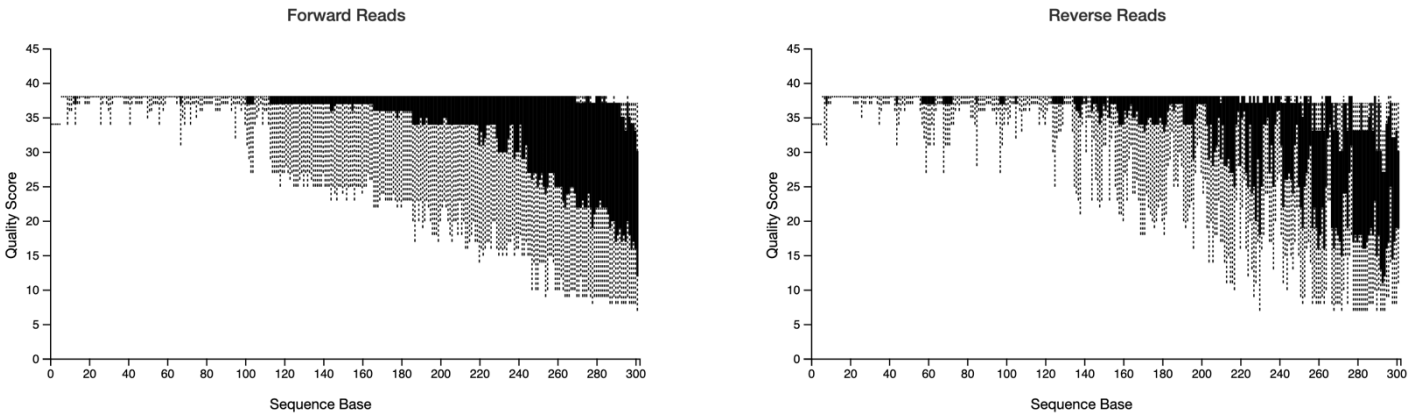


Figure 5 Phred scores – The Phred scores for each base of the forward and reverse reads.

Figure 5 illustrates an example of the quality scores for the forward and reverse reads. The reads are trimmed at the same beginning and end to maintain consistency. These locations are chosen because the quality typically deteriorates towards the end of the read, and the quality at the beginning is often slightly poorer. Therefore, trimming the first five bases at the beginning, 200 at the end for forward reads, and 190 for reverse reads has been decided. The reverse reads always have lower quality, primarily because the reagents and additives are depleted towards the end of sequencing. Therefore, the reverse read is always trimmed slightly earlier.

During the denoising process, replicates are also removed. These are read with exactly the same sequences, which often occur during PCR amplification when the same fragment appears multiple times in the library or binds to multiple spots on the flow cell.

During clustering, sequences that are identical or very similar (more than 97% similarity) are merged into single representative sequences. This process is also known as OTU (Operational Taxonomic Unit) picking. The final product of denoising and clustering is a feature table and representative sequence artefacts. These two of the most important artefacts in an amplicon sequencing workflow are used for many downstream analyses.

bron (Inferring microbial co-occurrence networks from amplicon data: a systematic evaluation)

2.2.2. Taxonomy classification

It is crucial to identify which organisms are present in the samples. This can be done by comparing the feature artefacts against a database containing sequences with known taxonomy. This comparison can be made in two ways: against an untrained database or using a trained classifier. These classifiers are trained to recognise which characteristics best distinguish each taxonomic group, adding an extra step to the classification process. The classifier is trained using the 16S gene. A classifier needs to be trained only once and can then be reused as often as necessary. The advantage of a trained classifier is that it is more specific in identifying organisms. The trained classifier was downloaded from Qiime2 and developed using the SILVA database (Quast et al., 2013) (Robeson et al., 2020) (Bokulich et al., 2018). It was decided to use a single classifier for the 16S gene rather than training individual classifiers for each unique primer set. This approach was chosen because various primer sets were used across different studies, and using one classifier helps maintain consistent results while saving considerable time during analysis.

2.3. R analyses

Once the reads had been analysed in QIIME2, the following files were needed to further analyse them in R: feature table artefact, phylogenetic diversity artefact, taxonomy artefact, and the metadata text file. These files were loaded and combined into a Phyloseq object using the "qiime2R" and "Phyloseq" packages. Bron package From the Phyloseq object, unwanted taxa such as Mitochondria, Chloroplasts, Unclassified Kingdom, Eukaryota, etc., were first filtered out. All incompletely classified OTUs were also supplemented. This meant that if an OTU was not classified at the genus level but was classified at the family level, it was labelled as "genus of family" at the genus level.

2.3.1. Measured vs theoretical

After loading the necessary files, the R analysis was performed. First, it was checked whether the measured abundance of the mock communities used in this study matched the theoretical abundance. The mock communities used in this study were in-house Wetsus 2022, in-house Wetsus 2023, Zymo equil, and Zymo Log10 mock communities. For each mock community, a tibble was created with the known bacterial species and their corresponding theoretical abundance. All mock samples from the different mock communities were filtered from the loaded sequence data and their associated measured abundance. Then, each mock community's measured and theoretical abundances were plotted against each other.

2.3.2. Contamination in Mock Communities

To determine if contamination was present in the mock samples, the mock sample was filtered out of the study. Then, all genera that were not expected in the mock were grouped and renamed to "other genera." This was done to easily distinguish between contamination and the genera expected to be present in the mock community. This process was performed for each study with different mock communities, which were then grouped. The abundance per genus and the mutated "other genera" were then displayed in a bar plot, with "other genera" in grey at the bottom to make the contamination easily visible.

2.3.3. Relationship between contaminations in mock communities and the abundance/prevalence of samples

To determine if OTUs with high abundance or prevalence in the samples were more likely to cause cross-contamination between samples and mock communities or if this had no impact, the prevalence of each sample OTU per study was calculated. Prevalence indicated whether a bacterium occurred in all samples or only in some. For example, a prevalence of 1 meant that the bacterium type was found in all samples, while a prevalence of 0.5 indicated that the bacterium was present in only half of the samples. Therefore, prevalence was examined to see if it affected contamination. The samples' maximum, mean, and median abundance were calculated per study. This was done to determine if the abundance of the samples influenced the contamination in the mock communities. The maximum, mean, and median abundances were chosen to prevent outliers from distorting the results.

For each study, the mock samples were filtered out, and the known genera were removed from these mock samples, leaving only the contamination OTUs. Then, the prevalence, maximum, mean, and median abundances were added to the filtered mock samples. Subsequently, the percentage of sample OTUs in the mock community was plotted against the prevalence, maximum, mean, and median abundance and statistically tested using ANOVA with the lmtest function.

2.3.4. Relationship between contamination in mock communities and DNA concentration

To determine if the slope of the plot where the percentage of sample OTUs in the mock community was plotted against the value (prevalence, maximum, mean, and median abundance) depended on the DNA concentration ratio and difference, the maximum, mean, and median DNA concentrations of the samples and mocks were calculated. The trend was determined by examining the relationship between the percentage of sample OTUs in the mock community and the value (prevalence, maximum, mean, and median abundance), depending on the study and sample characteristics (prevalence, maximum, mean, and median abundance). Using emtrends to calculate the slope of each combination of study and sample characteristics. The DNA concentration ratio and difference were also calculated. For the ratio, the mock DNA concentration was divided by the sample DNA concentration, and for the difference, the sample DNA concentration was subtracted from the mock DNA concentration. Two plots were created: one where the trend was plotted against the DNA concentration ratio and another where the trend was plotted against the DNA concentration difference.

2.4. Data Availability

All scripts created and used in this research have been uploaded to GitHub (https://github.com/stijnteunissen/microbialanalysis.git)

3. Results

This study examined whether contamination between samples and mock communities occurred during sequencing and quantified the degree of cross-contamination as a function of DNA concentration difference between samples and mock communities in published sequencing data. This was explored by first checking whether the proportions of bacteria in mock communities were the same between theoretical and measured abundance. Subsequently, it was determined what percentage of the mock community consists of contamination. Finally, the study examined whether there is a correlation between the detected contamination and the DNA concentration difference between samples and mock communities.

3.1. The measured abundance of genera in the mock samples matches the theoretical abundance

Various mock communities were utilised: in-house Wetsus 2022, in-house Wetsus 2023, Zymo equil, and Zymo Log10 mock communities. For each of these communities, there was a theoretical expectation of the abundance of the present bacteria based on the design data of the mock community. These theoretical abundances were then compared with the measured abundances to evaluate the accuracy of each type of mock community. This was visualised in Figure 6, where the theoretical and measured abundances for every kind of mock community are plotted against each other. This provides insight into the reliability of the mock communities.

…

Figuur 6

Bijschrift: Theoretical vs measured abundance – The theoretical versus measured abundances of 17 mock samples in four different mock communities. A) the in-house Wetsus 2022 mock community, B) the in-house Wetsus 2023 mock community, C) the Zymo Log 10 mock community, and D) the Zymo Equil mock community.

…

The different mock communities have been used multiple times across various studies. In Figures 6A-D, the differences between mock samples within a mock community are illustrated. Most points are not far apart and are close to the theoretical abundance, sometimes slightly above or below it. The bacteria that deviate the most from the theoretical abundance is *Brevibacillus* (Figure 6B), with a theoretical abundance of 50.78% and a measured abundance between 38.89% and 41.62%. *Sphingobium* also deviated from the theoretical abundance of 3.18% and a measured abundance between 6.84% and 11.17%. In Figueres 6A-B in-house Wetsus 2022 and 2023 mock communities, the lines are very close to each other. The Zymo log 10 mock community (Figure 6C) has been used only once across different studies but almost matches the theoretical abundance. The Zymo equil mock community (Figure 6D) shows the most variation between the mock samples. This mock community was used in internal Wetsus studies but also included data from external studies. The figure also illustrates that *Salmonella* is only found in the MCZ mock sample and not in the other mock samples. On the other hand, Lactobacillus is found in the other mock samples but not in the MCZ mock sample. The MCZ mock sample was used in an internal Wetsus study, while the other mock samples came from published external studies.

3.2. Contamination in Mock Communities

The mock communities should consist only of bacteria that have been intentionally included. If bacteria from other genera are found in the mock community, this is considered contamination. Figure 7 illustrates the percentage of the mock communities that consist of other genera (contamination) in grey.

…

Figuur 7

Bijschrijft: Summary of 16S rRNA gene sequencing taxonomic classification – bar plot of the different types of mock communities (in-house Wetsus 2022, in-house Wetsus 2023, Zymo Equil, and Zymo Log10). Each column represents a mock sample.

…

In Figure 6, it is discussed that the points of the in-house Wetsus 2022 and 2023 mock communities are close to each other and deviate only slightly from the theoretical abundance. In the bar plot of Figure 7, the in-house Wetsus 2022 and 2023 mock samples are consistent across different studies. It is also shown that a portion of the mock community consists of contamination, indicated in grey, ranging from 5.56% to 10.90% of the mock samples.

The Zymo log 10 mock community has been used only once, so it cannot be determined if it remains consistent across different studies, but there is 4.73% contamination visible in grey.

For the Zymo equil mock, Figure 6 already illustrates many variations between the different mock samples, which is also reflected in the bar plot (Figure 7). The bar plot shows that Salmonella is found only in mock sample MCZ, while Lactobacillus is found in the other mock samples but not in mock sample MCZ. This was also observed in Figure 6. Additionally, the bar plot shows that the contamination ranges from 0.75% to 20.89%.

3.3. Median abundance and prevalence affect the percentage of cross-contamination

Figure 7 illustrates the presence of contamination within the mock communities. This contamination is examined to determine if there is a relationship between the percentage of sample OTU in the mock communities and the genus abundance in the sample or the genus prevalence in the sample. This is shown in Figure 8.

…

Figure 8

Caption: Relationship between OTU abundance, prevalence, and cross-contamination – A scatterplot displaying the percentage of sample OTU in mock communities plotted against the values (maximum, mean, median abundance and prevalence). Each data point represents an OTU, with different studies distinguished by colours and sample characteristics (maximum, mean, median abundance, and prevalence) indicated by shapes.

…

The median abundance and prevalence are illustrated in Figure 8. The maximum and mean abundance are illustrated in Figure 11 in the appendix. The y-axis shows the percentage of sample OTU in the mock community, and the x-axis displays the values (percentage abundance and prevalence). Colours indicate the different studies, and the sample characteristics (maximum, mean, median abundance, and prevalence) are indicated by shapes. In the figure, each point represents an OTU, with a line drawn through the OTUs for each study separately. Most maximum, mean, and median abundance lines increase linearly, while the prevalence lines increase less steeply.

Based on the result from Figure 8 and the statistical test, it can be inferred that median abundance significantly affects the percentage of sample OTU in the mock community, which varies significantly per study (ANOVA: $F_{12, 56776} = 14{,}98$, $p < 10^{-15}$). Prevalence also significantly affects the percentage of sample OTU in the mock community, and this also varies significantly per study. The prevalence effect is significant but less strong than the median abundance (ANOVA: $F_{12, 54939} = 1{,}91$, $p = 0{,}029$).

3.4. Differences in DNA concentration affect the percentage of cross-contamination

The data presented in Figure 8 indicates a relationship between median abundance and the percentage of cross-contamination, as well as a relationship between prevalence and the percentage of cross-contamination. With this information, it is examined whether this is also associated with the DNA concentration ratio and difference. This is illustrated in Figures 9/10 and 12/13 in the appendix.

…

Figure 9

Caption: Relationship between trend and DNA concentration ratio – A scatterplot displaying DNA concentration ratio plotted against the trend. Different shapes represent sample characteristics (maximum, mean, median abundance, and prevalence) within each column. The colours represent each row's different DNA characteristics (maximum, mean and median DNA concentration).

….

In Figure 9, the x-axis represents the DNA concentration ratio, and the y-axis represents the trend. The trend is the slope of the lines in Figure 8, where the percentage of sample OTU in mock communities is plotted against the value (maximum, mean, median abundance, and prevalence). Colours indicate characteristics such as maximum, mean, and median DNA concentration, and the sample characteristics (maximum, mean, median abundance, and prevalence) are indicated by shapes. The grey areas around the lines represent the confidence intervals.

The median DNA concentration is illustrated in Figure 9. The maximum and mean DNA concentrations are illustrated in Figure 12 in the appendix. There is hardly any difference between the maximum, mean, and median DNA concentrations. The median DNA concentration is the least sensitive to outliers and is illustrated in Figure 9. It can be seen in Figure 9 that the lines are flat and that the confidence intervals are wide and cross zero. This, together with the statistical test that is conducted, indicates that there is no statistical relationship between the DNA concentration ratio and the trend.

…

Figure 10

Caption: Relationship between trend and DNA concentration difference – A scatterplot displaying DNA concentration difference plotted against the trend. Different shapes represent sample characteristics (maximum, mean, median abundance, and prevalence) within each column. The colours represent different DNA characteristics (median DNA concentration) for each row.

….

The median DNA concentration is illustrated in Figure 10. The maximum and mean DNA concentrations are illustrated in Figure 13 in the appendix. It can be seen in Figure 10 that the lines for mean and median abundance slope linearly downward, while the lines for maximum abundance and prevalence are flatter. The confidence intervals for mean and median abundance do not cross zero, while those for maximum abundance and prevalence do. The significance test indicates that the difference in DNA concentration between the mock and the sample is an essential factor influencing the trend, and this effect varies significantly by sample characteristic (ANOVA: $F_{3, 21} = 5{,}22$, $p = 0.008$). There is a significant difference between the slopes of median abundance and prevalence ($p = 0.009$).

4. Discussion

Cross-contamination is a major issue in sequencing studies, especially for sample types that have low microbial biomass. In dit onderzoek wordt daarom onderzocht naar de to quantify the degree of cross-contamination as a function of DNA concentration difference between samples.

https://www.cell.com/trends/microbiology/abstract/S0966-842X(18)30253-1?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0966842X18302531%3Fshowall%3Dtrue

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7849411/

4.1. The measured abundance of genera in the mock samples matches the theoretical abundance

4.2. Contamination in Mock Communities

4.3. Median abundance and prevalence affect the percentage of cross-contamination

4.4. Differences in DNA concentration affect the percentage of cross-contamination

Meseart vs theretical reden dat de lijnene niet precies lopen omdat het elke keer andere onderzoeken zijn naturlijk dusk omen niet precies overeen en door de contaminatie kunnen de verhoudingen verschuiven. Ook zijn er bij de equil versdchillende primers gebruik die waardoor er verschil in clasificatie zit een andere reden kan de gebruikte clasificatie methode zijn omdat deze over hele 16s gen is en niet voor elke primer apart is geclasficeerd omdat zo gelijk mogelijk resultaat te hebben

In de barplot is er dus contaminatie te zien dit kan contamintatie zijn van reagentia en zouy wat blijkt uit een artikel maar kan ook cross contamintatie zijn daar gaan ik naar kijken en versschillen bij de zymo equil kunnen nog steeds uit de zelfde reden bestaan als al genoed door verschil in primers of door de gebruik van de classifier

in figuren 6 en 7 was te zien dat de mock communiets bijna helemaal overeen komen met de theorie. Reden waarom de mock communities niet helemaal overeen komen kan zijn doordat er dus contamintie in de mock aanwezig is waardoor de propoerties veranderen. Een anderdere reden kan ook zijn dat er tijdens het onderzoek een classifier voor het hele 16S gen gebruikt is en niet een clasifier die specifiek voor een primer set is. dit is gedaan om het zo universeel mogelijk te houden omdat er verschillende sooreten primers zijn gebruikt maar hierdoor kan het wel zijn dat er de classificatie iets minder specifiek is. Ook was er te zien dat er in zymo equil mock veel verschil zit tussen de de verschillende mock samples die in verschillende studies is gebruikt. Een van de reden dat er verschil zit in dat er verschillende primer paren zijn gebruikt zo is er voor de sample MCZ de forward

This project aims to quantify the degree of cross-contamination as a function of DNA concentration difference between samples by leveraging the known information in mock communities in published sequencing data. The goal is to establish if a tailored mock community in terms of DNA concentration

(difference in DNA concentration towards zero, by normalising to the DNA concentration of a sample set) outperforms the current standard use of mock community DNA.

5. Conclusion

Het doel van het onderzoek was om te kijken of er een relatie zit tussen de hoeveelehid cross contaminatie en de dna concentraties. er is contaminatie in mock communities gevonden deze contamintie komt van cross contamintie vanuit samples. Belangrijke invloeden van de cross contaminatie zijn de abudanitie van de bacterien in de samples en de dna concentratie tussen mock en sampel met name de dna difference is belangrijk als het verschil tussen mock en smaple dna concentratei zo kelin mogelijk is is de hoeveelheid contaminatie het minst.

6. References

16S rRNA Sequencing. (n.d.). Retrieved February 20, 2024, from https://emea.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/16s-rrna-sequencing.html

Amplicon Sequencing Solutions. (n.d.). Retrieved February 27, 2024, from https://emea.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/amplicon-sequencing.html

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 1–17. https://doi.org/10.1186/S40168-018-0470-Z/TABLES/3

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019a). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology 2019 37:8*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019b). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology 2019 37:8*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Colovas, J., Bintarti, A. F., Mechan Llontop, M. E., Grady, K. L., & Shade, A. (2022). Do-it-Yourself Mock Community Standard for Multi-Step Assessment of Microbiome Protocols. *Current Protocols*, *2*(9). https://doi.org/10.1002/CPZ1.533

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., & Weyrich, L. S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, *27*(2), 105–117. https://doi.org/10.1016/J.TIM.2018.11.003

Elie, C., Perret, M., Hage, H., Sentausa, E., Hesketh, A., Louis, K., Fritah-Lafont, A., Leissner, P., Vachon, C., Rostaing, H., Reynier, F., Gervasi, G., & Saliou, A. (2023). Comparison of DNA extraction methods for 16S rRNA gene sequencing in the analysis of the human gut microbiome. *Scientific Reports 2023 13:1*, *13*(1), 1–12. https://doi.org/10.1038/s41598-023-33959-6

Illumina. (n.d.). *An introduction to Next-Generation Sequencing Technology*. Retrieved February 27, 2024, from www.illumina.com/technology/next-generation-sequencing.html

QIIME 2. (n.d.). Retrieved February 26, 2024, from https://qiime2.org/

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/NAR/GKS1219

R: The R Project for Statistical Computing. (n.d.). Retrieved April 17, 2024, from https://www.r-project.org/

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, *469*(4), 967. https://doi.org/10.1016/J.BBRC.2015.12.083

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., & Bokulich, N. A. (2020). RESCRIPt: Reproducible sequence taxonomy reference database management for the masses. *BioRxiv*, 2020.10.05.326504. https://doi.org/10.1101/2020.10.05.326504

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*(1). https://doi.org/10.1186/S12915-014-0087-Z

Shetty, S. A., Kuipers, B., Atashgahi, S., Aalvink, S., Smidt, H., & de Vos, W. M. (2022). Inter-species Metabolic Interactions in an In-vitro Minimal Human Gut Microbiome of Core Bacteria. *Npj Biofilms and Microbiomes 2022 8:1*, *8*(1), 1–13. https://doi.org/10.1038/s41522-022-00275-2

Tan, B. F., Ng, C., Nshimyimana, J. P., Loh, L. L., Gin, K. Y. H., & Thompson, J. R. (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Frontiers in Microbiology*, *6*(SEP), 1027. https://doi.org/10.3389/FMICB.2015.01027

Tourlousse, D. M., Narita, K., Miura, T., Ohashi, A., Matsuda, M., Ohyama, Y., Shimamura, M., Furukawa, M., Kasahara, K., Kameyama, K., Saito, S., Goto, M., Shimizu, R., Mishima, R., Nakayama, J., Hosomi, K., Kunisawa, J., Terauchi, J., Sekiguchi, Y., & Kawasaki, H. (2022). Characterization and Demonstration of Mock Communities as Control Reagents for Accurate Human Microbiome Community Measurements. *Microbiology Spectrum*, *10*(2). https://doi.org/10.1128/SPECTRUM.01915-21

What is index hopping? (n.d.). Retrieved February 26, 2024, from https://thesequencingcenter.com/knowledge-base/what-is-index-hopping/

7. Appendix

Figuur 11 value

Bijschrift:

Relationship between OTU abundance, prevalence, and cross-contamination – A scatterplot displaying the percentage of sample OTU in mock communities plotted against the values (maximum, mean, median abundance and prevalence). Each data point represents an OTU, with different studies distinguished by colours and sample characteristics (maximum, mean, median abundance, and prevalence) indicated by shapes.

Figure 12 DNA ratio

Relationship between trend and DNA concentration ratio – A scatterplot displaying DNA concentration ratio plotted against the trend. Different shapes represent sample characteristics (maximum, mean, median abundance, and prevalence) within each column. The colours represent different DNA characteristics (maximum, mean and median DNA concentration) for each row.

Figure 13 DNA diff

Relationship between trend and DNA concentration difference – A scatterplot displaying DNA concentration difference plotted against the trend. Different shapes represent sample characteristics (maximum, mean, median abundance, and prevalence) within each column. The colours represent different DNA characteristics (maximum, mean and median DNA concentration) for each row.