Identification and Mitigation of microbial contaminant reads from long-read metagenomic samples

Stefany Ayala-Montano[1], Ayorinde O. Afolayan[1], Raisa Kociurzynski[1], Ulrike Loeber[2,3,4,5], Sandra Reuter[1]

[1]Institute for Infection Prevention and Control, Medical Centre, University of Freiburg, Freiburg, Germany

[2] Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany.

[3] Charité – Universitätsmedizin Berlin, Berlin, Germany

[4] Experimental and Clinical Research Center (ECRC), a cooperation between the Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association and the Charité, Berlin, Germany

[5]German Centre for Cardiovascular Research (DZHK), Partner Site Berlin, Berlin, Germany

Abstract

Metagenomic sequencing has revolutionized our understanding of microbial communities, but the presence of contaminant DNA, particularly from the host, poses a significant challenge to accurate data interpretation. We present a methodology for the detection and removal of contaminant sequences in metagenomic datasets, focusing on microbial DNA as a primary contaminant. By integrating metrics such as the prevalence method proposed in decontam, and the coverage value per species per sample, we contributed to the remaining challenge of microbial contaminants that mislead the biological findings. We have emphasized the relevance of human DNA contamination removal (hDNA) in the laboratory and bioinformatically, especially in low-biomass environments.

We highlight that failure to account for human and microbial DNA can lead to erroneous conclusions about the microbial diversity and community composition in the metagenomes. Our approach provides a solution for contamination management for people with moderate bioinformatics skills to ensure more reliable and reproducible results in metagenomic research. The implications of this study for clinical microbiome and pathogen surveillance employing long-read sequencing are relevant and needed for comprehensive contamination control strategies.

Introduction

Metagenomic and microbiome studies have revolutionized our understanding of microbial communities, shedding light on their composition, functions, and interactions with hosts and the environment. These studies have broad applications, from uncovering ecological interactions to understanding the roles of microbes in human health and disease. Traditionally, microbiome research relied on 16S rRNA gene sequencing for taxonomic profiling, but this approach is limited in its applications. With advancements in sequencing technologies using high-throughput platforms such as Illumina technology and now Oxford Nanopore Technology (ONT), researchers can achieve strain resolution and gain more comprehensive insights into microbial dynamics and diversity. However, as the field has progressed, it is becoming increasingly evident that contamination remains a challenge, particularly for low-biomass samples, where external and internal contaminants can skew results and significantly impact the accuracy and reliability of data interpretation [1].

Low-biomass studies, such as those involving neonatal samples or other clinical specimens with limited microbial content, are especially susceptible to contamination. Contaminants can originate from external sources such as laboratory surfaces, reagents, and instruments/equipment, or internal sources, including cross-contamination, arising during sample processing or index switching/hopping in sequencing. These contaminants are particularly problematic in studies targeting specific taxa, such as Enterobacterales, as their presence may be falsely attributed to biological significance. For instance, in transmission studies, contamination could lead to erroneous conclusions about microbial transmission events. Similarly, in microbiome profiling/metataxonomic assignment, contaminants can shift taxonomic profiles, potentially masking true microbial signals and biasing ecological interpretations. While these issues may be less relevant in high-biomass samples, such as soil, where contaminants are more readily distinguishable, their impact in clinical and low-biomass studies necessitates stringent measures for identification and mitigation [2].

A major limitation in the field is the frequent omission of negative and technical controls, for instance, studies have shown that nearly 70% of the metagenomic studies do not include such controls, and only a fraction of those that sequence and

analyze them [3]. This oversight complicates the ability to distinguish true microbial signals from contaminant DNA. Additionally, certain contaminant taxa—over 60 of which have been identified in the literature—are context-dependent, varying based on the specific kits and laboratory environments used. These taxa collectively shape the "contabiome" [4] and their acknowledgment is crucial for interpreting results and making informed decisions in microbiome research [5]. Despite mitigation efforts, such as UV treatment, DNA purification, and segregated workflows, contamination remains a persistent challenge [5][7].

To address this, several tools have been developed for bacterial contaminant identification, including Decontam [3], SourceTracker [7], and MicrobIEM [8]. These tools employ different methodologies to differentiate true signals from contaminants. Decontam, for example, uses frequency-based methods (which assume that contaminant DNA is inversely correlated with input DNA) or prevalence-based methods (which compare the presence of taxa in negative controls versus biological samples) [9]. SourceTracker employs a Bayesian model to estimate the proportion of n sequences to a defined sink sample x by using the Gibbs distribution [7] and forms a part of the open-access QIIME tools for metagenome and 16S rRNA gene amplicons [9]. MicrobIEM incorporates the prevalence method of Decontam and provides options based on the ratio of the species mean relative abundance that requires environmental sampling [8]. However, most of these tools rely on sequencing data from negative or environmental controls, which are sometimes excluded from studies [5].

In this study, we investigated contamination in metagenomic samples obtained from low-biomass neonatal screening samples collected weekly from premature and term neonates in the neonatal intensive care unit (NICU) at the Unversity of Freiburg Medical Centre. These samples, analyzed using ONT long-read sequencing, presented unique challenges, as low-biomass samples are particularly prone to contamination due to their low microbial load, high host DNA content, and susceptibility to technical artifacts [12][13]. To tackle these challenges, we adopted and expanded the prevalence method proposed by Decontam. By incorporating a novel metric—coverage per species per sample—we developed a system called "Stop-Check-Go" for identifying and addressing potential external and internal contaminants. This system leverages the advantages of long-read sequencing, which enables more comprehensive genome and metagenome characterization, providing

critical insights into microbial diversity and its implications for health and disease [9][11].

Through this approach, we not only accounted for contamination but also highlighted the importance of implementing negative controls and integrating them into metagenomic analyses to enhance the reliability of results. By addressing contamination systematically, we aim to improve the accuracy of strain-level microbial profiling, particularly for Enterobacterales, where contamination poses unique challenges in understanding microbial transmission and microbiome dynamics in neonatal health [3][4].

Methods

Collection of Samples from Neonates at the Neonatal Intensive Care Unit (NICU)

Rectal and nasal swab samples were routinely collected from neonates admitted to the NICU within the framework of a study that investigated the emergence, spread, and dynamics of healthcare-associated bacteria with the intent of tracking the acquisition of pathogens in real time (TAPIR). We included sterile swabs and molecular grade water as weekly negative controls, which were processed and sequenced alongside the biological samples. All samples were enriched overnight with BHI in a proportion 1:2 sample:media at 37°C, 210 rpm. We validated the microbial growth by culturing on MacConkey, CHROMagar™/ CNA Agar (Biplate), and blood agar at 37°C for 24 h. A total of 1.351 samples were collected from 106 patients. Among them, 25 patients were microbiologically negative and remained so, as they were discharged before a second screening could be performed. Additionally, 126 negative control samples (from swabs and water) were analyzed weekly, bringing the total to 1,477 samples analyzed.

DNA Extraction, Library Preparation, and Metagenomic Sequencing

Samples were processed for human DNA removal by centrifuging at 10.000 rpm for 5 minutes and the incubation of the microbial pellet with 1 mL of molecular grade water at room temperature (RT) for 5 minutes, promoting cellular lysis. A final resuspension of the pellet in PBS 1X and 1:1 lysozyme + lysostaphin was used for DNA extraction using the kit High Pure PCR Template (Roche Molecular Diagnostics, Mannheim, Germany) following the manufacturer's protocol.

Libraries for long-read sequencing were prepared using the ligation sequencing kit (SQL-LSK109; ONT, Oxford, UK) and native barcoding kits (EXP-NBD104 & EXP-NBD114) for R9 Chemistry, as well as the SQK-NBD114-96 barcoding kit for R10 chemistry, following the manufacturer's protocol (available online in https://nanoporetech.com). In the R10 chemistry, we used an initial volume of 11 uL per sample in the library protocol to get the maximum possible concentration (low biomass samples ranged from too low to 3,5 uL after BHI enrichment). Final libraries were sequenced on a GridION X5 Mk1 sequencing platform. Sequence data acquisition, real-time base-calling, and demultiplexing of barcodes were conducted using the graphical user interface MinKNOW (v23.11.7) and the dorado basecall server (v7.2.13). Concurrently, Readfish (v23.1.1) was utilized during the sequencing run to actively "reject" sequencing reads corresponding to human DNA across individual nanopores [14].

Read Pre-processing and Removal of Human DNA Contaminants

Sequenced data underwent pre-processing using Porechop (v0.2.4; https://github.com/rrwick/Porechop) to trim off adapters. Trimmed reads were filtered using Filtlong (v0.2.1; https://github.com/rrwick/Filtlong) with a minimum read length threshold ("--min_length") of 1000. Afterwards, reads were mapped to the human genome (GRCh38) using Minimap2 (v2.24; using the "map-ont" parameter) [15] and human DNA contaminants were removed from reads using Samtools (v1.14) [16]. These steps, including read trimming, read filtration, and human DNA removal, were executed using custom Nextflow pipelines [17] (see Table S1). for the codes implemented in Nextflow pipelines). Sequences have been deposited in the ENA in project PRJEB82667.

Microbial Species Abundance Estimation and Visualization of Real-world Metagenomic Datasets

Kraken reports [18] were generated from the analysis of metagenomic sequence data, using Kraken v1.1.1 and MiniKraken DB_8GB. Microbial species abundance was estimated from kraken output using Bracken [19]. Microbial species abundance profiles were visualized in R, using Tidyverse (v2.0.0) [18], scales [21], and ggthemes [22].

The species identified as cross-contaminants were confirmed by MALDI-TOF, sequenced with MiSeq Illumina, and identified to the sequence type level using MLST [23]. Accordingly, the specific species was labeled contaminant in the metagenome database and removed from downstream analysis if the same sequence type was identified with TRACs [24]. Weeks associated only with library contamination (number of reads ≥ 1000) were repeated with the R10 chemistry flow cells with a loading volume of 11 uL, aiming for a good initial concentration of the biological samples.

Results

Implementation of a stop-check-go system

In this study, we identified three types of contaminants: (1) contaminants that appeared in low biomass samples (e.g., in microbiologically negative patients) at higher abundance than non-contaminated low biomass samples, both with low initial DNA concentration; (2) contaminants arising from cross-contamination during sample enrichment, where controls showed positive growth; and (3) contaminants introduced during DNA extraction and library preparation, now termed *library contamination*, irrespective of negative growth.

The species present in the controls (swab and water) were cross-checked with the species present in the samples per week by the direct comparison using the prevalence method published by Davis et. al 2018 [3]. They established that total sample DNA (T) is composed by contaminant (C) and true sample DNA (S), so $T = C + S$. Thus, in negative controls $S \sim 0$ and the probability of contaminant

detection is higher than the detection of contaminants in true samples when $S > 0$ [3]. This means that the species present in the controls are classified as contaminants if the prevalence is higher in the controls: $Prevalence\ controls > Prevalence\ samples$ as the species in the controls have absence of competing DNA in the sequencing process [3].

A combined file was created for the taxonomy reports from Kraken v1.1.1 and complemented with the binary microbiological growth results per sample (positive, negative) and the coverage per species (per sample). The coverage values were obtained from the assembly statistics from FLYE [25] from "--stats" parameter) for all the contigs per sample file. Followed by "kraken-translate" script [16] to get the taxonomic association within each contig, and therefore could associate the "contig_x" to a species before merging with the combined kraken file.

We used a minimum coverage of 20 according to literature [12], to classify the species as "*low*" for coverage < 20X, and "*ok*" for coverage of 20X or above. We calculated the difference of the fraction reads (DFR) per species based on the prevalence method that included the reads fraction of the control (FRC), and the fraction reads of the sample (FRS). The results were classified as "stop" if $DFR = FRC - FRS > 0$, and "go" when $DFR < 0$, meaning that the fraction of the species in the sample is higher than the control and the abundance is not given by chance.

Read Pre-processing and Removal of Human DNA contaminants

Given that swab samples may also contain the patient's epithelial cells, the human DNA (hDNA) contamination was as high as 90% of the sample, therefore it was crucial to reduce the concentration of the hDNA by a lysis protocol [26] before DNA extraction and library preparation. The concentration of the hDNA contaminant decreased significantly with this approach (see Table S2), obtaining on average 76.63% reduction of hDNA as contaminant (IQR: 64.80, 88.01%). The number of human and microbial reads across the samples per week of screening and the negative controls are displayed in Fig. 1. The remaining hDNA contaminant reads were filtered out with Kraken v1.1.1 during the speciation pipeline (Table S1, i).

Classification of non-hDNA contaminants in the system stop-check-go

Three parameters were included in the algorithm: coverage, DFR, and contamination reason (either cross-contamination, library contamination 1K or 10K). The conditions were formulated using the function “case_when” from the dplyr R package, meaning that each condition was evaluated independently. Table 1 summarizes the system decision parameters and the classification into the corresponding category: “stop”, “check”, or “go”.

The classification system strictly excluded potential contaminants, stopping samples if any one parameter indicated an issue (Table 1). For instance, samples were rejected if cross-contamination or library contamination with 10K reads was detected (parameter_3). Similarly, species found in microbiologically negative patients were classified as "stop," as no sequence features were expected. If parameter_1 were included in the second (2) and third conditions (3), the samples would fail due to low coverage, consistent with the presumed low initial DNA concentration.

Table 1. Summary of the decision for the system stop check and go that includes the parameters coverage, DFR, and microbiological results

	Parameter_1	**Parameter_2**	**Parameter_3 (auxiliary)**	**Outcome**
Condition	**Coverage**	**Difference Fraction Reads (DFR)**	**Contamination reason control**	**Decision System**
(1)	low	red	-	stop
(2)	-	-	cross-contamination or lib*. contamination 10K	stop
(3)	-	-	patient microbiologically negative	stop
(4)	ok	red	-	stop
(5)	low	green	no reason	check
(6)	ok	green	no reason	go
(7)	ok	green	library contamination 1K	go

Lib*. Contamination 10K refers to species in controls where the number of reads was higher than 10.000. Spaces denoted with a hyphen were parameters not evaluated in the conditions.

We plotted the reads of the different sample types (water, swab, anus, and nose), and found significant differences in all comparisons (Fig. 3), indicating that all sample types have individual signatures inherent to them. We found the smallest differences between the category of controls and the group of samples (p-adj, Fig. 3). Despite the significant differences, the average number of reads is higher than expected, pre-filtering the data for the known-species contaminant will contribute to decreasing this number, however, we did not find useful this approach as some contaminants are Enterobacterales. We intended to solved this by sequencing with MiSeq (short-read) and then carefully highlighting the same sequence type as cross-contamination after TRACs [24] and MLST prediction [23].

Figure 4 depicts the relation between the average coverage and the number of reads per species in each sample. Using coverage as one of the parameters for the classification of the species, we confirmed that the negative controls are below this defined cutoff (in blue, Fig. 4), except one control above this threshold (water, Fig. 4A) as it represented cross-contamination with *Ralstonia pickettii* species. Approximately 47% of species in the samples had a coverage that fell below our defined cutoff (samples, Fig. 4B). We furgher tested the relationship between the DNA concentration (high: >101 ng/µl, low < 4 ng/µl) and the rejection rate (“stop”), acceptance (“go”), and indecision (“check”) with the Pearson correlation for non-parametric data. We found a significant relationship (p adj < 0.001) but a weak correlation (□-go: 0.23, □-stop: -0.18, □-check: -0.27), which is consistent with the acceptance and rejection rates not being linked to the library concentration, as these rates are similar.

Identification of non-human contaminants and visualization of negative controls

We quantified cross-contamination in 5.53% of the total dataset (n total=4101), which we identified at species level and sequence type (ST). These samples were classified

with the corresponding species as “stop”. On the other hand, library contamination with a number of reads higher than 1 000 (1K) or reads higher than 10 000 (10K) represented 19.04%, and 2.93%, respectively, while 72.57% of the data showed no signs of contamination. We plotted the distribution of species in the control samples per week in Fig. 2. Contamination is visible by a change from a minimally diverse set of species to a dominance of particular species, for example, weeks 18-24 are dominated by *Escherichia coli* contaminant whereas weeks 5, 6, 32, and 33 show a more expected distribution of species. The end of the screening period occurred in week 56 (fourth week, 2024), but we decided to follow the last patient being colonized with Enterobacterales, in this case *Klebsiella oxytoca.* In the coming weeks (57-62), we only sequenced that sample and the negative controls in a single library preparation, this explains the strong similarity between the profiles in the negative controls on those weeks (Fig. 2).

For the weeks where library contamination was suspected, *E. coli* and *Staphylococcus* sp. were significantly abundant and therefore a strong library contaminant (Fig. 2). Upon repetition of the libraries starting from the same DNA extraction, there was a change in the taxonomic profiles after repeating the libraries, however, the read abundance for the different species remained high (> 1K). This change in taxonomic profile supports the suspected contamination of the original samples, therefore we were able to correct contamination in 80% of the repeats dataset (Fig. 2). The high read abundance maybe attributed to the use of the Readfish tool for selectively rejecting the reads mapped to the human reference genome (GRCh38), and indirectly enriching all microbial reads.

Relative abundance of species classified in the system stop-check-go

Figure 5 shows a general overview of relative frequency in the classification system go and stop at the genus level by weeks. This gives a percentage that reflects the contribution of the species relative to the total dataset. The species with coverage “ok” and DFR “green” with no signs of contamination in the controls and/or library contamination 10K were classified as “go”. *Staphylococcus* sp., and *Klebsiella* sp., showed the highest relative abundance in the system “go” and “stop” (Fig. 5). Figure S1 similarly depicts the species classified as “check” because the coverage

(parameter_2) was below the defined threshold where *Staphylococcus* sp., *Enterococcus* sp., *Klebsiella* sp., and *E.coli* had the uppermost frequencies in this classification system.

Figure S2 highlights the abundance of species colored by the sample site, following the principle that the true sample $S \sim 0$, the near-zero counts and the negative bars corresponded to species suspected as contaminants such as *Salmonella* sp. in the nose.

The most abundant species were *S. aureus*, *S. epidermidis,* and *S. capitis* followed by *Streptococcus* sp., *Klebsiella pneumoniae,* and *K. oxytoca,* along with Enterobacter cloacae complex.

The system classified the majority of the samples as contaminants, with 51% in the "stop" category (n=2093, n total=4104), 19% of the species to "check" (n=778), and 30% to "go" (n=1233). The species in the check system were reclassified as "to-stop-after-check" after the confirmation that 33% of the species present in the samples belonged to microbiologically negative patients (n=258).

Figure S3 displays the percentage of filtered species separated by source and the corresponding average coverage distribution. For the filtered species, there were filtered bacteria that were not shared between the sources, such as *Bacteroides* sp., and *Enterococcus faecium* in the anus. In the nose, *Streptococcus* sp. was almost exclusively removed except for *S. parasanguinis*, which was present in both sources (Fig. S3). There were also species that we classified as "go" in a 100% of the samples, such as *Acinetobacter soli, A. schindleri*, *K. quasipneumoniae, K. variicola,* and *Salmonella enterica* (Fig. S3).

Discussion

We developed a classification system to identify bacterial contaminants below a defined coverage threshold and complemented with coverage together with microbiological growth results of the control samples. The three parameters included in the system helped to determine those patients that are microbiologically negative but do show significant bacterial reads as likely contaminated samples. Metagenomic

studies frequently do not carry out microbiological growth controls, as their samples may include unculturable species, and hence, a coverage threshold can be employed. However, the coverage statistics by species (per sample) in the metagenomes require a big bioinformatics infrastructure for running assembly processes. In addition, the coverage association to each species demands the use of a Kraken database and the assemblies (in our case from FLYE). The completeness of the genomes through binning may also present an alternative option reported [27] but not explored yet in this study.

The system stop-check-go was useful for discriminating potential contaminants from those meaningful species in the biological sample. The species classified in the "check" system due to coverage below 20 need further individual analysis to determine if they belong to low biomass samples that could not be increased in the DNA concentration despite the enrichment step. In these cases, we could study the positive linear correlation between the initial DNA concentration and the number of reads expected for non-contaminant species as reported in the literature [3]. The presence of the species in the stop-check-go system should be context-depended and can be extensively evaluated, for example, *Rothia* and *Prevotella (Fig. S1, Fig. S2) are* frequently part of the core microbiota in clinical microbiomes, and the baseline of lung microbiomes rather than contaminants [26]. Similarly, as we are particularly interested in tracking Enterobacterales, species such as *Klebsiella* sp., and *E.coli* with peak frequencies in the "check" system could be extensively analyzed. One approach is to compare the detailed microbiological reports and the ST (sequenced with MiSeq Illumina from pure culture) with TRACs [24] results from patients' metagenomes. Then, if there is an agreement between these results, selectively include only this species in our dataset regardless of the low coverage for further transmission events analyses.

We have emphasized in this paper the removal of microbial contaminants, even so, the human DNA (hDNA) contaminants removal is also a critical step in ensuring the accuracy and reliability of metagenomic studies [27], especially in low-biomass samples such as swabs [2]. It is highlighted that in host-associated metagenomes, such as human or mouse samples, human DNA contamination was the most prevalent, f.e. it was found that 72% of the metagenomes in one study had at least

one potential human contamination [27]. Given that the samples in our collection often contain a substantial amount of host DNA—primarily from epithelial cells—the challenge of managing hDNA contamination cannot be overstated. Failure to address this issue risks diluting microbial signals, obscuring the detection of clinically relevant organisms, and increasing the likelihood of false-positive results.

Our approach underscores the importance of integrating both experimental and computational strategies to address microbial and hDNA contamination, especially in low biomass samples. While physical reduction of hDNA through a lysis protocol reduced contamination levels substantially, computational filtering using Kraken allowed for further refinement. This layered strategy ensures a cleaner dataset for microbial analyses, enabling researchers to better resolve the microbial diversity and to detect low-abundance pathogens with confidence. The benefits of systematic contaminant removal extend beyond technical accuracy. In clinical research, precise microbial profiling is essential for understanding host-pathogen dynamics, tracking transmission events, and informing patient care. For instance, the presence of residual hDNA or microbial contaminants in datasets could mislead efforts to identify clinically relevant organisms or obscure microbial community interactions. By prioritizing the removal of hDNA contaminants, we reduce noise in the dataset and create a clearer picture of the microbial landscape, while addressing microbial contaminants ensures the reliability of downstream taxonomic and functional analyses.

Microbial contaminants can also distort the taxonomic profiles, overstating the presence of certain taxa (e.g., Enterobacterales in our study), and lead to erroneous conclusions about pathogen transmission or microbiome structure. This is the primary reason why we decided to complement our metagenomic analyses from long-read (Nanopore) with short-read sequencing (Illumina), which relies on culturing methods where employing selective media agar is an advantage. Importantly, not all studies incorporate sufficient controls to distinguish contaminants from true community members, increasing the risk of misinterpretation. Including negative controls and applying robust computational tools, as demonstrated here, can help mitigate these issues. These strategies are particularly critical in low-biomass settings but can be broadly applied across diverse metagenomic studies. As

sequencing technologies evolve, the ability to effectively manage contaminants will remain central to producing high-quality data and uncovering meaningful biological insights.

We argue that using coverage as one of the classification parameters outperforms the usage of simply the read counts as the coverage adjusts for read length and genome size. This accounts for bacteria with, for example, small genome sizes which might be falsely excluded when only the read counts were used. The analysis of species from controls indicates that a high threshold needs to be set to exclude the majority of the species in case when read counts are used as the classification parameter. This might possibly lead to a large amount of falsely excluded species. In contrast, coverage provides a reliable measure as most of the species from the negative controls are excluded. However, as a simple threshold is not a sufficient measure to detect contamination, species with a coverage above the threshold can also fail the classification system or vice versa, depending on the outcome of other parameters.

We found the inclusion of the prevalence method robust for the identification of bacterial contaminants, so we compared the prevalence (only) method to the ratio method proposed by MicrobIEM [8] (data not shown), and the agreement percentages varied from 90.23% to 98.89% for all the categories in the system. As a limitation, we state that due to technical obstacles the data presented in this paper represents 53% of the total collection (n=28/52 weeks). Also, we assume that the repeated samples had the best library quality because 1) the given lower error rate and the improvement in performance of R10 chemistry, and 2) a rise in the loading volume (initial concentration).

Conclusion

The use of the negative controls, validation with microbiological growth, and well-defined thresholds for the variables/parameters included in the system stop-check-go were good approaches to assess the contaminant species. The algorithm described in this paper is a methodology that included laboratory and bioinformatics work and

did not explicitly model species contamination. It is important to exhaustively remove microbial and human DNA contamination in the laboratory and bioinformatically as both often represent misleading results with weak biological meaning. In the context of our aim to track Enterobactales on the Neonatal Intensive Care Unit, the enrichment step was advantageous to enrich the microbial reads for those species, and helped in most cases to classify those as in the “go” system. We are aware that the metagenomes in this study did not represent the actual microbiome of the patients due to the enrichment step. The prevalence method published in the decontam algorithm [3] continues to be a valuable approach to identify potential contaminants but might be complemented with additional metrics for increasing the confidence in the results.

List of Figures, Tables, and Supplementary Files (Figures and Tables)

Fig. 1. Quantification of human and microbial reads in the samples, categorized by chronological week of screening, source and negative control groups. On the x-axis (representing weeks), 1_y2, 2_y2, and 3_y2 correspond to the initial weeks of the year 2024.

Fig. 2. Taxonomic profile of chronologically-ordered controls (swab and water). Double white stars on stacked bars represent species with ≥10,000 reads assigned to them, while a single white star indicates species with 1,000 to 10,000 reads. Species with less than 1,000 reads are denoted by no star. Gray represents unassigned reads, and black other microorganisms. Samples termed as repeats correspond to library repetitions with R10 flow cell chemistry.

Fig. 3 Average number of reads (log10) differs significantly between sample types. Comparison with Wilcoxon test evidenced in statistical significance between the groups (p. adj. < 0.05).

Fig. 4. Average coverage per species in the metagenome versus the number of total reads. The blue dots indicate species with coverage below the 20-read threshold. Panel **A** displays the coverage for controls, while Panel **B** presents the coverage for samples.

Fig. 5. Overall contribution of species classified as "go" and "stop" within the system, showing the relative frequency ("Rel_freq") of bacterial genera in the collected samples. Darker regions represent higher relative abundances, while values below 0.01 are not labeled.

Fig. S1. Species abundance colored by sample site. The near-zero count and the negative bars indicate potential contaminants where the total count in control reads exceeded the count in sample reads, following the principle $S \sim 0$.

.Fig. S2. Overview of the species classified as “check” in the system and the relative abundance (indicated as "Rel_freq") of the bacterial genera in the collected samples. Darker regions indicate higher relative abundances. Values below 0.01 are not labeled.

Fig. S3. Percentage of the filtered species, separated by source and the corresponding average coverage distribution. The species with a coverage below the threshold (< 20) are coloured in blue. Species with a coverage higher than 400 are not shown (max. coverage 1500).

Table S1. Code availability for the specified processes that were implemented in Nextflow pipelines [17].

Bioinformatic process	**Source**
Read trimming and filtering	https://github.com/ayoraind/ONT_adapter_removal_and_read_filtration
Human DNA contaminant removal	https://github.com/ayoraind/hDNA_removal_and_mapping_stats
Assembly	https://github.com/ayoraind/flye
Taxonomic classification (i. Kraken, ii.	i) Kraken: https://github.com/ayoraind/kraken ii) https://github.com/ayoraind/bracken

Bracken)	
Pairwise transmission distances	https://github.com/ayoraind/tracm_nextflow

Table S2. Average in the proportion of the reads mapped to the human reference (GRCh38). Unmapped reads refers to microbial reads.

References

[1] L. Cornet and D. Baurain, “Contamination detection in genomic data: more is not enough,” *Genome Biology*, vol. 23, no. 1, p. 60, Feb. 2022, doi: 10.1186/s13059-022-02619-9.

[2] J. M. Martí, "Recentrifuge: Robust comparative analysis and contamination removal for metagenomics," *PLoS Computational Biology*, vol. 15, no. 4, p. e1006967, Apr. 2019. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006967.

[3] N. M. Davis, D. M. Proctor, S. P. Holmes, D. A. Relman, and B. J. Callahan, “Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data,” *Microbiome*, vol. 6, no. 1, p. 226, Dec. 2018, doi: 10.1186/s40168-018-0605-2.

[4] H. Jurasz, T. Pawłowski, and K. Perlejewski, “Contamination Issue in Viral Metagenomics: Problems, Solutions, and Clinical Perspectives,” *Frontiers in Microbiology*, vol. 12, 2021, doi: 10.3389/fmicb.2021.745076.

[5] Y. C. Lou *et al.*, “Using strain-resolved analysis to identify contamination in metagenomics data,” *Microbiome*, vol. 11, no. 1, p. 36, Mar. 2023, doi: 10.1186/s40168-023-01477-2.

[6] R. Eisenhofer, J. J. Minich, C. Marotz, A. Cooper, R. Knight, and L. S. Weyrich, “Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations,” *Trends in Microbiology*, vol. 27, no. 2, pp. 105–117, Feb. 2019, doi: 10.1016/j.tim.2018.11.003.

[7] D. Knights *et al.*, “Bayesian community-wide culture-independent microbial source tracking,” *Nature Methods*, vol. 8, no. 9, pp. 761–763, Sep. 2011, doi: 10.1038/nmeth.1650.

[8] C. Hülpüsch *et al.*, “Benchmarking MicrobIEM – a user-friendly tool for decontamination of microbiome sequencing data,” *BMC Biology*, vol. 21, no. 1, p. 269, Nov. 2023, doi: 10.1186/s12915-023-01737-5.

[9] C. Kim, M. Pongpanich, and T. Porntaveetus, “Unraveling metagenomics through long-read sequencing: a comprehensive review,” *Journal of Translational Medicine*, vol. 22, no. 1, p. 111, Jan. 2024, doi: 10.1186/s12967-024-04917-1.

[10] R. Henry *et al.*, “Into the deep: Evaluation of SourceTracker for assessment of faecal contamination of coastal waters,” *Water Research*, vol. 93, pp. 242–253, 2016, doi: https://doi.org/10.1016/j.watres.2016.02.029.

[11] L. Zhang *et al.*, “Advances in Metagenomics and Its Application in Environmental Microorganisms,” *Frontiers in Microbiology*, vol. 12, 2021, doi: 10.3389/fmicb.2021.766364.

[12] J. Shen, A. G. McFarland, V. B. Young, M. K. Hayden, and E. M. Hartmann, “Toward Accurate and Robust Environmental Surveillance Using Metagenomics,” *Frontiers in Genetics*, vol. 12, 2021, doi: 10.3389/fgene.2021.600111.

[13] J. R. Bedarf *et al.*, England. *Much ado about nothing? Off-target amplification can lead to false-positive bacterial brain microbiome detection in healthy and Parkinson’s disease individuals.*, (Mar. 26, 2021). doi: 10.1186/s40168-021-01012-1.

[14] A. Payne, N. Holmes, T. Clarke, R. Munro, B. J. Debebe, and M. Loose, “Readfish enables targeted nanopore sequencing of gigabase-sized genomes,” Nature Biotechnology, vol. 39, no. 4, pp. 442–450, Apr. 2021, doi: 10.1038/s41587-020-00746-x.

[15] H. Li, ‘Minimap2: Pairwise alignment for nucleotide sequences’, *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, Sep. 2018, doi: 10.1093/bioinformatics/bty191.

[16] H. Li *et al.*, ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[17] P. DI Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, ‘Nextflow enables reproducible computational workflows’, *Nature Biotechnology 2017 35:4*, vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.

[18] Wood, D. E., and Salzberg, S. L., "Kraken: Ultrafast metagenomic sequence classification using exact alignments," *Genome Biol.*, vol. 15, pp. 1–12, 2014.

[19] Lu, J., Breitwieser, F. P., Thielen, P., and Salzberg, S. L., "Bracken: Estimating species abundance in metagenomics data," *PeerJ Comput. Sci.*, vol. 2017, p. e104, 2017.

[20] Wickham, H., Averick, M., and Bryan, W. C., "Welcome to the tidyverse," *J. Open Source Softw.*, vol. 4, p. 16, 2019.

[21] Scales Functions for Visualization, "Scales," [Online]. Available: https://scales.r-lib.org/. [Accessed: Apr. 4, 2024].

[22] Extra Themes, Scales and Geoms for ggplot2 (ggthemes)," [Online]. Available: https://jrnold.github.io/ggthemes/. [Accessed: Apr. 4, 2024].
[23] K. Jolley and M. Maiden, "This publication made use of the PubMLST website," *BMC Bioinformatics*, vol. 11, p. 595, 2010. [Online]. Available: https://pubmlst.org/. The website, hosted at the University of Oxford, was funded by the Wellcome Trust.
[24] A. O. Afolayan, S. A. Montano, I. J. Akintayo, L. Duarte dos Santos, and S. Reuter, “Assessing the performance of current strain resolution tools on long-read metagenomes,” *bioRxiv*, 2024, doi: 10.1101/2024.11.20.624313.

[25] M. Kolmogorov *et al.*, “metaFlye: scalable long-read metagenome assembly using repeat graphs,” *Nature Methods*, vol. 17, no. 11, pp. 1103–1110, Nov. 2020, doi: 10.1038/s41592-020-00971-x

[26] M. Shehadul Islam, A. Aryasomayajula, and P. Selvaganapathy, "A review on macroscale and microscale cell lysis methods," *Micromachines (Basel)*, vol. 8, p. 83, 2017.

[27] A. K. T. H. Labadie, F. A. R. Silva, and R. B. V. P. Smith, "Genome completeness through binning in metagenomes: A review of methods and applications," *J. Microbiol. Methods*, vol. 136, pp. 55-66, Mar. 2018. doi: 10.1016/j.mimet.2018.01.010.

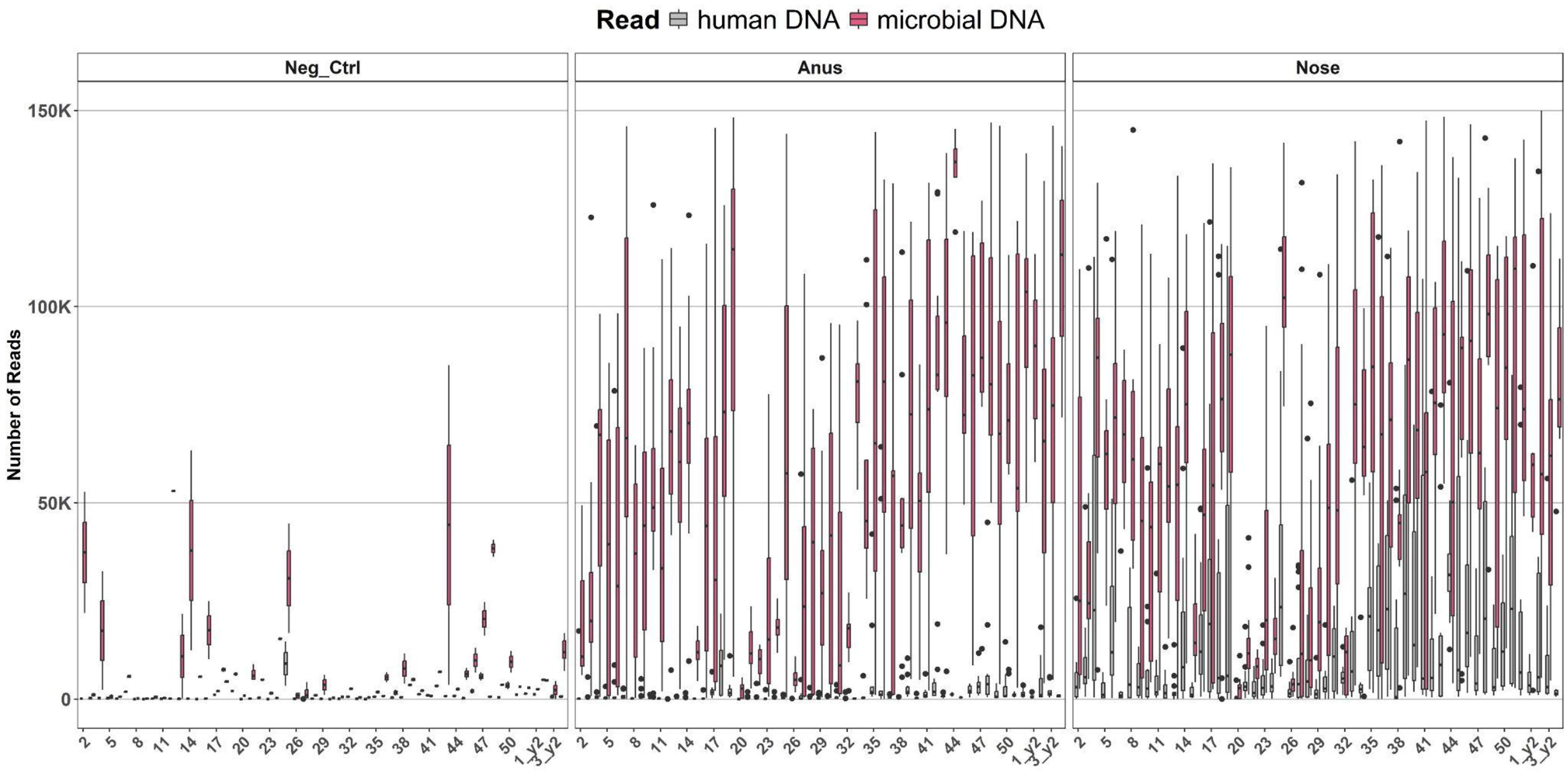
Read human DNA microbial DNA
Neg_Ctrl
Anus
Nose
Number of Reads
150K
100K
50K
0
2 5 8 11 14 17 20 23 26 29 32 35 38 41 44 47 50 1_y2 3_y2

Fig. 1

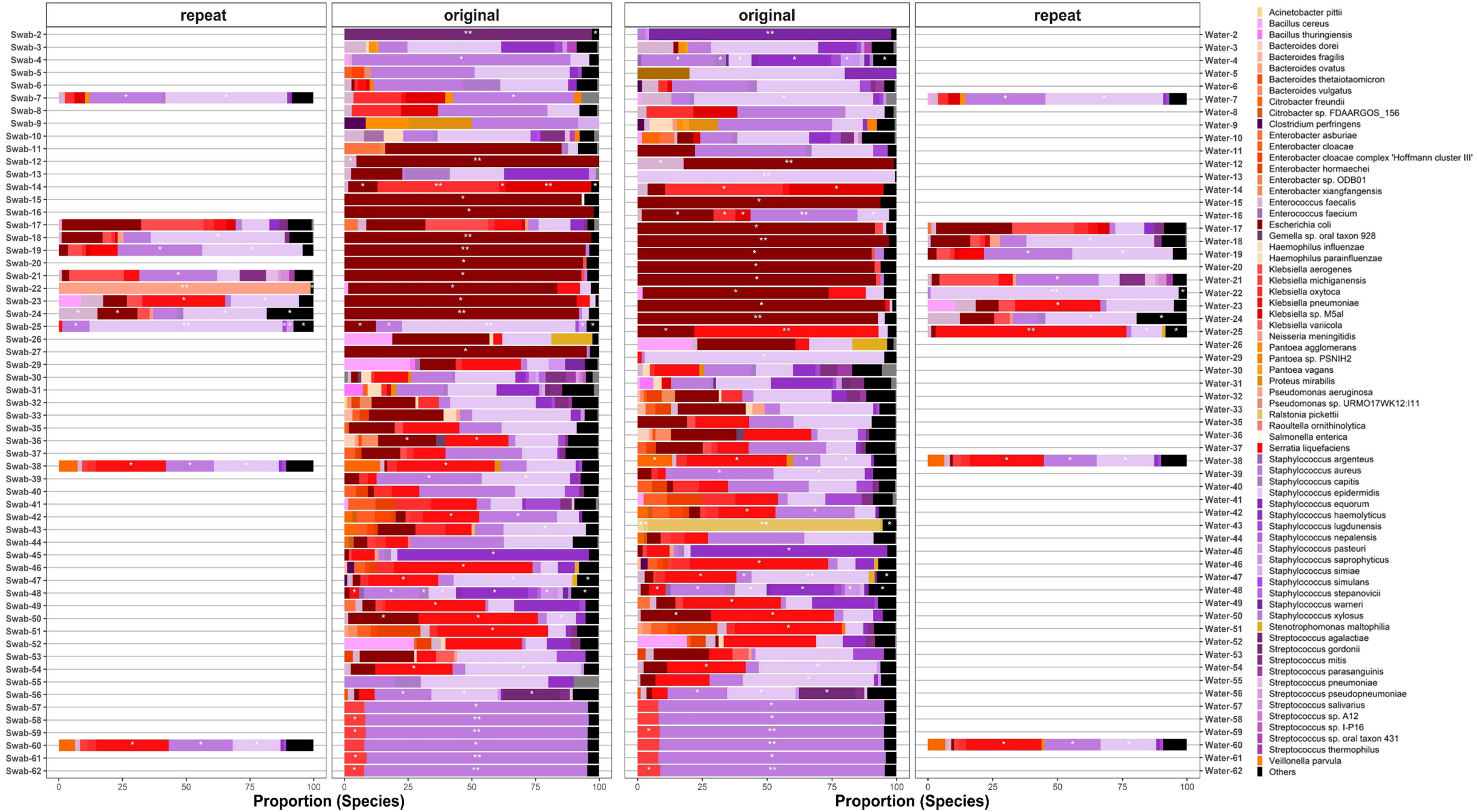
repeat
original
original
repeat
Swab-2
Swab-3
Swab-4
Swab-5
Swab-6
Swab-7
Swab-8
Swab-9
Swab-10
Swab-11
Swab-12
Swab-13
Swab-14
Swab-15
Swab-16
Swab-17
Swab-18
Swab-19
Swab-20
Swab-21
Swab-22
Swab-23
Swab-24
Swab-25
Swab-26
Swab-27
Swab-29
Swab-30
Swab-31
Swab-32
Swab-33
Swab-35
Swab-36
Swab-37
Swab-38
Swab-39
Swab-40
Swab-41
Swab-42
Swab-43
Swab-44
Swab-45
Swab-46
Swab-47
Swab-48
Swab-49
Swab-50
Swab-51
Swab-52
Swab-53
Swab-54
Swab-55
Swab-56
Swab-57
Swab-58
Swab-59
Swab-60
Swab-61
Swab-62
Water-2
Water-3
Water-4
Water-5
Water-6
Water-7
Water-8
Water-9
Water-10
Water-11
Water-12
Water-13
Water-14
Water-15
Water-16
Water-17
Water-18
Water-19
Water-20
Water-21
Water-22
Water-23
Water-24
Water-25
Water-26
Water-29
Water-30
Water-31
Water-32
Water-33
Water-35
Water-36
Water-37
Water-38
Water-39
Water-40
Water-41
Water-42
Water-43
Water-44
Water-45
Water-46
Water-47
Water-48
Water-49
Water-50
Water-51
Water-52
Water-53
Water-54
Water-55
Water-56
Water-57
Water-58
Water-59
Water-60
Water-61
Water-62
0
25
50
75
100
Proportion (Species)
Proportion (Species)
Acinetobacter pittii
Bacillus cereus
Bacillus thuringiensis
Bacteroides dorei
Bacteroides fragilis
Bacteroides ovatus
Bacteroides thetaiotaomicron
Bacteroides vulgatus
Citrobacter freundii
Citrobacter sp. FDAARGOS_156
Clostridium perfringens
Enterobacter asburiae
Enterobacter cloacae
Enterobacter cloacae complex 'Hoffmann cluster III'
Enterobacter hormaechei
Enterobacter sp. ODB01
Enterobacter xiangfangensis
Enterococcus faecalis
Enterococcus faecium
Escherichia coli
Gemella sp. oral taxon 928
Haemophilus influenzae
Haemophilus parainfluenzae
Klebsiella aerogenes
Klebsiella michiganensis
Klebsiella oxytoca
Klebsiella pneumoniae
Klebsiella sp. M5al
Klebsiella variicola
Neisseria meningitidis
Pantoea agglomerans
Pantoea sp. PSNIH2
Pantoea vagans
Proteus mirabilis
Pseudomonas aeruginosa
Pseudomonas sp. URMO17WK12:I11
Ralstonia pickettii
Raoultella ornithinolytica
Salmonella enterica
Serratia liquefaciens
Staphylococcus argenteus
Staphylococcus aureus
Staphylococcus capitis
Staphylococcus epidermidis
Staphylococcus equorum
Staphylococcus haemolyticus
Staphylococcus lugdunensis
Staphylococcus nepalensis
Staphylococcus pasteuri
Staphylococcus saprophyticus
Staphylococcus simiae
Staphylococcus simulans
Staphylococcus stepanovicii
Staphylococcus warneri
Staphylococcus xylosus
Stenotrophomonas maltophilia
Streptococcus agalactiae
Streptococcus gordonii
Streptococcus mitis
Streptococcus parasanguinis
Streptococcus pneumoniae
Streptococcus pseudopneumoniae
Streptococcus salivarius
Streptococcus sp. A12
Streptococcus sp. I-P16
Streptococcus sp. oral taxon 431
Streptococcus thermophilus
Veillonella parvula
Others

Fig. 2

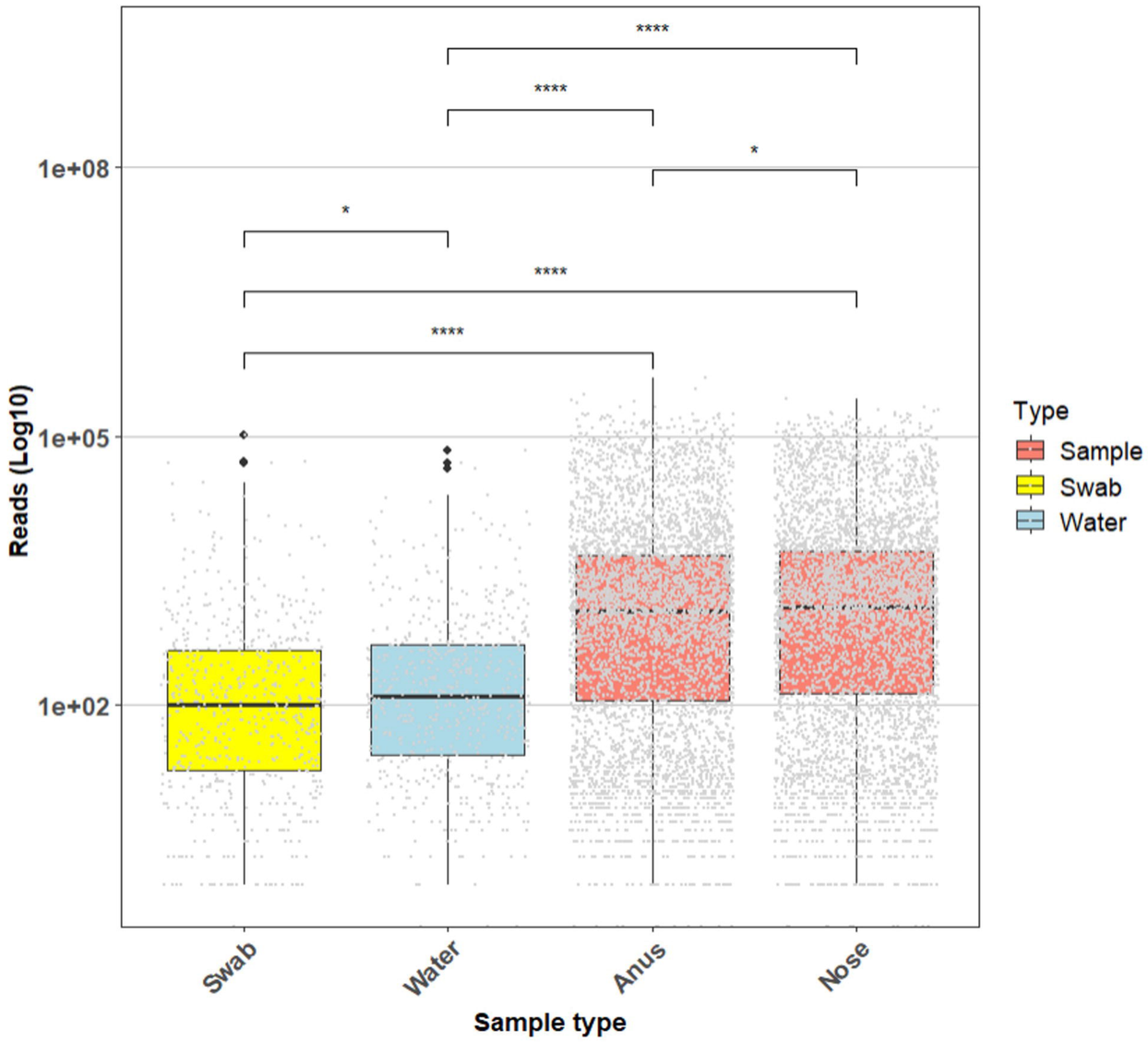
****
****
*
*
****
****
1e+08
1e+05
1e+02
Reads (Log10)
Swab
Water
Anus
Nose
Sample type
Type
Sample
Swab
Water

Fig. 3

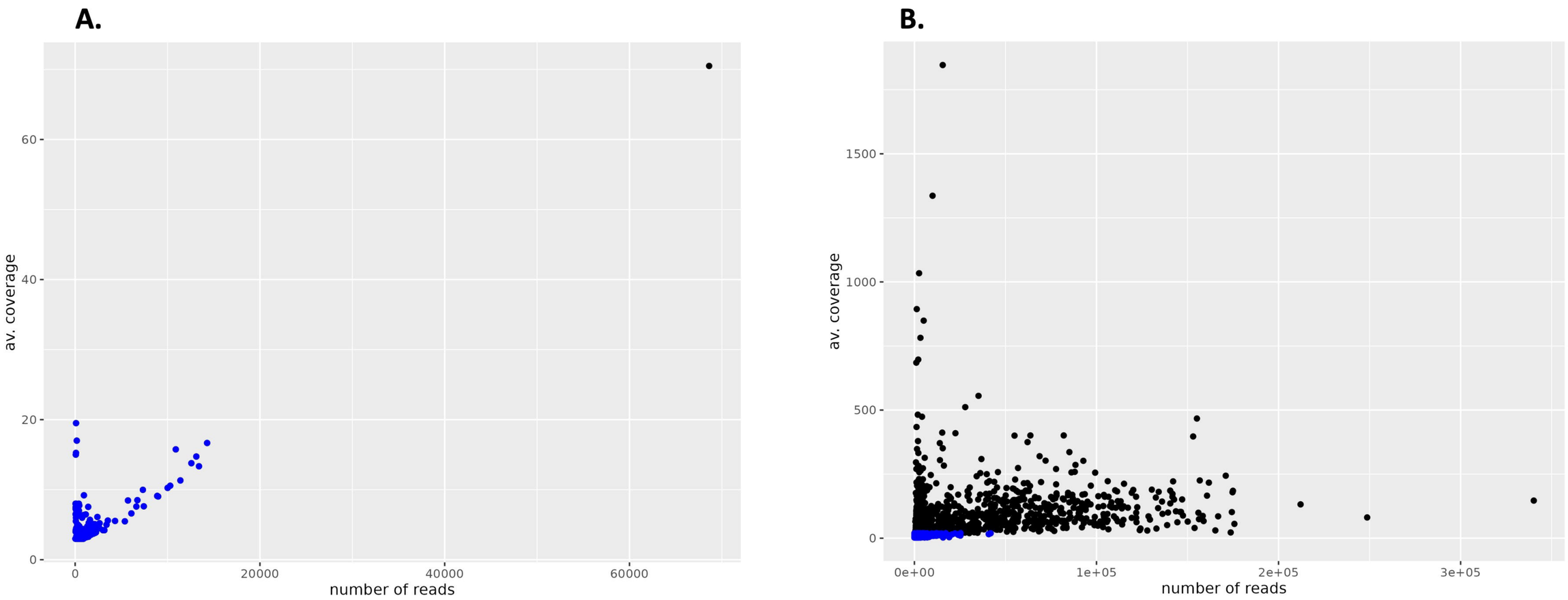
A.
av. coverage
0
20
40
60
0
20000
40000
60000
number of reads
B.
av. coverage
0
500
1000
1500
0e+00
1e+05
2e+05
3e+05
number of reads

Fig. 4

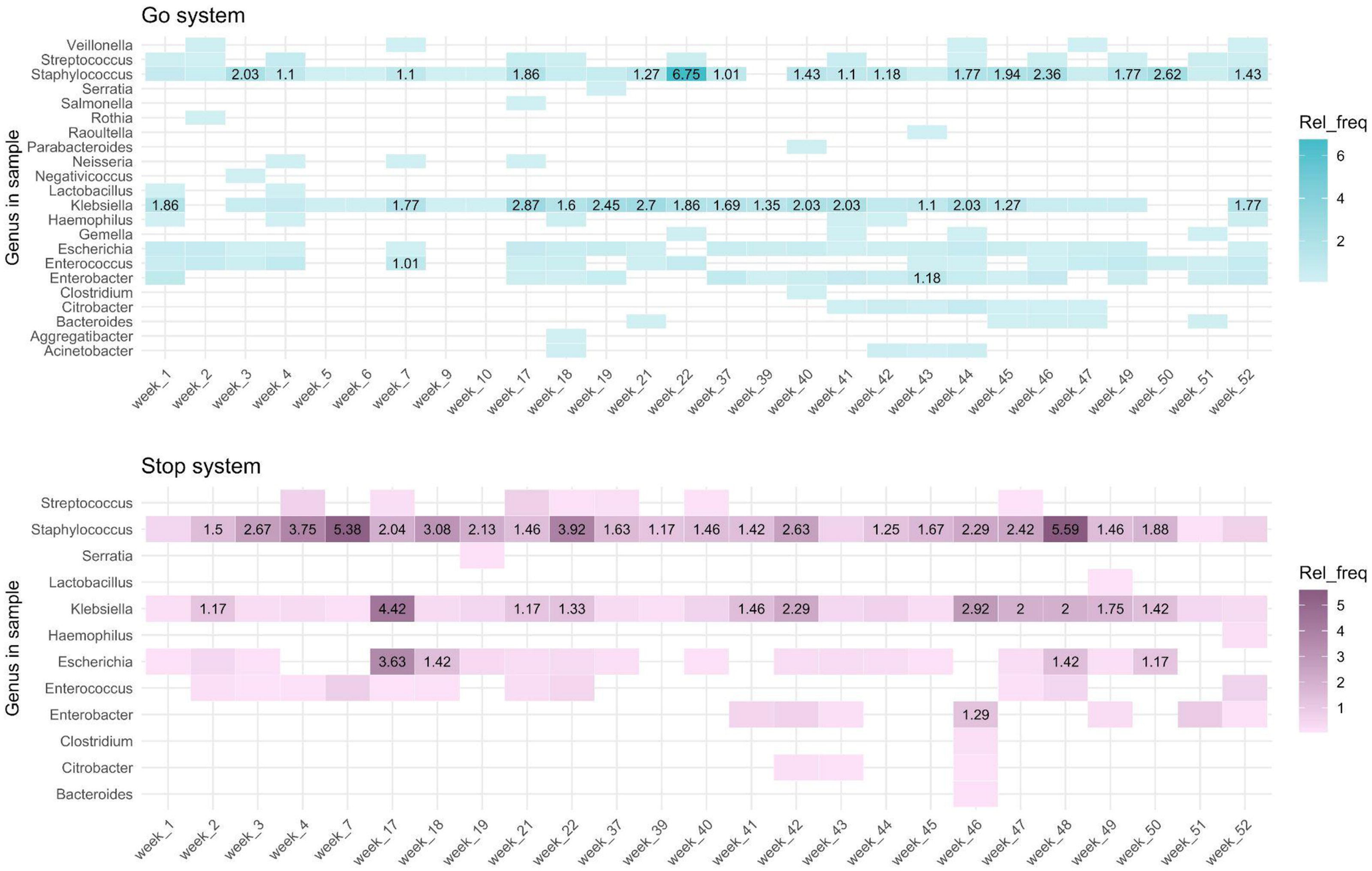
Go system
Genus in sample
Veillonella
Streptococcus
Staphylococcus
Serratia
Salmonella
Rothia
Raoultella
Parabacteroides
Neisseria
Negativicoccus
Lactobacillus
Klebsiella
Haemophilus
Gemella
Escherichia
Enterococcus
Enterobacter
Clostridium
Citrobacter
Bacteroides
Aggregatibacter
Acinetobacter
week_1 week_2 week_3 week_4 week_5 week_6 week_7 week_9 week_10 week_17 week_18 week_19 week_21 week_22 week_37 week_39 week_40 week_41 week_42 week_43 week_44 week_45 week_46 week_47 week_49 week_50 week_51 week_52
Rel_freq
6
4
2
Stop system
Genus in sample
Streptococcus
Staphylococcus
Serratia
Lactobacillus
Klebsiella
Haemophilus
Escherichia
Enterococcus
Enterobacter
Clostridium
Citrobacter
Bacteroides
week_1 week_2 week_3 week_4 week_7 week_17 week_18 week_19 week_21 week_22 week_37 week_39 week_40 week_41 week_42 week_43 week_44 week_45 week_46 week_47 week_48 week_49 week_50 week_51 week_52
Rel_freq
5
4
3
2
1

Fig. 5

Year	Week	Proportion human	Proportion microbial
2023	2	6,78	93,22
2023	3	12,20	87,80
2023	4	17,55	82,45
2023	5	11,81	88,19
2023	6	16,49	83,51
2023	7	5,58	94,42
2023	8	5,55	94,45
2023	9	18,01	81,99
2023	10	11,62	88,38
2023	11	7,66	92,34
2023	12	2,00	98,00
2023	13	2,92	97,08
2023	14	6,87	93,13
2023	15	4,70	95,30
2023	16	12,54	87,46
2023	17	21,66	78,34
2023	18	24,78	75,22
2023	19	13,74	86,26
2023	20	9,83	90,17
2023	21	13,78	86,22
2023	22	15,00	85,00
2023	23	9,94	90,06
2023	24	11,85	88,15
2023	25	10,57	89,43
2023	26	17,70	82,30
2023	27	22,88	77,12
2023	28	11,30	88,70
2023	29	10,23	89,77
2023	30	12,34	87,66
2023	31	9,85	90,15
2023	32	24,68	75,32
2023	33	7,72	92,28

2023	34	8,00	92,00
2023	35	21,76	78,24
2023	36	19,64	80,37
2023	37	20,20	79,80
2023	38	6,92	93,08
2023	39	20,87	79,13
2023	40	13,96	86,04
2023	41	24,22	75,78
2023	42	11,64	88,36
2023	43	6,58	93,42
2023	44	19,10	80,90
2023	45	13,49	86,51
2023	46	16,51	83,49
2023	47	11,58	88,42
2023	48	11,22	88,78
2023	49	11,03	88,97
2023	50	13,75	86,25
2023	51	20,54	79,46
2023	52	9,84	90,16
2024	1_y2	14,14	85,86
2024	2_y2	25,20	74,80
2024	3_y2	20,36	79,64
2024	5to10	4,73	95,27