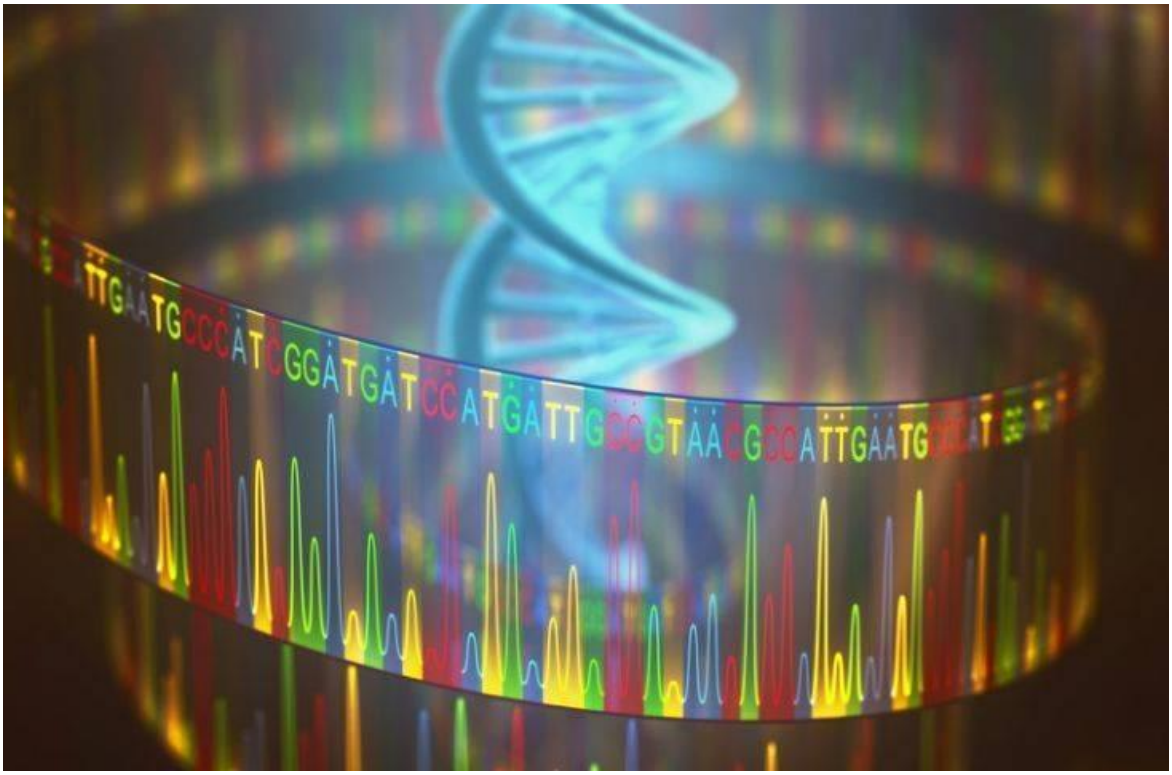
Towards tailored mock communities for quality-controlled microbiome profiling



Stijn Teunissen (000024753)

Supervisor Van Hall Larenstein: Wouter Suring & Rik Veldhuis

Supervisor Wetsus: Dr. Ir. Pieter van Veelen

Date & place: 6/5/2024, Leeuwarden

NHL STENDEN university of applied sciences

van hall larenstein university of applied sciences

LS&T LEEUWARDEN

wetsus european centre of excellence for sustainable water technology

Towards tailored mock communities for quality-controlled microbiome profiling

Student name: Stijn Teunissen (000024753)

Supervisor Van Hall Larenstein: Wouter Suring & Rik Veldhuis

Supervisor Wetsus: Dr. Ir. Pieter van Veelen

Education and school: Biotechnology on Van Hall Larenstein, Leeuwarden

Company: Wetsus, Leeuwarden

Date & place: 6/5/2024, Leeuwarden

Image Source: Public Domain

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 an elective course in biological data science. I was able to use this knowledge 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, 5 June 2024

Samenvatting

High-throughput sequencing van 16S rRNA gen-amplicons is een breed toepasbare methode geworden voor het in kaart brengen van complexe microbiële gemeenschappen. Echter tijdens sequensen van microbiële gemeenschappen kan contaminatie optreden. Contaminatie betreft ongewenste DNA-fragmenten die niet afkomstig zijn van samples, wat kan leiden tot onjuiste conclusies. Contaminatie kan optreden tijdens het DNA-isolatieproces, maar kan ook het gevolg zijn van cross-contaminatie tussen samples. Dit is voornamelijk een probleem bij samples met een lage biomassa. Dit project heeft als doel de mate van cross-contaminatie te kwantificeren, waarvan wordt aangenomen dat deze een functie is van DNA-concentratieverschillen tussen samples. Deze studie werd uitgevoerd door sequencing data en DNA-concentraties van samples uit gepubliceerde studies en studies die binnen Wetsus zijn uitgevoerd, te verzamelen en te analyseren met behulp van bioinformatica tools zoals QIIME2 en R. Door gebruik te maken van mock communities als positieve sequencing controles, werd de mate van contaminatie onderzocht. Eerst werd de terugwinning van mock taxa beoordeeld, waaruit bleek dat niet alle taxa waren gedetecteerd. Vervolgens werd de contaminatie van bacteriën afkomstig van samples in de mock community geëvalueerd, die varieerde van 0,75% tot 21% met een mediane van 6,7% tussen de studies. Voor contaminant ASVs in de mock communities werd geanalyseerd of het percentage contaminatie afhankelijk was van de ASV-abundantie in het sample of de ASV-prevalentie in de samples. De resultaten toonden aan dat de mediane ASV-abundantie over samples significant het percentage sample ASV in de mock community voorspelde, de hellingen van de mediane abundantie in monsters versus waargenomen abundantie in de mock community varieerden significant tussen de studies. Prevalentie over samples voorspelde ook significant het percentage sample ASV in de mock community, en dit varieerde ook significant per studie. De hellingen van de maximale, gemiddelde en mediane abundantie en prevalentie werden gerelateerd aan de DNA-concentratie ratio (sample/mock DNA-concentratie) en verschil (sample - mock DNA-concentratie) om te onderzoeken of DNA-concentratie invloed heeft op contaminatie. Het verschil in DNA-concentratie tussen de mock community en de sample, maar niet de ratio, had een significant effect op de trend (contaminatiehelling). Dit effect varieerde significant per monsterkenmerk (maximale, gemiddelde, mediane abundantie en prevalentie). Er kan geconcludeerd worden dat DNA-concentratieverschillen tussen sampels de mate van cross-contaminatie in amplicon-sequencing beïnvloeden. Voor toekomstige sequencing studies wordt geadviseerd om de DNA-concentraties tussen samples en mock communities consistent te houden om cross-contamintie te minimaliseren. Een beperking van deze studie is het beperkte aantal mock samples in de totale dataset. Voor toekomstig onderzoek zouden meer studies met mock communities aan de dataset toegevoegd kunnen worden om het bewijs nog sterker te maken. Verder onderzoek zou ook kunnen uitwijzen wat er gebeurd is met de bacteriën die wel in de mock community hadden moeten zitten, maar niet gedetecteerd zijn. Het kan onderzocht worden of de bacteriën daadwerkelijk aanwezig waren in de mock community of dat ze niet geclassificeerd zijn door het gebruik van verschillende primers of andere redenen.

Abstract

High-throughput sequencing of 16S rRNA gene amplicons has become a widely applicable method for profiling complex microbial communities. However, contamination can bias diversity estimation, which often occurs during the sequencing process. Contamination refers to unwanted DNA fragments that do not originate from the sample, potentially leading to incorrect conclusions. Contamination can occur during the DNA isolation and due to cross-contamination between samples. This is mainly a problem in samples with low biomass. This project aims to quantify the degree of cross-contamination, which is hypothesised to be a function of DNA concentration differences between samples. This study was conducted by collecting and analysing sequencing data and DNA concentrations from samples in published studies and studies conducted within Wetsus using bioinformatics tools such as QIIME2 and R. By leveraging mock communities as positive sequencing controls, the degree of contamination was investigated. First, the recovery of mock taxa was assessed, which indicated that not all that had been detected. Then, contamination of sample-derived bacteria was evaluated in the mock community, which ranged from 0,75% to 20,9% and had an overall median of 6,7% among studies. For contaminant ASVs in the mock communities, it was analysed if the percentage of contamination depended on the ASV abundance in the sample or the ASV prevalence in the samples. The results showed that median ASV abundance across samples significantly predicted the percentage of sample ASV in the mock community. However, the slopes of median abundance in samples vs. observed abundance in the mock community varied significantly among studies. Prevalence across samples also significantly predicted the percentage of sample ASV in the mock community, and this also varies significantly per study. The maximum, mean, median abundance and prevalence slopes were related to the DNA concentration ratio (sample/mock DNA concentration) and difference (sample – mock DNA concentration) to investigate whether DNA concentration influences contamination. The difference in DNA concentration between the mock community and the sample, but not the ratio, significantly affected the trend (contamination slope). This effect varies significantly per sample characteristic (maximum, mean, median abundance, and prevalence). It can be concluded that DNA concertation differences among samples influence the degree of cross-contamination in amplicon sequencing. For future sequencing studies, keeping DNA concentrations among samples and mock communities consistent is advised to minimise cross-contamination. A limitation of this study is the limited number of mock samples in the total dataset. For future research, more studies with mock communities could be added to the dataset to strengthen the evidence. Future research could also investigate what happened to the bacteria that should have been in the mock community but were not detected. It could be investigated whether the bacteria were present in the mock community or were not classified due to the use of different primers or other reasons.

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 helps develop innovations and provides 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) (Stoler & Nekrutenko, 2021).

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. 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 signal microbial diversity. 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) (Stoler & Nekrutenko, 2021) (Caporaso et al., 2011).

1.2. 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) (Stoler & Nekrutenko, 2021).

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' (Stoler & Nekrutenko, 2021).

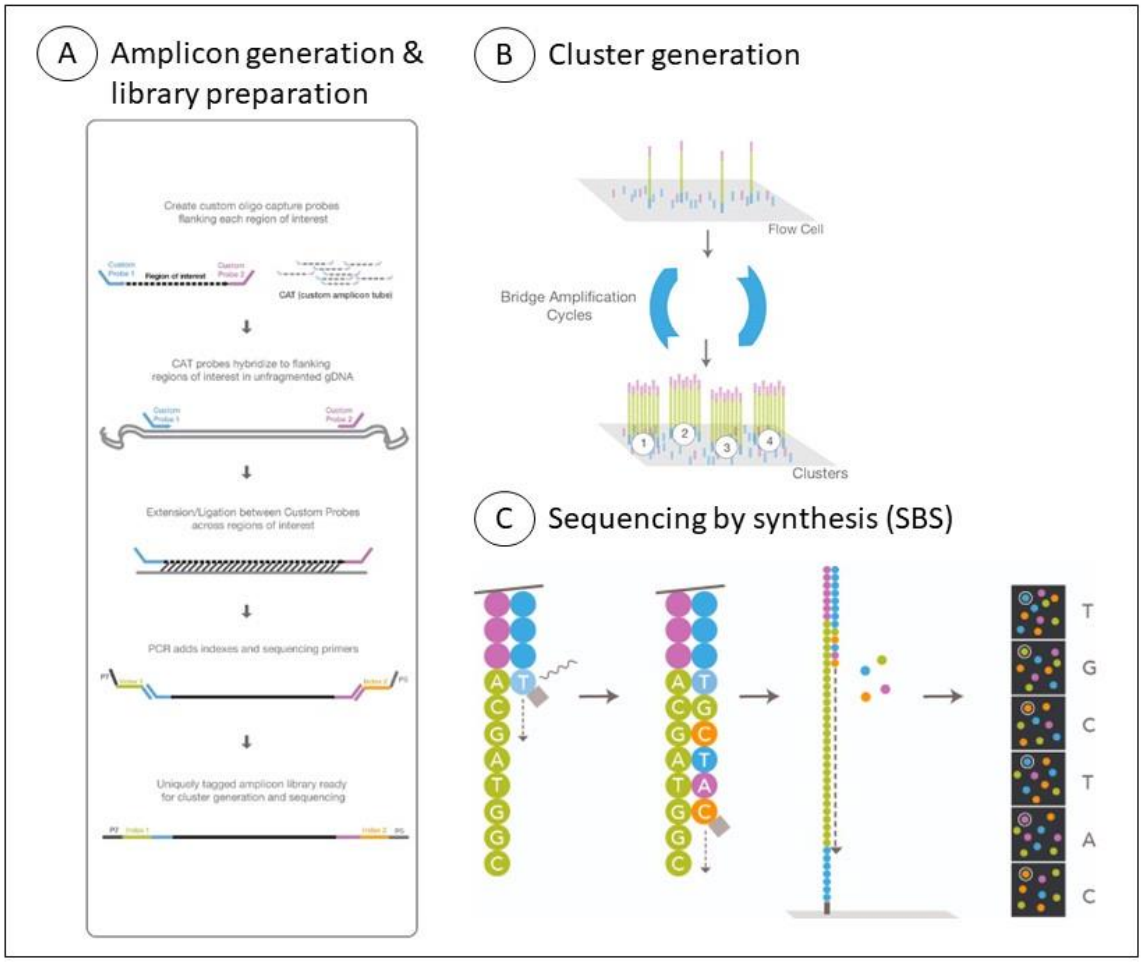


Figure 1 Amplicon sequencing workflow overview - Illumina sequencing includes three steps. (A) Amplicon generation & library preparation, (B) Cluser generation, and (C) Sequencing by synthesis (SBS) (n.d., 2017.).

Illumina sequencing uses the sequencing by synthesis (SBS) process; chemically modified nucleotides 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.) (Stoler & Nekrutenko, 2021).

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.

1.2.1. Multiplexing

Multiple samples can be sequenced simultaneously in a single run, 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.

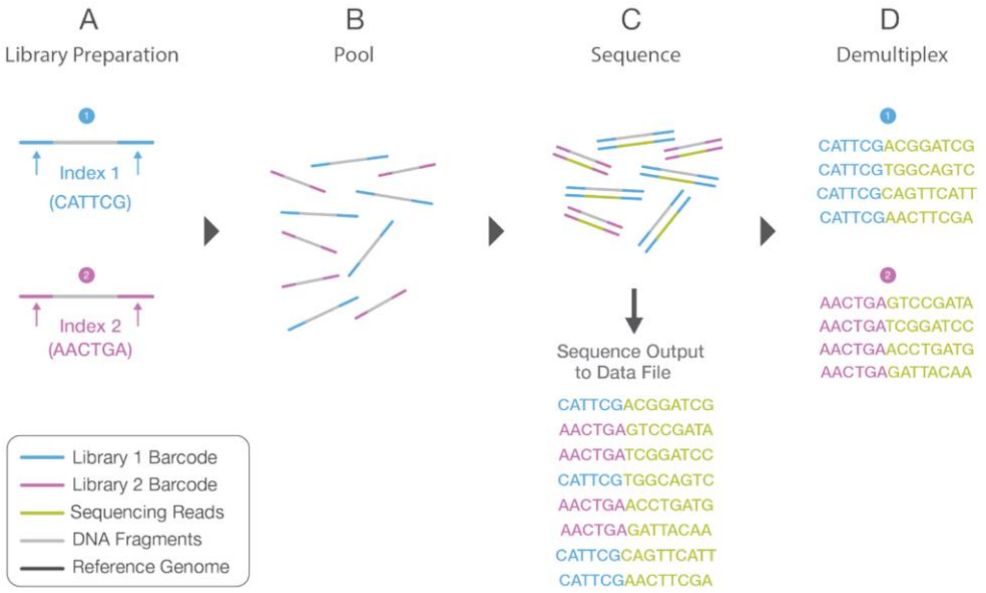


Figure 2 Multiplexing - Multiplexing of multiple libraries. (A) Unique index sequences are assinged to each DNA fragment during library preparation. (B) Multiple libraries are pooled together. (C) Libraries are sequenced. (D) The sequence data is demultiplexed. (n.d., 2017.)

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). 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 (Holm et al., 2019).

1.3. Contamination during Sequencing

If contamination occurred during sequencing, unwanted DNA fragments were found in the sample. This could lead to inaccuracies in the analysis or false positive results, resulting in incorrect conclusions. During microbiome profiling, two main types of contamination could arise. Contaminant DNA could originate from many sources, and cross-contamination could occur 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 a low microbial biomass, such as water samples. Contamination during DNA extraction is another common source of DNA contamination. Cross-contamination between samples can occur during sample processing and 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'. (Holm et al., 2019) (Eisenhofer et al., 2019) (Weissensteiner et al., 2021) (Salter et al., 2014).

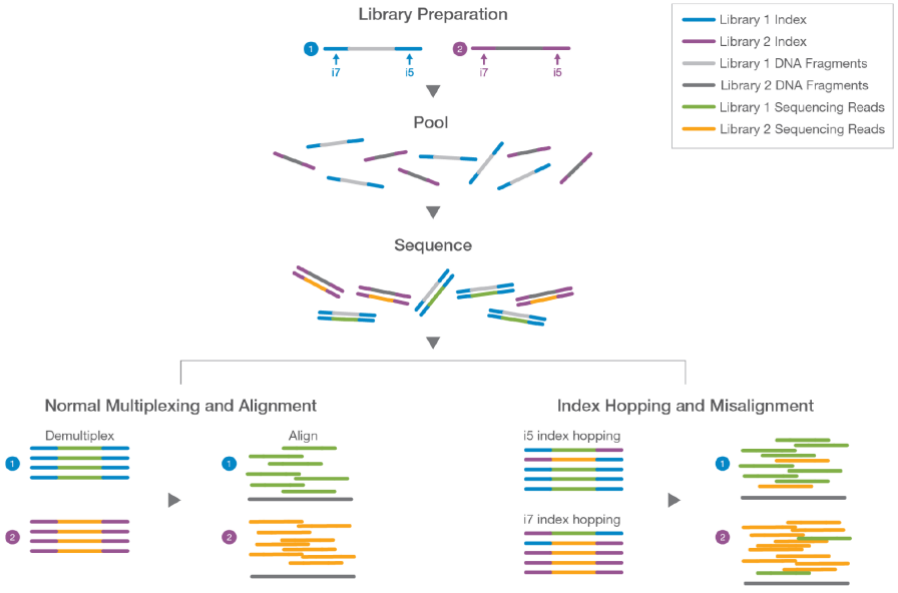


Figure 3 Index hopping - Shows the library preparaion process, sample pooling, sequencing, and indexing for two individual samples. The difference between index hopping and normal multiplexing is shown (n.d., 2017.)

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).

1.4. Mock Community

Microbiome samples typically have unknown compositions, and thus, it is inherently impossible to determine the accuracy of a sample microbial profile. Mock communities are artificial microbial communities constructed from bacterial strains or whole cells mixed in known quantities and proportions to serve as positive controls to microbiome sequencing experiments, providing insights into the precision of estimated taxa proportions. Mock communities also allow for estimating the proportion of false positive detections likely resulting 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 their environment or other settings. Second, these mock communities can be expensive. Alternatively, it is also possible to create a 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. In contrast, Log10 mock communities distribute bacteria on a logarithmic scale based on tenfold differences in relative abundance. This mock community is particularly useful for testing and calibrating a broad range of bacterial concentrations, from very high to very low, simulating more typically the complex and variable communities in natural environments. Similarly, Log2 mock communities distribute bacteria on a logarithmic scale by a power 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. Project aims

This project aims first to quantify the degree of cross-contamination from samples to mock communities. Then, assuming contamination is a random process, the project investigates if contamination is a function of DNA concentration difference between samples and mock communities by using 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 at least a single mock community, 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 are collected.

In this study, the final data set consisted of seventeen mock samples from four different mock communities (in-house Wetsus 2022, in-house Wetsus 2023, Zymo Log10, and Zymo equal) 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 from samples to mock, examine whether contamination in mock communities is influenced by the abundance and prevalence of bacteria in samples, and estimate the cross-contamination effects of DNA concentration differences between samples and mock communities.

2. Methods

In this study, the package QIIME2 (version 2022.11) (Bolyen et al., 2019) was used, and R software (version 4.3.2) (R Core Team, 2023) to analyse the raw and processed published sequencing data. The data analysis workflow is shown in Figure 4. QIIME 2 is a powerful bioinformatics package explicitly designed for microbiome analysis, offering a wide range of features for processing, analysing, and visualising NGS data.

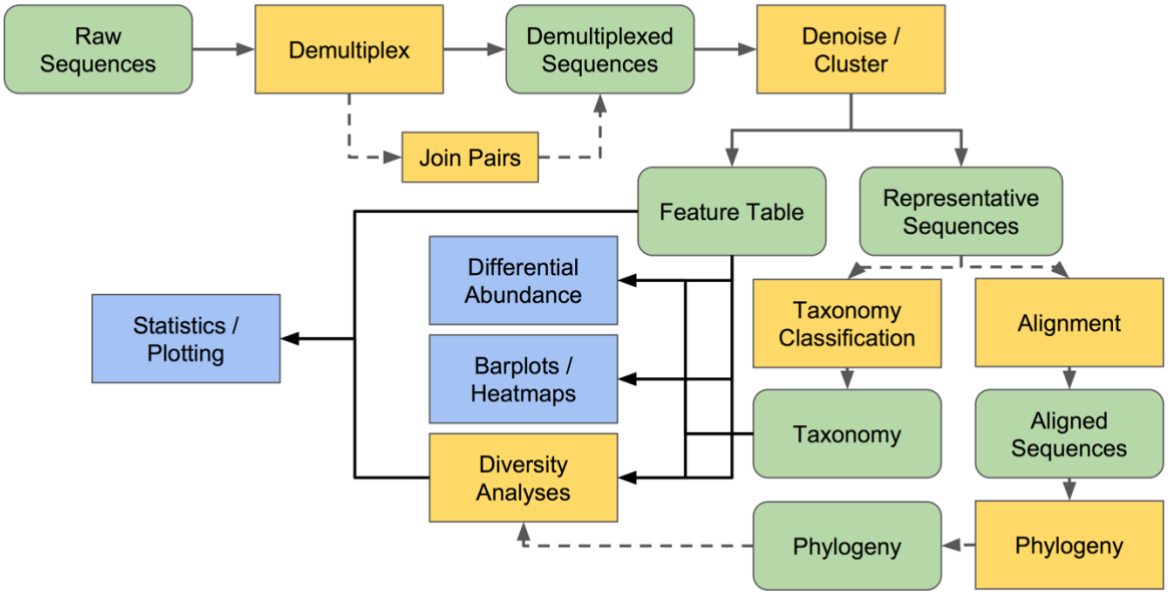


Figure 4 Data analyse workflow - The workflow to analse raw sequence data using QIIME2 and R (Bolyen et al., 2019).

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., 2019).

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

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. This is done using the package DADA2 (Callahan et al., 2016) in QIIME2. During denoising, reads are denoised into amplicon sequence variants (ASVs) with two goals: reducing sequencing errors and dereplicating sequences. This is done by filtering out noisy sequences, correcting errors in sequences, removing

chimeric sequences, removing singletons, joining paired-end reads, and then dereplicating those sequences.

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 (99.9% confidence) is used as the lower threshold.

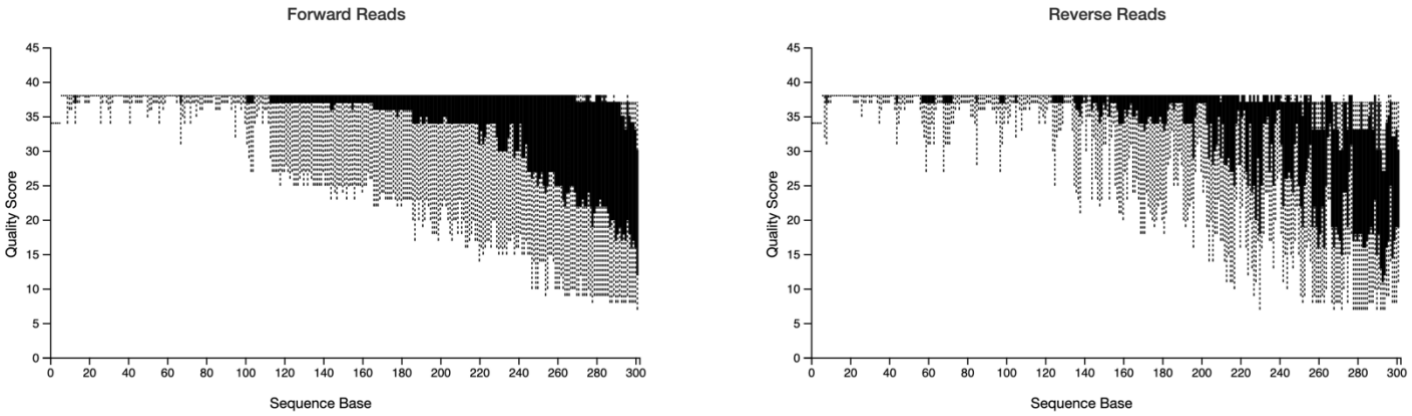


Figure 5 Phred scores – The Phred scores for each base of the forward and reverse reads.

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 bases at the end for forward reads, and 190 bases for reverse reads has been decided. The reverse reads always have lower quality (Figure 5), primarily because the reagents and additives are depleted towards the end of sequencing. Therefore, the reverse read is always trimmed slightly earlier.

During the error correction in sequences, sequences with likely errors are corrected to probable correct sequences based on base frequency probability modelling. Chimeric sequences are artificial sequences that arise during the PCR process. This happens when fragments from different DNA molecules are glued together, creating a new, non-natural sequence. These chimeric sequences can disrupt the analysis, can be detected and are therefore removed. Singletons are sequences that appear only once in the dataset. These may be the result of sequencing errors and are, therefore, also removed. After that, the forward and reverse reads are joined together through an overlapping region. Finally, sequences are dereplicated. This is the process of combining identical sequences in the dataset into a single representative sequence. Each unique sequence is represented only once, along with a frequency indicating how often this sequence occurs in the dataset. The final product of denoising is a FeatureTable[Frequency] (feature table) and FeatureData[Sequence] (representative sequences) artefact. These two of the most important artefacts in an amplicon sequencing workflow are used for many downstream analyses (Kishore et al., 2023) (Callahan et al., 2016) (Bolyen et al., 2019).

2.2.2. Phylogenetic diversity analyses

Phylogenetic diversity is a measure of biodiversity that incorporates phylogenetic differences between species and is stored in a Phylogeny[Rooted] artefact. This artefact is created using the FeatureData[Sequence] artefacts and an alignment tool such as MAFFT (Katoh & Standley, 2013), followed by a tree construction tool such as FastTree (Price et al., 2010). The resulting artefact is used to perform phylogenetic diversity measurements, which provide insight into microbial communities' evolutionary relationships and diversity within and between samples (Bolyen et al., 2019).

2.2.3. Taxonomy classification

It is crucial to identify which organisms are present in the samples. This can be done by comparing the feature sequence artefacts against a database containing sequences with known taxonomy. This comparison can be using a trained machine learning model called a classifier. This classifier is trained to recognise which sequence 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 version 138.1 database (Quast et al., 2013) (Robeson et al., 2020) (Bokulich et al., 2018). For methodological consistency, it was decided to use a classifier trained on the complete 16S gene sequence rather than classifiers trained 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 (Bokulich et al., 2018) (Bolyen et al., 2019).

2.3. R analyses

Once the reads had been analysed with QIIME2, the following files were needed to analyse them in R further: 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 (McMurdie & Holmes, 2013). From the Phyloseq object, unwanted taxa such as Mitochondria, Chloroplasts, Unclassified Kingdom, Eukaryota, etc., were first filtered out. All incompletely classified ASVs were also supplemented. This meant that if an ASV 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 mock community abundance

After loading the necessary files, the R analysis was performed. It was checked whether the measured abundance of the mock communities used in this study matched their theoretical abundance in the created mock mixtures. The mock communities used in this study were in-house Wetsus 2022, in-house Wetsus 2023, Zymo Equal, and Zymo Log10 mock communities. For each mock community, a tibble was created with the known bacterial genera 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. All genera not expected in the mock were grouped and renamed "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 every study with different mock communities, which were then grouped per mock type. The abundance per genus and the "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

It was then determined if ASVs 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 ASV per study was calculated. Prevalence indicates the fration of samples in a study with a bacterium. 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, bacterial prevalence was examined to assess if prevalence affected contamination in the mock community. The maximum, mean, and median abundance of bacteria were calculated per

study. This was done to determine whether the abundance variable of the samples influenced the contamination in the mock communities. The maximum and median abundances were chosen to assess the effect of outliers from distorting the results from average abundance.

The mock samples were subsets for each study, and the contaminant bacteria in these mock samples were selected for further analysis. The prevalence, maximum, mean, and median abundances of contaminant bacteria were added to the mock samples subset for comparison. The percentage of sample ASVs in the mock community was plotted against the prevalence, maximum, mean, and median abundance and statistically tested using linear regression with the lm function.

2.3.4. Relationship between contamination in mock communities and DNA concentration

To determine if the slope of the percentage of sample ASVs in the mock community 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 mock communities were calculated. For each sample ASV characteristics (prevalence, maximum, mean, and median abundance), the trend was estimated by examining the relationship between the percentage of sample ASVs and the percentage in the mock community for every study using an interaction term. To calculate the slope of each combination of study and sample characteristics, the emtrends function of the emmeans package (Lenth, 2023) was used. 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 and interpreted: one where the regression slope was plotted against the DNA concentration ratio and another where the trend was plotted against the DNA concentration difference.

2.4. Data Availability

The code used for the R analysis in this study has been uploaded to GitHub (https://github.com/stijnteunissen/microbialanalysis.git). The files required to run the script and the script used for data analysis in QIIME2 are also available on GitHub.

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. It was determined what percentage of the mock community consists of contamination. The amount of contamination in mock communities is also examined, as the abundance and prevalence of sample genera influence it. Finally, it was examined whether there is a correlation between the slope of contamination and the DNA concentration difference between samples and mock communities.

3.1. The measured abundance matches the theoretical abundance in the mock community.

Various mock communities were utilised: in-house Wetsus 2022, in-house Wetsus 2023, Zymo Equal, and Zymo Log10 mock communities. Table 1 shows which bacteria are theoretically expected to be present in each type of mock community. 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 mock community are plotted against each other. This provides insight into the reliability of the mock communities.

Table 1 Theoretical bacteria in the mock community at genus level.

Mock community	Theoretical bacteria
In-house Wetsus 2022	Serratia
	Massilia
	Brevibacillus
	Bacillus
	Weizmannia
	Lysinibacillus
	Peribacillus
	Bordetella
In-house Wetsus 2023	Brevibacillus
	Weizmannia
	Lysinibacillus
	Bacillus
	Sphingobium
	Serratia
	Streptomyces
	Burkholderia
Zymo (Equal & Log10)	Listeria
	Pseudomonas
	Bacillus
	Saccharomyces
	Escherichia
	Salmonella
	Lactobacillus
	Enterococcus
	Cryptococcus
	Staphylococcus

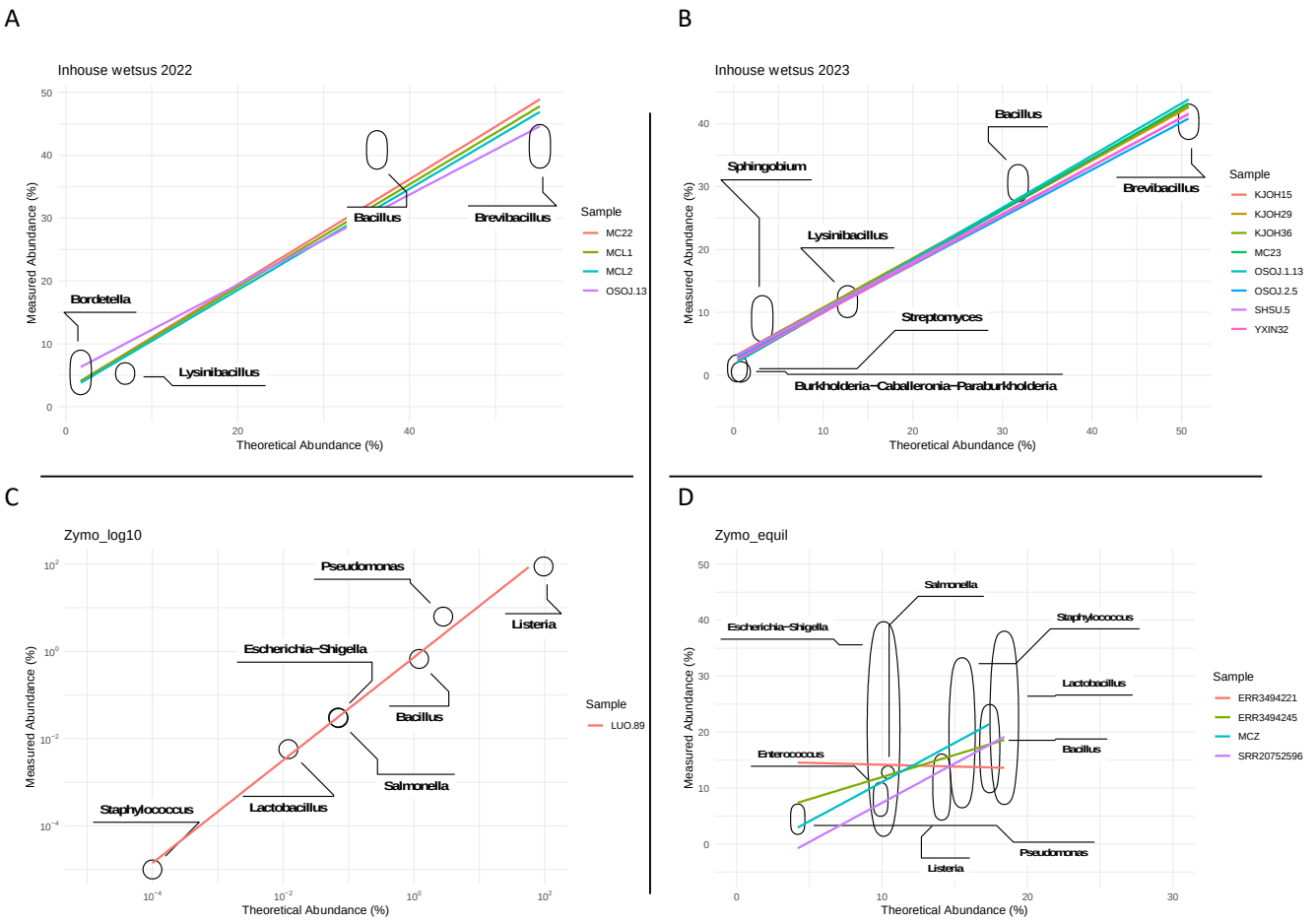


Figure 6 Theoretical vs measured abundance – The theoretical versus measured abundances of 17 mock samples in four different mock communities. A) the in-house Wetsus 2022 mock community, B) the in-house Wetsus 2023 mock community, C) the Zymo Log10 mock community, and D) the Zymo Equil mock community.

The different mock communities were used multiple times across various studies. Figures 6A-D illustrate the differences between mock samples within a mock community. In Figure 6A, in-house Wetsus 2022 mock community was compared with Table 1. The following bacteria, which should have been present theoretically, were not found in the mock community: *Serratia*, *Massilia*, *Weizmannia* and *Peribacillus*. The bacteria that were found were closely clustered per study but could deviate from the theoretical abundance. The theoretical abundance of *Brevibacillus* was 55,2%, but the measured abundance ranged from 39% to 43%, which was lower than expected. The theoretical abundance of Bordetella was 1,7%, but it ranged from 3,4% to 7,5%, which was higher than expected.

Figure 6B, the in-house Wetsus 2023 mock community, was compared with Table 1. Bacteria *Serratia* and *Weizmannia* that should have been present were not found in the mock community. Additionally, the points between the studies were closely clustered but could deviate from the theoretical abundance. *Brevibacillus* had a theoretical abundance of 50.8%, but it was found between 39% and 42%, which was lower than expected.

The Zymo Log10 mock community (Figure 6C) was used only once but almost matched the theoretical abundance. Only *Enterococcus* and *Staphylococcus* were expected, according to Table 1, but were not found. The fungus and yeast *Cryptococcus* and *Saccharomyces* were also not found, and this was as expected.

The Zymo Equil mock community (Figure 6D) showed the most variation between the mock samples. This mock community was used in a Wetsus study (MCZ) and included data from external studies

(SRR20752596, ERR3494245, and ERR3494221). When Figure 6D was compared with Table 1, *Salmonella* was only found in the MCZ mock sample and not in the other studies. Conversely, *Lactobacillus* was found in SRR20752596, ERR3494245, and ERR3494221 but not in the MCZ mock sample. This is not what was expected in theoretical *salmonella,* and *Lactobacillus* should be present in all mock samples. The bacteria found in all studies showed significant variation. For example, *Escherichia* was found between 2,5% and 38,6%, while theoretically, it should have been around 10,1%. For *Staphylococcus*, the points ranged between 7,6% and 32,2%, with a theoretical abundance of 15,5%. *Lactobacillus* found in SRR20752596, ERR3494245, and ERR3494221 varied between 8,2% and 36,9% but had a theoretical abundance of 18,4%.

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, these are considered contamination. Figure 7 illustrates the percentage of the mock communities that consist of 'other genera' (contamination) in grey.

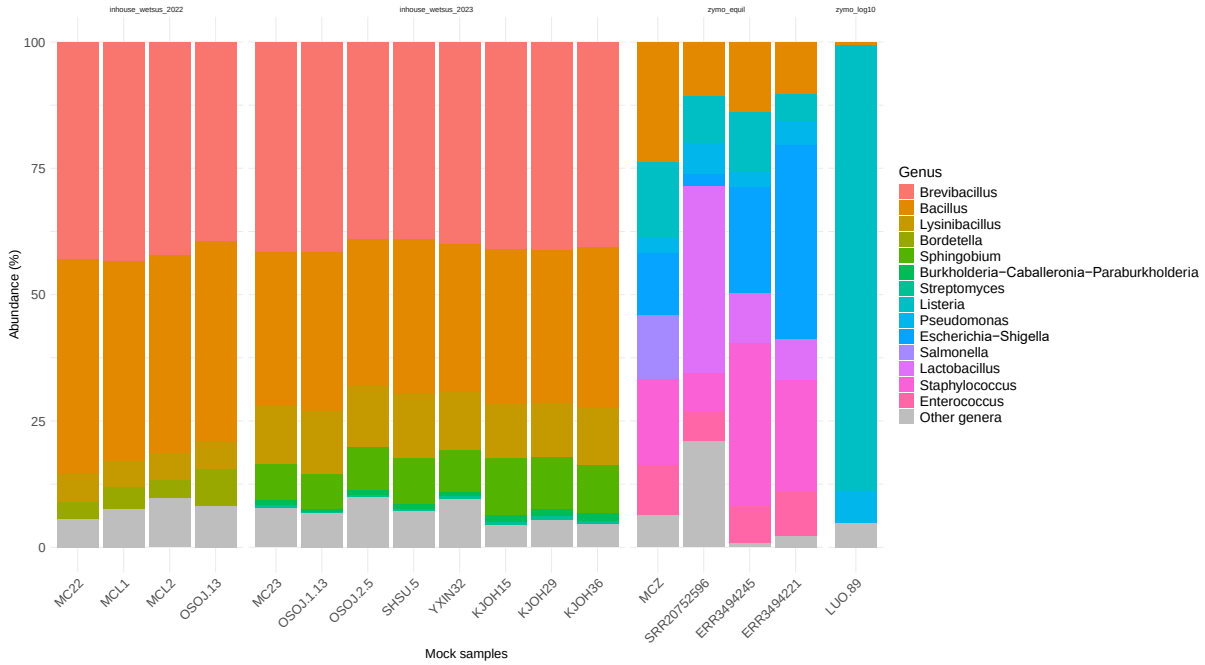


Figure 7 Summary of 16S rRNA gene sequencing taxonomic classification – bar plot of the different types of mock communities (in-house Wetsus 2022, in-house Wetsus 2023, Zymo Equal, and Zymo Log10). Each column represents a mock sample.

Figure 7 showed that certain bacteria, which were theoretically expected to be present in the mock community, were missing. For in-house Wetsus 2022, the bacteria *Serratia*, *Massilia*, *Weizmannia*, and *Peribacillus* were missing. For in-house Wetsus 2023, *Serratia* and *Weizmannia* were missing. For Zymo Log10, *Enterococcus* and *Staphylococcus* were missing. For Zymo Equal, *Lactobacillus* was not found in the MCZ mock sample, and *Salmonella* was not found in the SRR20752596, ERR3494245, and ERR3494221 mock samples. This confirmed the findings from Figure 6.

In Figure 6, it was discussed that the points of the in-house Wetsus 2022 and 2023 mock communities were close to each other and deviated only slightly from the theoretical abundance. In the bar plot of Figure 7, the in-house Wetsus 2022 (MC22, MCL1, MCL2, and OSOJ.13) and in-house Wetsus 2023 (MC23, OSOJ.1.13, OSOJ.2.5, SHSU.5, YXIN32, KJOH15, KJOH29, and KJOH36) mock samples were consistent across different studies. This was expected, as little difference was anticipated between the mock communities used across various studies. It was also shown that a portion of the mock community consisted of contamination, indicated in grey, ranging from 5,6% to 10,9% of the mock samples.

The Zymo Log10 (LUO.89) mock community was used only once, so it could not be determined if it remained consistent across different studies, but there was 4.7% contamination visible in grey. For the Zymo Equal (MCZ, SRR20752596, ERR3494245, and ERR3494221) mock samples, Figure 6 already illustrated a large variation between the different mock samples, which was also reflected in the bar plot (Figure 7). Additionally, the bar plot showed that the contamination ranged from 0,75% to 20,9%. The overall median of all the mock samples was 6,7% contamination among studies. This confirmed that contamination occurred during sequencing.

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 ASV in the mock communities and the genus abundance in the sample or the genus prevalence in the sample. This is illustrated in Figures 8 and 9.

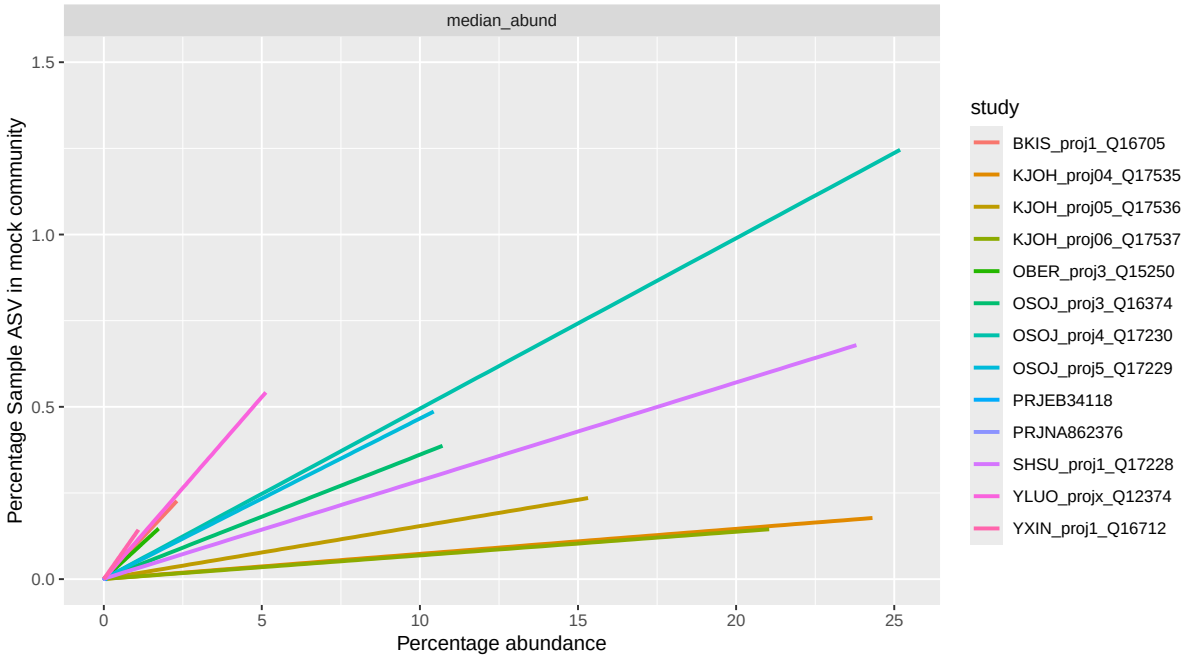


Figure 8 Relationship between ASV abundance and cross-contamination – A scatterplot displaying the percentage of sample ASV in mock communities plotted against the value (median abundance). Each data point represents an ASV, with different studies distinguished by colours.

The median abundance is illustrated in Figure 8. The maximum and mean abundance are illustrated in Figure 11 in the appendix. Figures 8 and 11 showed a relationship between the abundance of ASV in the sample and the contamination in the mock community. This relationship was observed because the lines sloped upward. The steepness of the lines varied per study, and the length of the lines also varied per study. The length of the lines was due to the variation in ASV abundance in the studies. The relationship between the abundance of ASV in the sample and the contamination in the mock community was expected since a high abundance of ASV present in the sample was more likely to cause contamination than a low abundance of ASV. This was confirmed with a statistical test, indicating that median abundance significantly affected the percentage of sample ASV in the mock community, which varies significantly per study (ANOVA: $F_{12, 56776} = 14{,}98$, $p < 10^{-15}$).

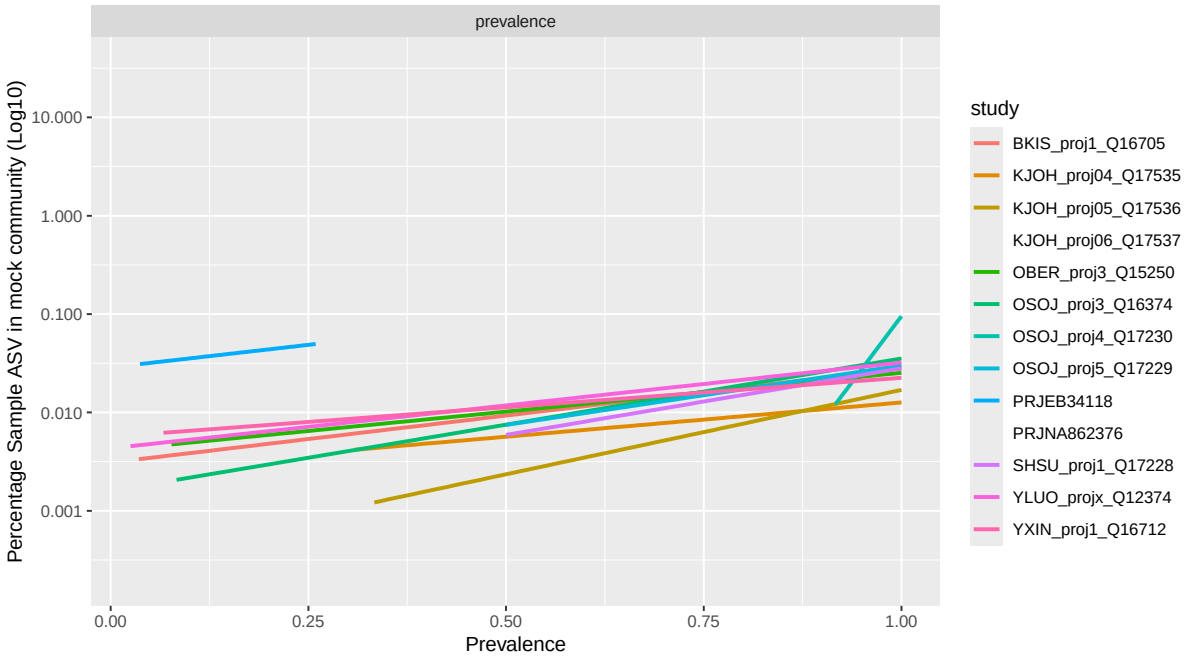


Figure 9 illustrates the prevalence. For prevalence, there was also a relationship between the amount of contamination and prevalence. This relationship is less strong than that of abundance. This was evident from the less steep lines in the plot with wider confidence intervals. This was also expected because a low-abundance ASV with a high prevalence does not necessarily cause more contamination than a high-abundance ASV with a lower prevalence. This was confirmed with a statistical test, indicating prevalence significantly affected the percentage of sample ASV in the mock community, which varied significantly per study. The prevalence effect was significant but less strong than the median abundance effect (ANOVA: $F_{12,\ 54939}$ = 1,91, p = 0,029). The slopes of the lines for both abundance and prevalence differed significantly per study. Therefore, these slopes were plotted against the difference and ratio in DNA concentrations between samples and mock communities. The expectation is that DNA concentration is related to the slope.

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

The data presented in Figures 8 and 9 indicated a relationship between median abundance and the percentage of cross-contamination, as well as a relationship between prevalence and the percentage of cross-contamination. The maximum, mean, median abundance and prevalence slopes are expected to be related to the DNA concentration ratio (sample/mock DNA concentration) and the DNA concentration difference (sample – mock DNA concentration) to investigate whether DNA concentration influences contamination. This was plotted in Figures 10/11 and 13/14 in the appendix.

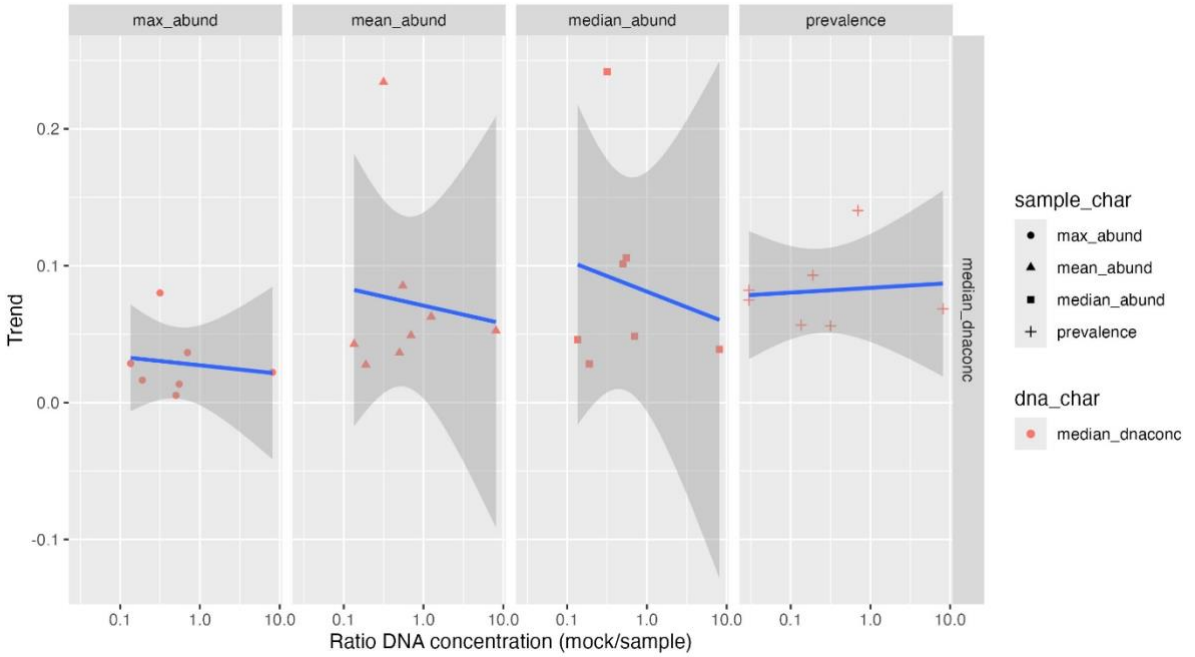


Figure 9 Relationship between trend and DNA concentration ratio – A scatterplot displaying DNA concentration ratio plottea against the trend. Different shapes represent sample characteristics (maximum, mean, median abundance, and prevalence) within each column. The colours represent each row's different DNA characteristics (maximum, mean and median DNA concentration).

The median DNA concentration is illustrated in Figure 10. The maximum and mean DNA concentrations are illustrated in Figure 13 in the appendix. Since the difference between the maximum, mean, and median DNA concentrations was small, it was decided to present the median DNA concentration because it is the least sensitive to outliers. In Figure 10, it was illustrated that the DNA concentration ratio was hardly related to the trend (slope) of the maximum, mean, median abundance, and prevalence. This was evident as the lines were not steep, and the confidence intervals were wide and crossed zero. This and the statistical test indicated that the DNA concentration ratio was unrelated to the trend. This was not as expected, as the DNA concentration ratio was thought to have more influence on the trend.

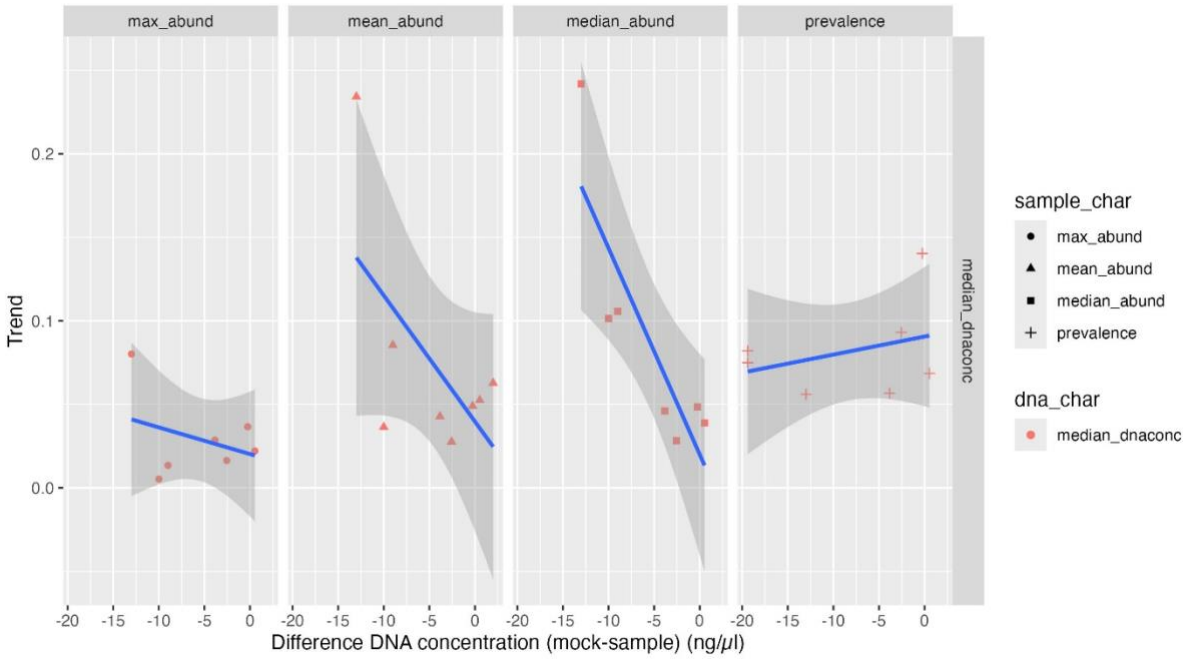


Figure 10 Relationship between trend and DNA concentration difference – A scatterplot displaying DNA concentration difference plotted against the trend. Different shapes represent sample characteristics (maximum, mean, median abundance, and prevalence) within each column. The colours represent different DNA characteristics (median DNA concentration) for each row.

The median DNA concentration is illustrated in Figure 11. The maximum and mean DNA concentrations are illustrated in Figure 14 in the appendix. In Figure 11, it was shown that the DNA concentration difference was related to the trend (slope) of the mean and median abundance. The lines for mean and median abundance sloped linearly downward. This meant that as the difference in DNA concentration decreased, the slope between abundance and contamination became flatter. The lines for maximum abundance and prevalence were flatter. The confidence intervals for mean and median abundance did not cross zero, while those for maximum abundance and prevalence did. The significance test indicated that the difference in DNA concentration between the mock and the sample was an essential factor influencing the trend, and this effect varied significantly by sample characteristic (ANOVA: $F_{3, 21} = 5{,}22$, $p = 0{,}008$). There was a significant difference between the slopes of median abundance and prevalence ($p = 0.009$). This was consistent with the expectation that the DNA concentration difference would affect contamination. Therefore, the DNA concentration difference did influence the slope, while the DNA concentration ratio did not.

4. Discussion

Cross-contamination is a significant issue in sequencing studies, especially for sample types with low microbial biomass. This study investigates the quantification of cross-contamination as a function of DNA concentration differences between samples and mock communities.

4.1. The measured abundance matches the theoretical abundance in the mock community

A comparison was made of whether the measured abundance of the mock communities aligns with the theoretical abundance, as shown in Figure 5. It was found that not all bacteria expected in the mock community were detected. The expected bacteria for the in-house Wetsus 2022 mock community were *Serratia, Massilia, Brevibacillus, Bacillus, Weizmannia, Lysinibacillus, Peribacillus* and *Bordetella*. However, *Serratia, Massilia, Weizmannia* and *Peribacillus* were not found in the mock community. There are several explanations for this. *Weizmannia coagulans* belonged to the *Bacillaceae* family and was formerly known as *Bacillus coagulans* according to Son et al., (2023). Therefore, *Weizmannia* was likely identified as *Bacillus*. *Peribacillus* also belong to the *Bacillaceae* family and previous members of the genus *Bacillus,* according to Manetsberger et al., (2023). Therefore, *Peribacillus* was likely identified as *Bacillus*. This was corrected by reviewing the percentage ratios. However, *Serratia* and *Massilia* were not detected at all. There could be two reasons for this. First, the primers and classification may not correctly identify the genus. Since primers within the 16S gene were used to identify nearly all bacteria but not specifically *Serratia* or *Massilia*, they might have been misclassified. Second, they might not have been added to the mock in the first place due to a mistake during its preparation, although this is difficult to verify.

For the in-house Wetsus 2023 mock community, the composition is slightly different, including *Brevibacillus, Weizmannia, Lysinibacillus, Bacillus, Sphingobium, Serratia, Streptomyces* and *Burkholderia*. Again, Weizmannia was not separately identified and fell under *Bacillus*. *Serratia* was not found, but 'genus of *Yersiniaceae*' was detected and might represent *Serratia* that was not identified at the genus level. *Serratia* belongs to the family *Yersiniaceae*. This was not the case in the in-house Wetsus 2022 mock community, where 'genus of *Yersiniaceae*' was not found.

The Zymo mock communities consist of *Listeria, Pseudomonas, Bacillus, Escherichia, Salmonella, Lactobacillus, Enterococcus, Staphylococcus, Cryptococcus* and *Saccharomyces*. In both Zymo Log10 and Zymo Equal, *Cryptococcus* and *Saccharomyces* were not found, as expected, since they are fungi and yeast, not bacteria. Only bacteria are expected due to the sequencing of the 16S gene.

In the Zymo Log10 mock community, two additional bacteria, *Enterococcus* and *Staphylococcus*, were not found. This is likely because the mock is divided in Log10, with these bacteria added in such low amounts (0,00067% and 0,0001%) that they fell below the detection threshold.

The Zymo Equal had the most variation among the mock samples. This is probably due to the different primers used in the published articles based on these mock samples. The primers used for the mock sample MCZ were 515F and 926R, for SRR20752596 341F and 806R, and for ERR3494245 and ERR3494221 515F and 806R. Different primers are better at identifying different groups of bacteria, leading to variation. For example, *Salmonella* was only found in mock sample MCZ, while *Lactobacillus* was found in all mock samples except MCZ. The study by Abellan-Schneyder et al., (2021), it is mentioned that taxonomic classification can differ significantly between different variable regions on the 16S gene. The use of different primers can thus explain the difference in taxonomy because these primers target different regions of the 16S gene.

The expected percentages were corrected for all bacteria not found in the mock communities, as shown in Figure 5. Most mock samples reasonably matched the expectations after correction. Some discrepancies could be due to contamination in the mock samples, altering the ratios. The contamination in the mock samples is shown in Figure 6. This contamination is expected from other samples due to cross-contamination and contamination during isolation and from reagents.

4.2. Contamination in Mock Communities

A bar plot was used to determine the presence of contamination in the mock samples (Figure 6). This illustrated that all samples were contaminated, ranging from 0,75% to 20,9%, with an overall median of 6,7% among studies. There is a significant variation among the mock samples, possibly due to isolation methods, reagents used, lab procedures, barcodes, or the sequencing apparatus. Contamination can come from cross-contamination between samples, isolation procedures, or reagents, as described by Salter et al., (2014). The two mock samples with the slightest contamination were ERR3494245 and ERR3494221. These samples came from a published article, making it difficult to determine if they genuinely had little contamination or had been pre-filtered, as this was not clearly described in the article.

Contamination in the mock communities suggests contamination is likely also present in the other samples. However, it is harder to identify because the samples exact composition is unknown. The contamination found in this study is consistent with other research where contamination was also found in samples after sequencing. In the article by Goig et al., (2020), they found around 10% contamination per sample during whole-genome sequencing. The article by Dyrhovden et al., (2021), mentioned that contamination occurs during 16S amplicon sequencing, varying between 2% and 32% depending on the biomass and dilution. It is also mentioned that one of the reasons for contamination could be cross-contamination, although this is difficult to determine.

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

Next, it was determined whether the contamination in mock samples is related to the percentage of sample ASV in the mock community and the genus abundance or prevalence in the samples, as shown in Figure 8. It was found that median abundance significantly affects the percentage of sample ASV in the mock community, varying significantly per study (ANOVA: $F_{12, 56776} = 14{,}98$, $p < 10^{-15}$). Prevalence also significantly affects the percentage of sample ASV in the mock community and varies significantly per study. The prevalence effect is significant but less strong than the median abundance (ANOVA: $F_{12, 54939} = 1{,}91$, $p = 0{,}029$). The article by Dyrhovden et al.Dyrhovden et al., (2021) supported the findings and mentioned that bacteria with dominant abundance cause contamination consistently across all controls.

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

Finally, the slope in Figure 8 was examined to determine if it is related to the DNA concentration ratio and difference. Figure 9 shows that the lines have no steep slopes, and the confidence intervals are wide and cross zero. This and the statistical test indicate no statistical relationship between the DNA concentration ratio and the trend. Figure 10 shows that the mean and median abundance lines slope linearly downward, while those 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, varying 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 ($p = 0{,}009$). This suggests that the difference in DNA concentrations is more important than the DNA concentration ratio. The results could be more reliable if more mock samples with varying DNA concentrations were added to the dataset.

4.5. Future research

A limitation of this study is the limited number of mock samples in the total dataset. For future research, more studies with mock communities could be added to the dataset to strengthen the evidence. Future research could also investigate what happened to the bacteria that should have been in the mock community but were not detected. It could be investigated whether the bacteria were present in the mock community or were not classified due to the use of different primers or other reasons.

5. Conclusion

This project aimed first to quantify the degree of cross-contamination from samples to mock communities. Then, assuming contamination is a random process, the project investigates if contamination is a function of DNA concentration difference between samples and mock communities by using published sequencing data. This study found that contamination exists in mock communities, with some of this contamination originating from cross-contamination between samples. The main influences on cross-contamination are the abundance of bacteria in the samples and the DNA concentration differences between the mock and the sample. The DNA concentration difference also has a significant effect. Therefore, it is advised for future sequencing studies to keep the DNA concentrations among samples and with mock communities consistent to minimise cross-contamination.

6. References

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., List, M., & Neuhaus, K. (2021). Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *MSphere*, *6*(1). https://doi.org/10.1128/MSPHERE.01202-20/SUPPL_FILE/MSPHERE.01202-20-ST005.PDF

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. (2019). 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

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/NMETH.3869

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

Dyrhovden, R., Rippin, M., Øvrebø, K. K., Nygaard, R. M., Ulvestad, E., & Kommedal, Ø. (2021). Managing Contamination and Diverse Bacterial Loads in 16S rRNA Deep Sequencing of Clinical Samples: Implications of the Law of Small Numbers. *MBio*, *12*(3). https://doi.org/10.1128/MBIO.00598-21

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

Goig, G. A., Blanco, S., Garcia-Basteiro, A. L., & Comas, I. (2020). Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. *BMC Biology*, *18*(1). https://doi.org/10.1186/S12915-020-0748-Z

Holm, J. B., Humphrys, M. S., Robinson, C. K., Settles, M. L., Ott, S., Fu, L., Yang, H., Gajer, P., He, X., McComb, E., Gravitt, P. E., Ghanem, K. G., Brotman, R. M., & Ravel, J. (2019). Ultrahigh-Throughput Multiplexing and Sequencing of >500-Base-Pair Amplicon Regions on the Illumina HiSeq 2500 Platform. *MSystems*, *4*(1). https://doi.org/10.1128/MSYSTEMS.00029-19/SUPPL_FILE/MSYSTEMS.00029-19-SF007.PDF

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

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772. https://doi.org/10.1093/MOLBEV/MST010

Kishore, D., Birzu, G., Hu, Z., DeLisi, C., Korolev, K. S., & Segrè, D. (2023). Inferring microbial co-occurrence networks from amplicon data: a systematic evaluation. *MSystems*. https://doi.org/10.1128/MSYSTEMS.00961-22/SUPPL_FILE/MSYSTEMS.00961-22-S0010.PDF

Manetsberger, J., Caballero Gómez, N., Soria-Rodríguez, C., Benomar, N., & Abriouel, H. (2023). Simply Versatile: The Use of Peribacillus simplex in Sustainable Agriculture. *Microorganisms*, *11*(10). https://doi.org/10.3390/MICROORGANISMS11102540

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*, *8*(4), e61217. https://doi.org/10.1371/JOURNAL.PONE.0061217

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3). https://doi.org/10.1371/JOURNAL.PONE.0009490

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 Core Team. (2023). *R: A Language and Environment for Statistical Computing*. 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

Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

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

Son, B., Kim, Y., Yu, B., & Kong, M. (2023). Isolation and Characterization of a Weizmannia coagulans Bacteriophage Youna2 and Its Endolysin PlyYouna2. *Journal of Microbiology and Biotechnology*, *33*(8), 1050–1056. https://doi.org/10.4014/JMB.2303.03021

Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics*, *3*(1). https://doi.org/10.1093/NARGAB/LQAB019

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

Weissensteiner, H., Forer, L., Fendt, L., Kheirkhah, A., Salas, A., Kronenberg, F., & Schoenherr, S. (2021). Contamination detection in sequencing studies using the mitochondrial phylogeny. *Genome Research*, *31*(2), 309–316. https://doi.org/10.1101/GR.256545.119/-/DC1

7. Appendix

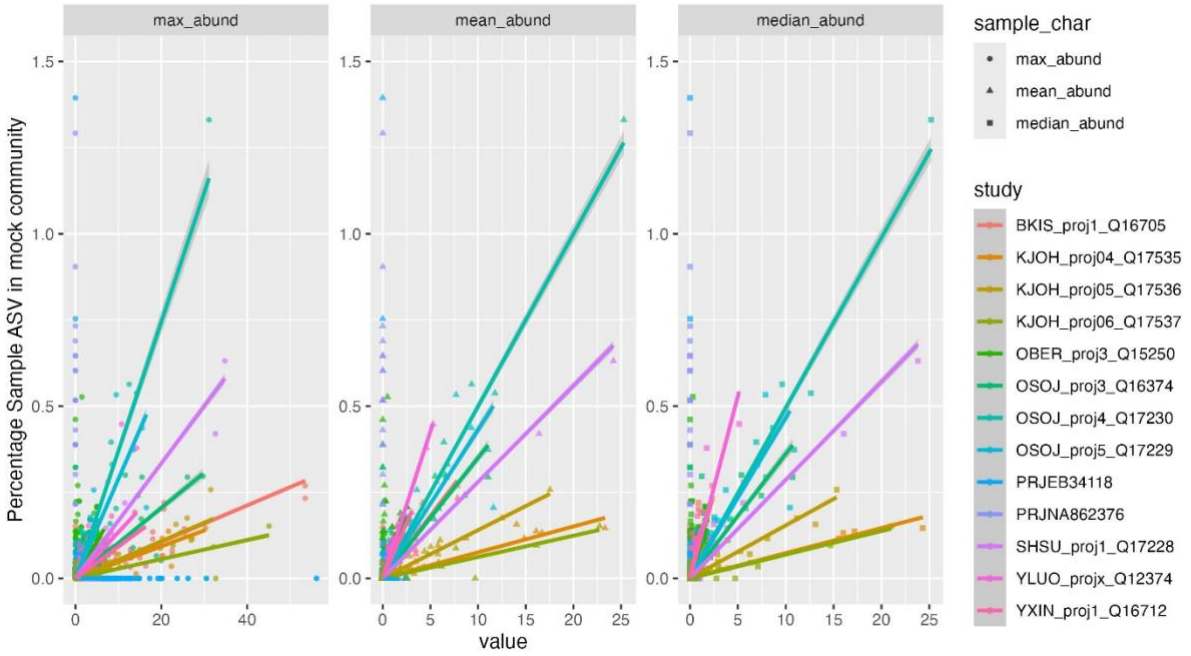


Figure 11 Relationship between ASV abundance, prevalence, and cross-contamination – A scatterplot displaying the percentage of sample ASV in mock communities plotted against the values (maximum, mean, median abundance and prevalence). Each data point represents an ASV, with different studies distinguished by colours and sample characteristics (maximum, mean, median abundance, and prevalence) indicated by shapes.

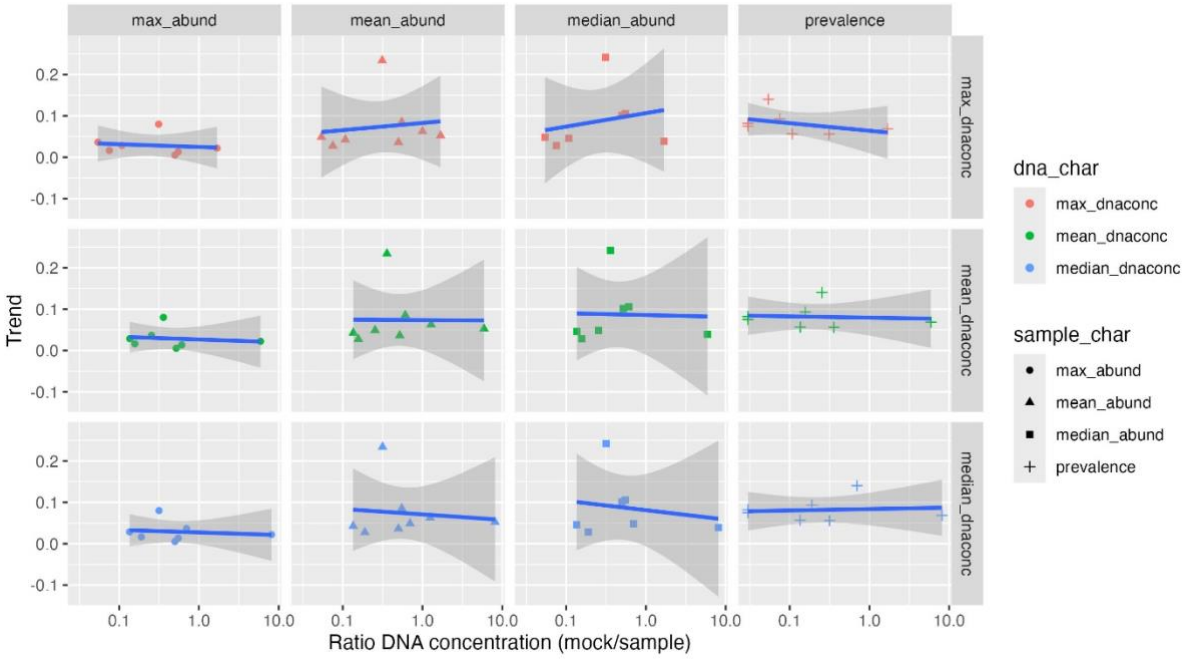


Figure 12 Relationship between trend and DNA concentration ratio – A scatterplot displaying DNA concentration ratio plottea against the trend. Different shapes represent sample characteristics (maximum, mean, median abundance, and prevalence) within each column. The colours represent different DNA characteristics (maximum, mean and median DNA concentration) for each row.

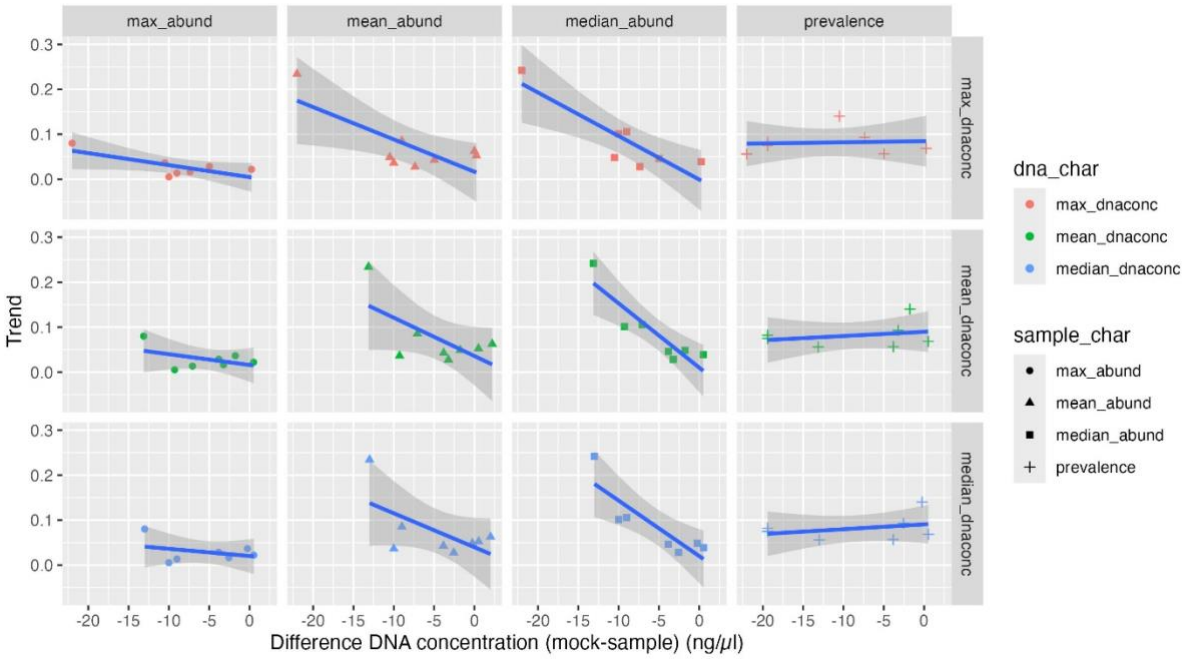


Figure 13 Relationship between trend and DNA concentration difference – A scatterplot displaying DNA concentration difference plotted against the trend. Different shapes represent sample characteristics (maximum, mean, median abundance, and prevalence) within each column. The colours represent different DNA characteristics (maximum, mean and median DNA concentration) for each row.