Mitigation and detection of putative microbial contaminant reads from long-read metagenomic datasets

Stefany Ayala-Montaño[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 of clinical samples has significantly enhanced our understanding of microbial communities. However, microbial contamination and host-derived DNA remain a major obstacle to accurate data interpretation. Here, we present a methodology called 'Stop-Check-Go' for detecting and mitigating contaminants in metagenomic datasets obtained from neonatal patient samples (nasal and rectal swabs). This method incorporates laboratory and bioinformatics work combining a prevalence method, coverage estimation, and microbiological reports. We compared the 'Stop-Check-Go' decontamination system with other published decontamination tools, and commonly found poor performance in decontaminating microbiologically negative patients (false positives). We emphasize that host DNA decreased by an average of 76% per sample using the lysis method, and was further removed during post-sequencing analysis, and microbial putative contaminants were classified to 'Stop' in nearly 60% of the dataset. The 'Stop-Check-Go' system was developed to address the specific need of decontaminating low-biomass samples, where existing tools—primarily designed for short-read metagenomic data—showed limited performance.

Introduction

Metagenomics has established a framework for better understanding the composition and interactions within microbial communities, shedding light on clinical applications, metabolic functional interactions, and the long-term role of microbes in human health and disease. Traditionally, microbiome research relied on 16S rRNA gene sequencing, but this approach has limitations in resolving lower-order taxonomic classifications, and additional discriminatory power is required [1]. 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, such as those involving neonatal samples or other clinical specimens with limited microbial content, where external and internal contaminants can skew results and significantly impact the accuracy and reliability of data interpretation [2]. Contaminants can originate from external sources such as laboratory surfaces, reagents, sample handling, and instruments/equipment, or internal sources, including cross-contamination, arising during sample processing or index switching/hopping in sequencing [3] [4]. 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 [5].

A major limitation in the field is the frequent omission of negative and technical controls. Studies have shown that nearly 70% of the metagenomic studies do not include such controls, and of those that do include them, only a fraction sequence and analyze them [6]. 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’ [7] and their acknowledgment is crucial for interpreting results and making informed decisions in microbiome research [8]. Contamination is challenging to identify as it involves microorganisms also present in the samples, and regardless of the mitigation techniques such as UV irradiation, purification, or the separation of working areas, complete elimination is rarely achievable [6] [3] [8] [9].

To address this issue, several tools have been developed for bacterial contaminant identification, including Decontam [6], SourceTracker [10], SCRuB [11], and MicrobIEM [12]. These tools employ different methodologies to differentiate true signals from contaminants, relying on sequencing data from negative or environmental controls—[8]. Decontam 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) [13]. SourceTracker and SCRuB employ a Bayesian model to estimate the proportion of *n* sequences to a defined sink sample *x* by using the Gibbs distribution [10] and forms part of the open-access QIIME tools for metagenome and 16S rRNA gene amplicons [13]. MicrobIEM incorporates the prevalence method of Decontam and provides options based on the ratio of the species mean relative abundance that requires environmental sampling [12].

In this study, we investigated contamination in metagenomic samples obtained from low-biomass screening samples collected weekly from premature and term neonates in the neonatal intensive care unit (NICU) at the Medical Centre – University of Freiburg. These samples were analyzed using ONT long-read sequencing and presented 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 [14] [15]. To address these challenges, we developed a system called ‘Stop-Check-Go’ by adopting the prevalence method in Decontam and incorporating a novel metric—coverage per species per sample. 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 [13] [16].

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, to understand microbial transmission and microbiome dynamics in neonatal health [15] [7].

Methods

Collection of samples, microbial growth and preservation

Rectal and nasal swab samples were routinely collected from neonates admitted to the Neonatal Intensive Care Unit (NICU) within the framework of a study investigating 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 in Brain Heart Infusion (BHI) broth at a 1:2 ratio of sample to media at 37°C, 210 rpm. Microbial growth was validated by culturing on MacConkey agar, CHROMagar™/CNA agar (biplate), and blood agar, incubated at 37°C for 24 hours. Positive-growth samples were reported and the culturable bacteria on MacConkey agar were associated with Enterobacterales and therefore identified using MALDI-TOF. The bacteria of interest were purified, and both sweep and pure cultures were cryopreserved in horse blood at -80°C. These were available in cases where confirming true positives and ruling out cross-contamination was needed.

Removal of Human DNA and DNA Extraction

Human DNA (hDNA) was removed from enriched broth samples by centrifuging at 10.000 rpm for 5 minutes, and the microbial pellet was incubated with 1 mL of molecular grade water at room temperature (RT) for 5 minutes, promoting cellular lysis [17]. 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 (Catalog no. 11796828001).

Library Preparation and Metagenomic Sequencing

On average, we collected 24 samples per week, corresponding to 12 patients with two sets of samples, which were processed and distributed on two identical flow cells. 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 (from the beginning of the project until week 13). The SQK-NBD114-96 barcoding kit for R10 chemistry was used from week 14, following the manufacturer's protocol (available online at https://nanoporetech.com). In the R10 chemistry, we used an initial volume of 11 µL per sample in the library protocol to get the maximum possible concentration (low biomass samples ranged from too low to 3.5 ng/µL 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 Dorado basecall server (v7.2.13). From week 17, Readfish (v23.1.1) was run simultaneously with the start of the sequencing to actively target and 'reject' sequencing reads corresponding to hDNA (GRCh38, T2T-CHM13v2) across individual nanopores [18]. Weeks with a final library concentration below 7.0 ng/µL and negative growth in the controls were repeated with R10 chemistry flow cells. The 11µL loading volume per sample was chosen for a good initial concentration and to minimize the risk of contamination based on the prevalence method (refer to results section: Implementation of the Stop-Check-Go system).

Read Pre-processing and Removal of Human DNA (hDNA) 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, T2T-CHM13v2) using Minimap2 (v2.24; using the 'map-ont' parameter) [19] and hDNA reads were removed using SAMtools (v1.14) [20]. These steps, including read trimming, read filtration, and hDNA removal, were executed using custom Nextflow pipelines [21] (see Table S1. for the codes implemented in Nextflow pipelines).

Microbial Species Abundance Estimation of Real-world Metagenomic Datasets

Metagenomic composition reports were generated using Kraken v1.1.1 [22] with the MiniKraken database (DB_8GB), and the microbial abundance was estimated from this output using Bracken [23]. The taxonomic profiles per sample per week were visualized in R, using Tidyverse (v2.0.0) [24], scales [25], and ggthemes [26] to validate the microbiological reports. Those controls associated with positive growth were purified, confirmed by MALDI-TOF, sequenced with the Nextera DNA Flex Library Preparation Kit (Illumina, MiSeq), and the species were identified to the sequence type level using MLST [27] [28]. We labeled the contaminants in the Kraken metagenomic reports and excluded them from downstream analyses if the same sequence type for the corresponding week was identified with TRACS [28].

Stop-Check-Go system development and identification of non-hDNA Contaminants

The Stop-Check-Go system was designed to include two main quantitative parameters and one qualitative parameter: Difference Fraction Reads (DFR), Coverage, and the presence of microbial contaminants in negative controls. Individual taxonomic reports were combined and used to calculate the 1) Difference of the Fraction Reads (DFR) per species per sample based on the prevalence method that included the reads fraction of the control (FRC) of species x, and reads fraction in the sample (FRS), $DFR = FRC - FRS$. The 2) Coverage values were obtained from the assembly statistics text file from FLYE [29] by the '--stats' parameter for *n* contigs per sample file. We used the 'Kraken-translate' script [22] to obtain the taxonomic association within each contig and therefore could associate the 'contig_n' to a species. Binary microbiological growth results per sample (positive, negative), including controls, were added to the extensive combined Kraken file, which incorporated the three parameters to classify and identify potential contaminants.

Comparison of the Stop-Check-Go with other decontamination tools

The ‘Stop-Check-Go’ system was compared across five weeks with decontamination tools Decontam [5], and microBIEM [10]. Each method generated a list of putative contaminants ‘tax_id’ that was used to decontaminate the dataset, where each ‘tax_id’ was located in the respective tables containing the observation counts for each taxonomic unit (otu_table, phyloseq object). A new otu_table was generated from all decontamination tools, where zero was assigned to every putative contaminant otu (in the corresponding positions). These decontamination outputs were normalized using the Normalization.R script (author: Ilias Lagkouvardos), and Jensen-Shannon Diversity were calculated to quantify the communities’ species distribution. Non-parametric Spearman correlation and Mantel statistics were calculated for each paired combination of the decontamination tools.

Data Availability

Sequences have been deposited in the ENA in project PRJEB82667.

Results

Collection Overview

A total of 1,351 samples were collected from 302 patients, where 182 (60%) stayed one week on the ward, 53 (18%) stayed two, 20 (7%) three, and 47 (15%) more than three weeks. Overall, 58% of the patients (n=177) were colonized by Enterobacterales, and 25 patients were found to be microbiologically negative and were discharged before a second screening. Two negative controls from sterile swabs and molecular grade water were processed weekly, bringing the total to 1475 samples analyzed.

Human DNA removal

We evaluated the presence of human DNA (hDNA) contaminants in our samples as hDNA usually misleads the interpretation of the results [30] [31]. We have applied three different approaches for the removal. First, a lysis protocol [28] in the samples as hDNA represented in some cases 90% of the reads. The lysis helped to achieve on average a reduction of 67% of hDNA reads, which combined with the enrichment step reduced up to 96%. Second, during sequencing, the Readfisth Python package (Table S1, A.) selectively ejected the reads from the flow cell pores matching the human DNA

reference (GRCh38). We quantified that the content of hDNA in the samples was higher in the nose than the anus, and the proportion of matching hDNA reads across the weeks is provided in Table S2 (column C-E). Third, in post-sequencing analyses, after Porechop and Filtlong (Table S1, B.), we filtered the reads that mapped against the reference GRCh38 with minimap2. The remnant hDNA and microbial DNA reads are presented in Fig. S1, from which we only retained the microbial DNA for downstream analyses using SAMtools (Table S2, column F, H) [20].

genomic DNA was extracted from these strains using the High Pure PCR Template Preparation Kit (Roche) according to manufacturer's protocol. DNA libraries were then prepared with the Nextera DNA Flex Library Preparation Kit (Illumina). Pairedend sequencing (2×150 reads, 300 cycle v2) was conducted on an Illumina MiSeq platform (Illumina).

The residual number of reads in sample types might be associated with contamination

We sequenced 25 negative patients and 105 negative-growth controls (85%, n total= 124), and regardless we reported negative growth on agar plates, we quantified reads associated with the introduction of microbial contaminants. We found that most of the species in the samples had less than 1,000 (1K) reads, however, certain species were contributing with a large number of reads (> 1K). *Staphylococcus epidermidis*, *S. aureus*, *Klebsiella pneumoniae, Escherichia coli* contributed with reads between 1K – 10K, and *S. epidermidis* exhibited reads between 10K – 100K. We checked the weeks with reads in negative samples and found a homogeneous distribution across the dataset, although the biggest effect was attributable to weeks 22, 23, 24 and 24, 25, 40 (nose and anus, respectively). This elevated number of reads in the negative samples and controls led us to compare the reads abundance among the sample types. We found significant differences (p-adj.) in the controls (swab, water) versus the samples (anus, nose) (Fig. 1). We further separated the sample types by positive or negative growth and showed that regardless of the high number of reads in the negative samples, the number in the positive samples is markedly different with statistical significance (p-adj.).

Contaminant detection and source attribution

The high number of reads found in the negative-growth samples justified the need to distinguish the contamination sources, and to develop a strategy for dealing with potential microbial contaminants. We identified three types of contaminants: (1) cross-contamination either before or during the enrichment step, (2) contaminants introduced during DNA extraction and library preparation, and (3) contaminants that appeared in low biomass samples at higher abundance, now termed *library contamination.* The contamination (1) could be confirmed by using microbiological reports as we observed growth on these agar plates, purified the bacteria and sequenced them with MiSeq (Illumina). We reported 15 weeks with cross-contamination (24%), but only 3 weeks (5%) with Enterobacterales in swab/water: week 12*, E. coli* ST8186, week 15, and 16, *K. oxytoca* ST 189. The contamination (2) and (3) remained challenging to classify and separate from the species with true biological meaning, so we quantified the potential contaminants using read numbers.

From the observation and quantification of negative controls, we expected that samples with low or near zero contamination have a mixed profile in the taxonomic composition, with low number of reads (< 1K, denoted as 'no star'). We also expected the species proportion to show variation in the library preparation, examples of expected profiles for the controls (regardless of the type: water or swab) are weeks 8 -10, 30-35, and 52-53 (Fig. 2). The weeks 17-19 and 21-25 are possible examples of contamination (2) and (3) due to dominance of a single species, in this case *E. coli,* and the associated negative-growth reports after culturing.

Library repetition improved the DNA concentration and taxonomic profiling

The weeks that did not follow the expected mixed pattern were further investigated, and we found that 70% of the libraries (n=7/10) had low library concentrations with low individual DNA extraction concentrations. In addition, we extracted the reads of the overrepresented *E. coli* in the controls with more than 10K reads (in most cases), however we were unable to assign a particular ST for any of these (*E. coli*). The combination of low DNA concentration, negative growth in the controls, and unavailable STs, led us to select seven weeks with this particular library contamination. The library preparation for these weeks was repeated (including the controls) with the

R10 chemistry, Table 2 summarizes the final library DNA concentration (before sequencing) of the originals and the repeats. We included five additional libraries with either low DNA concentration (week 7) or good concentration but presence of Enterobacterales (> 1K) and negative-growth (week 25, 38). Two additional libraries were included as negative controls (Enterobacterales < 1K, week 39, 60) for a total of 12 repeated libraries. There was an outstanding improvement in the library concentration in 75% of the repeated libraries, except for weeks 25, 38 and 39, which had good concentration in the original libraries (see Table 1). The taxonomic profiles of the original and the repeated libraries are displayed in Fig. 3A, where in the cases with *E. coli* overrepresentation, the controls changed back to mixed profiles. An exception was week 22, in which after the repetition we could observe cross-contamination with *Pseudomonas aeruginosa* in the swab, which was not associated with positive-growth during the culturing step.

We closed the study in week 56, but we followed the last patient until discharged as we reported colonization by *K. oxytoca* from week 54, however, due to the low number of samples (per week) we multiplexed all in a single library. This facilitated identification of the strong batch effect in the libraries that were sequenced together as displayed in Fig. 3B. In addition, the same effect was observed for week 60 (Fig. 3C) after being multiplexed with week 38, where the composition of the taxonomic profiles are nearly identical in both weeks.

Table 1. DNA concentration of the weeks in the first (original) and second (repeated) library preparation with R10 flow cells.

Week	DNA concentration original library	DNA concentration repeated library (R10)	DNA concentration improvement (%)
7	A: 22.66, B: 6.08	39.00	A: 172, B: 641
17	5.52	25.08	454
18	3.61	21.81	604
19	4.73	28.00	592
21	6.54	67.82	1037
22	5.02	A: 33, B: 6.0*	A: 657, B: 120
23	A: 2.15, B: 1.93	37.65	A: 1751, B: 1951
24	19.31	30.43	158
25**	56.6	14.15	-25
38**	63.01	30.83	-49
39**	72.0	57.6	-20

60	38.40	48.00	125

*Week 22 was split into two pools (A, and B). In pool B, 3/7 of the patients were microbiologically negative, which strongly diluted the final library concentration. **Regardless of concentration, the libraries were repeated due to the suspicion of library contamination in the controls. A, and B, referred to Pool A and Pool B respectively, which were loaded in different flow cells.

Classification of putative microbial contaminants in the 'Stop-Check-Go' system

The contaminants during DNA extraction, library preparation and sequencing that are difficult to identify and the observation of a high number of reads in the samples we presumed as negative, encouraged us to develop a classification system. This system called 'Stop-Check-Go' included microbiological reports, coverage, and the weekly negative controls for validating the species presence with a method already published [6]. The species validation occurred by a direct comparison of the species in the samples with the species in the controls (by week) using the prevalence method proposed by Davis et. al 2018 [6]. The authors established that the total sample DNA (T) is composed of contaminant (C) and true sample DNA (S), so $T = C + S$. Thus, in negative controls where $S \sim 0$, the probability of contaminant detection is higher than the detection of contaminants in true samples when $S > 0$ [6]. This means that species present in the samples are classified as contaminants if species prevalence is higher in the controls: $Prevalence\ controls > Prevalence\ samples$. This follows the assumption that the species in the controls have competing DNA during the sequencing process, and therefore it is a good portrait of species contaminants for the corresponding week [6]. We checked this prevalence by calculating the difference of the fraction reads (DFR) per species per sample and the values were classified as 'Stop' if $DFR = FRC - FRS > 0$, and 'Go' when $DFR < 0$. Afterwards, we classified the species according to the coverage after assembly and error correction (Flye, Medaka, respectively) with a minimum (2) coverage of 20 as proposed by the literature [32]. Thus, the species were classified as '*low'* for coverage $< 20\mathrm{X}$, and '*ok*' for coverage of $20\mathrm{X}$ or above, and we complemented this with an auxiliary third parameter from microbiological reports. These reports were useful to classify the contamination reason of the controls, either cross-contamination (before DNA extraction) or library contamination (1K or 10K).

These three parameters were used to formulate the conditions of the classification system using the function 'case_when' from the dplyr R package, each condition was evaluated independently. Table 2 summarizes the system decision parameters and the classification into the corresponding category: 'Stop', 'Check', or 'Go'. Overall, the classification system strictly excluded species with a lower proportion of reads compared to the species in the negative controls $DFR > 0$ (conditions 1, 2), and microbiologically negative samples (4), as the reads might arise from previous steps where the contamination was present. Species with a high number of reads in the negative controls (library contamination 10K) were considered as putative contaminants if the coverage was low (condition 3). Species in the 'Check' category (conditions 5 - 8) were examined individually with the microbiological reports and only Enterobacterales found in pure culture were classified to 'Go'. Three weeks (5%) were contaminated with Enterobacterales: *K. oxytoca* ST186 (week 15, 16) and *E. coli* ST8186 (week 12). The other non-Enterobacterales contaminant species (*Staphylococcus* sp., and *Enterococcus* sp.) were classified to 'Stop'.

Table 2. Summary of the decision for the system stop check and go that includes the parameters DFR, coverage, and microbiological results

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

(10)	go	ok	library contamination 1K	go

Library contamination 10K* refers to species in controls where the number of reads was higher than 10.000. '-' denoted parameters not evaluated in the conditions.

Dynamic classification of species across samples

Using read counts is an approach frequently employed for classification thresholds of putative contaminants in metagenome and microbiome studies [4], [33], [34], but the use of coverage has the advantage of accounting for the genome size and read length. Consequently, species with a small genome size (e.g. *Staphyloccocus* sp.) can be falsely filtered when using read counts as they have a small number of reads but still good coverage. We explored the relationship between the coverage and the number of reads within the metagenomes by classifying the samples based on microbiological reports. Microbiologically positive samples (Fig. 4A), microbiologically negative (Fig. 4B), and controls (Fig. 4C), and also reported the proportion of filtered species solely based on a coverage below 20 (Fig. 4D). The differences in read quality can be observed by the broad distribution of the dots in Fig. 4A, and it is clear that even in negative samples, there are reads that can be misclassified based only on coverage (Fig. 4B). In the controls, we determined that all the samples, reported as microbiologically negative were under the threshold of 20. However, using only the coverage will create false positives in the samples where the controls had a high number of reads and good coverage (Fig. 4C), commonly observed in cross-contamination or library contamination 10K. On average, 50% of the species present in the metagenomes would be removed from the positive samples (Fig. 4D) without the inclusion of additional parameters, and approximately 20% of the microbiologically negative samples could not be filtered. The profiles and proportion of putative contaminants filtered using the defined coverage cutoff of 20 are presented in Fig. 4E, where the species profile changed weekly. However, a common backbone of *Staphylococcus* sp., *Streptococcus* sp., and *Veillonella* sp., had the highest removal rate independently of the week.

When we applied the classification system 'Stop-Check-Go' to the dataset, the species that fell into conditions (1), (2), (3), (4) from Table 3 were classified to 'Stop' and species into conditions (9), (10) to 'Go'. The overview of genus relative frequency per

week is presented in Fig. 5, where 20 species were classified simultaneously in both categories (57.14%, ntotal=35). *Staphylococcus* sp., *Enterococcus* sp., *Klebsiella* sp., and *E. coli* contributed with the highest rejection ('Stop') and inclusion rates ('Go') across the dataset, where *Staphylococcus* sp. and *Enterococcus* sp. did not follow a specific pattern. However, the average relative frequency of *E. coli* in the category to 'Stop' was higher in the weeks with suspected library contamination compared to the rest of the dataset (32.82%, 10.20%, respectively). Similarly, *K. pneumoniae* was observed with higher frequency in the second part of the study, but the effect on library contamination was underestimated as the dominance in the controls was not complete like *E. coli.* The Enterobacterales in 'Check' were compared with the microbiological reports to determine the proportion that could be confirmed to 'Go'. A total of 17.45% (204, ntotal=1169 samples in 'Go') were confirmed as positive by microbiology for *Klebsiella* sp. (*K. aerogenes*: 4, *K. oxytoca*: 22, *K. pneumoniae*: 97, *K. variicola*: 2) and *Escherichia coli* (n=79).

The 'Stop-Check-Go' system was compared with Decontam and MicroBIEM methods in the decontamination of 'tax_ids' per week, and the correlation between the (dis)similarity of the microbial communities (Fig. 6). A Pearson correlation below 0.3 indicating a high level of dissimilarity between the microbial community structures was further evaluated. We found a relative decontamination agreement of 66,67% and 75,56% between the 'Stop-Check-Go' (System) and Decontam, MicroBIEM, respectively. However, the 'tax_ids' that were not decontaminated by Decontam and MicroBIEM had low levels of agreement (14,29% and 4,17% respectively). The comparison of the Decontam_System, and MicroBIEM_System is outlined in Fig. S3 A, B, respectively. We identified that the major limitation of the methods was the decontamination of microbiologically negative patients (MNP). We quantified that MNP falsely not decontaminated by Decontam represented 22.45% of the dataset (n=11/49, Fig. S3 A). Likewise, MNP not decontaminated by MicroBIEM represented 58,33% of the samples (n=14/24, Fig. S3 B). The prevalence inspection of positive samples and negative controls by Decontam with a threshold of 0.1 (default) theoretically splits true from the false contaminants (Fig. S4 A). However, this inspection in our dataset (after removing MNP and weeks with suspected cross-contamination) showed limitations in differentiating true samples from contaminants even with stricter thresholds like 0.5 (Fig. S4 B).

Discussion

Contaminant species are always present in metagenomic studies, usually introduced from reagent kits, environmental sources, and cross-contamination when samples are routinely handled before sequencing. This study addresses the challenge of microbial contamination in long-read metagenomic datasets derived from neonatal nasal and rectal swabs. We developed a comprehensive workflow encompassing pre-sequencing, sequencing, and post-sequencing stages to detect putative contaminants exclude putative human and microbial contaminant DNA. We introduced the 'Stop-Check-Go' system that aims to mitigate the introduction of contaminants in downstream analyses using species prevalence, coverage, microbiological reports to refine the identification of microbial taxa with biological meaning. Therefore, we complement our metagenomic analyses with short-read sequencing (Illumina) to validate the presence of Enterobacterales (which we survey) and confirm with better precision the cross-contaminant species in the controls.

Addressing contamination has gained relevance due to the increase of metagenomic studies applications and the use of reproducible decontamination tools is an emerging need. Previous studies in host-associated metagenomes, such as human or mouse samples, have found human DNA contamination to be the most prevalent, for example, 72% of the metagenomes in one study had at least one potential human contamination [35]. To date, there are 51 different tools to remove host contamination [36] [37], and the microbial decontamination tools vary depending on the compositionality of the dataset, the sequencing technology, and the specific needs of the metagenomic studies. However, most decontamination tools are designed and tested in short-read metagenomic microbiomes [38] [39] [40] [6] [10], which pose an evident limitation to use in our dataset. Also, the identification of contaminants in low biomass samples adds a layer of difficulty, where the tools have been published
a limited detection when separating controls from true samples [41].

In our present study, we wish to emphasize the use of negative controls to mitigate putative contaminants through the detection of common shared species per sequencing run. Previous studies reported costs and time as limitations [41]. We calculated that using two negative controls represented in our study 0.78% (< 1%) of

the sequencing costs (calculation based on one flow cell, demultiplexed with 15 samples). In the absence of negative controls some decontamination tools use a list of common species reported as contaminants from extraction kits and environmental sources [9] [41] [42] [43], under the assumption that reads matching these external references are considered contaminants. We showed changes weekly in the taxonomic profiles relating to things like lot changes, hence, in agreement with Salter et al. [3], we advise against comparing with external controls or sources.

As sequencing technologies evolve, the ability to effectively manage contaminants will remain central to producing high-quality data and uncovering meaningful biological insights. The exclusion of sufficient negative controls to distinguish contaminants from real community species increases the risk of misinterpretation. Metagenomic studies frequently do not carry out microbiological growth controls [44], [45], [46], as their samples may include unculturable species, and hence, a coverage threshold can also be employed. We considered that the investigation of Enterobacterales and the enrichment step we performed were advantages compared with other studies that investigated the abundant fraction of non-culturable microbes.

The three parameters included in the 'Stop-Check-Go' classification system differentiate those patients that showed a significant number of bacterial reads as putative contaminants. We argue that using coverage as one of the classification parameters outperforms the usage of the read counts as the coverage adjusts for read length and genome size. The analysis of species in controls indicates that a proper threshold needs to be set to exclude most of the misclassified species when read counts are used as the classification parameter. Coverage in our dataset provided reliable classifications as 80% of the species in microbiologically negative samples were excluded, but it is well complemented by the prevalence method [6] and microbiological reports. Using the combination of these three parameters we could observe in our dataset that for the same week, species can be simultaneously classified as 'Go' and as putative contaminants to 'Stop' (Fig. 5). The decontamination tools we have tested – Decontam and MicroBIEM – generate a list of putative contaminant 'tax_ids' in the dataset that are removed globally. We highlight that the 'Stop-check-go' (System) provides a combination of 'tax_ids', and 'sample_ids', which makes the decontamination sample-specific. The decontamination agreement

between the System, Decontam, and MicroBIEM is on average 71%, represented in red colors in all the weeks (Fig. 6). The largest differences between the decontamination tools were attributed to Decomtam and MicroBIEM failing to exclude putative contaminants present in high abundance in microbiologically negative patients.

The decontamination is also influenced by the complexity of the metagenomic samples. We observed negative controls with a low number of reads and coverage above 10x, characteristic of low-complexity metagenomes where the species are over-represented (most likely library contaminants). In medium to high-complexity metagenomes, the species will require high-quality reads to end up in good coverages and lower rejection rates. We found the prevalence method robust enough when complemented in our study for the identification of bacterial contaminants. We compared the results of the prevalence method alone with the ratio method proposed by MicrobIEM [10] (data not shown), and the agreement percentages varied from 90.23% to 98.89% for all the categories in the 'Stop-Check-Go' system.

Decontamination efforts in metagenomic studies are particularly relevant when considering clinical applications, where time-sensitive decisions often need to be made. In many cases, traditional culture methods are performed, but their results become available later. In contrast, metagenomic sequencing can provide rapid insights into the microbial composition of samples. Therefore, minimizing contamination is essential to ensure the reliability of the sequencing data, as clinicians may need to rely on this information to guide early diagnosis and treatment decisions before culture results are available.

As contamination is not avoidable, mitigation strategies and resources should be employed. Whilst it is crucial to address microbial contaminants introduced during sample handling or from extraction and library preparation kits, an equally strict removal of hDNA contaminants is a critical step in ensuring the accuracy and reliability of metagenomic studies [35]. We employed strategies to mitigate contamination rather than discarding our data. The first strategy was to perform a lysis protocol before DNA extraction, and the use of Readfish from week 17 after we detected the introduction of contaminants in low biomass samples, reducing the sequenced reads matching the

human DNA. A second strategy was to cryopreserve positive Enterobacterales samples, and third, to sequence them with MiSeq (Illumina) to confirm the species and sequence type. Foruth, including weekly negative controls and applying robust computational tools helped to mitigate the rate of false positives and false negatives, especially in low-biomass samples [4].

We acknowledge some limitations in our study. We could not test the effect of additional measures for contamination mitigation as the final volume of the samples was less than 700 μL, and only processes after DNA extraction could be repeated. Also, the samples negative for Enterobacterales from microbiology reports were not cryopreserved, and if the reports from metagenomic sequencing disagreed, it was not possible to recover the sample. We assumed that the repeated libraries 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). Although this was false for two repeated weeks (25, 38, 39), from which we used the original library as the repetition did not increase the library concentration, and there was no significant improvement in the controls' taxonomic profile (Fig. 3).

Conclusions

The parameters described in this study included laboratory and bioinformatics work and did not model species contamination. We aimed to mitigate the introduction of contaminants and to keep species with biological significance that can be falsely excluded by the decontamination tools tested. The validation with microbiological growth reports, the prevalence method [6], and coverage from assemblies using FLYE and Kraken-translate were key when assessing contaminant species. We are aware that the metagenomes in this study did not represent the actual microbiome of the patients due to the enrichment step, but served to increase the microbial ratio in our low-biomass samples from neonates. We found the conceptualization of the 'Stop-Check-Go' system useful because we could individually examine the species classified as 'Check' in the context of tracking Enterobacterales. We are also aware that examining individual cases may involve a trade-off with the desired high-throughput processing. However, it is well recognized that metagenome analysis poses a

significant challenge due to the complexity of the samples where precision and biological relevance should take precedence over speed.

Acknowledgment

Special thanks to Leonardo Duarte dos Santos and Ifeoluwa J. Akintayo for their technical assistance and for helping me with the sample collection, microbiological reports, DNA extraction, and library preparation. This work was funded by the Federal Ministry of Education and Research (BMBF) under grant number 01KI2018. Without their financial support, this research would not have been possible.

Author Contributions

Conceptualization: S.R. Data curation: S.A.M., A.O.A. Data Analysis and Investigation: S.A.M., R. K., A.O.A. Project administration and supervision: S.R. Resources: S.R. Visualization: S.A.M. Writing – original draft: S.A.M., A.O.A., R.K. Review & editing: All authors.

Conflicts of interest

The authors have no conflicts of interest to declare.

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

Fig. 1. Average number of reads (log10) between the sample types: swab, water (controls), anus, nose (samples). Shading light indicates microbiologically negative samples. The Wilcoxon signed-rank test between all sample types (globally), the controls, and the samples (only) (p. adj. < 0.05).

Fig. 2. Taxonomic profile of chronologically-ordered controls (swab and water).. Species with less than 1,000 reads are denoted by no star. 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. Black represents other microorganisms.

Fig. 3. Taxonomic profiles of libraries with low concentration libraries (week 7 to week 24) where library contamination was suspected. Libraries with good concentrations (weeks 25, 38, 39) were used as controls **A.** Original and repeated libraries (swab and water). Species with less than 1,000 reads are denoted by no star. 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. Black represents other microorganisms. Samples termed as repeats correspond to library repetitions with R10 flow cell chemistry. **B.** Taxonomic profiles of the negative controls for the last patient in this study across six weeks. **C.** Batch effect in the repeated sequencing libraries of different patient samples.

Fig. 4. Average coverage per species in the metagenome versus the total number of reads. The blue dots indicate species with coverage below the 20 threshold. **A.** The coverage for microbiologically positive samples. **B.** Coverage for microbiologically negative samples. **C.** Controls. **D.** Fraction of filtered species based only on coverage. **E.** Taxonomic profile of the filtered species based only on coverage.

Fig. 5. Overview of genus classified as 'Go' and 'Stop' within the system, showing the relative frequency of bacterial genera in the collected samples. Darker regions represent higher relative abundances, while values below 0.3 are not labeled.

Fig. 6. Pearson correlation of the microbial ecology distribution in the samples from Jensen-Shannon Diversity index (JSD) between the decontamination methodologies 'Stop-Check-Go' (System), MicroBIEM and Decontam.

Fig. S1. Number of reads in human and microbial DNA chronologically orderered for the sources sampled: anus and nose.

Fig. S2. The most abundant species in the dataset classified as 'Stop' and 'Go' in the system and the relative abundance of the bacterial species. Values below 0.1 are not labeled.

Fig. S3. Tree diagram comparing the results for the paired decontamination tools tested. ND refers to 'not decontaminated', decontaminated (D), **A.** Decontam (deco), 'Stop-Check-Go' (sys) **B.** MicroBIEM (micro) 'Stop-Check-Go' (sys).

Fig. S4. Result of the prevalence method application (only) from Decontam **A.** PS dataset (published in https://benjjneb.github.io/decontam/vignettes/decontam_intro.html, section 7) with the default threshold (0.1). **B.** In our dataset, excluding microbiologically negative patients (MNP) and weeks with suspected cross-contamination (with a stricter threshold of 0.5)

Table S1. Code availability for the specified processes that were implemented in Nextflow pipelines [19].

	Bioinformatic process	**Source**
A.	Selective depletion of humand DNA (Readfish)	https://github.com/LooseLab/readfish?tab=readme-ov-file
B.	Read trimming and filtering	https://github.com/ayoraind/ONT_adapter_removal_and_read_filtration
C.	Human DNA contaminant removal	https://github.com/ayoraind/hDNA_removal_and_mapping_stats
D.	Assembly	https://github.com/ayoraind/flye
E.	Taxonomic classification (i. Kraken, ii. Bracken)	i) Kraken: https://github.com/ayoraind/kraken ii) https://github.com/ayoraind/bracken
F.	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 refer to microbial reads.

References

[1] J. L. O'Callaghan, D. Willner, M. Buttini, F. Huygens and E. S. Pelzer, "Limitations of 16S rRNA Gene Sequencing to Characterize Lactobacillus Species in the Upper Genital Tract," *Frontiers in Cell and Developmental Biology,* vol. 9, no. 641921, 2021.

[2] L. Cornet and D. Baurain, "Contamination detection in genomic data: more is not enough," *Genome Biology,* vol. 23, no. 1, p. 60, 2022.

[3] S. J. Salter, M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P. Turner, J. Parkhill, N. J. Loman and A. W. Walker, "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses," *BMC biology,* vol. 12, pp. 1-12, 2014.

[4] J. J. Minich, J. G. Sanders, A. Amir, G. Humphrey, J. A. Gilbert and R. Knight, "Quantifying and understanding well-to-well contamination in microbiome research," *MSystems,* vol. 4, no. 4, p. 10, 2019.

[5] J. M. Martí, "Recentrifuge: Robust comparative analysis and contamination removal for metagenomics," *LoS computational biology,* vol. 15, no. 4, p. e1006967, 2019.

[6] 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, pp. 1-14, 2018.

[7] H. Jurasz, T. Pawłowski and K. Perlejewski, "Contamination issue in viral metagenomics: problems, solutions, and clinical perspectives," *Frontiers in Microbiology,* vol. 12, p. 745076, 2021.

[8] Y. C. Lou, J. Hoff, M. R. Olm, J. West-Roberts, S. Diamond, B. A. Firek, M. J. Morowitz and J. F. Banfield, "Using strain-resolved analysis to identify contamination in metagenomics data," *Microbiome,* vol. 11, no. 1, p. 36, 2023.

[9] 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, 2019.

[10] D. Knights, J. Kuczynski, E. S. Charlson, J. Zaneveld, M. C. Mozer, R. G. Collman, F. D. Bushman, R. Knight and S. T. Kelley, "Bayesian community-wide culture-independent microbial source tracking," *Nature methods,* vol. 8, no. 9, pp. 761-763, 2011.

[11] G. I. Austin, H. Park, Y. Meydan, D. Seeram, T. Sezin, Y. C. Lou, B. A. Firek, M. J. Morowitz, J. F. Banfield, A. M. Christiano, I. Pe'er, A. C. Uhlemann, L. Shenhav and T. Korem, "Contamination source modeling with SCRuB improves cancer phenotype prediction from microbiome data," *Nature Biotechnology,* vol. 41, no. 12, pp. 1820-1828, 2023.

[12] C. Hülpüsch, L. Rauer, T. Nussbaumer, V. Schwierzeck, M. Bhattacharyya, V. Erhart, C. Traidl-Hoffmann, M. Reiger and A. U. Neumann, "Benchmarking MicrobIEM–a user-friendly tool for decontamination of microbiome sequencing data," *BMC biology,* vol. 21, no. 1, p. 269, 2023.

[13] 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, 2024.

[14] 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, p. 600111, 2021.

[15] J. R. Bedarf, N. Beraza, H. Khazneh, E. Özkurt, D. Baker, V. Borger, U. Wüllner and F. Hildebrand, "Much ado about nothing? Off-target amplification can lead to false-positive bacterial brain microbiome detection in healthy and Parkinson's disease individuals," *Microbiome,* vol. 9, pp. 1-15, 2021.

[16] L. Zhang, F. Chen, Z. Zeng, M. Xu, F. Sun, L. Yang, Z. Bi, Y. Lin, Y. Gao, H. Hao, W. Yi, M. Li and Y. Xie, "Advances in Metagenomics and Its Application in Environmental Microorganisms," *Frontiers in microbiology,* vol. 12, p. 766364, 2021.

[17] M. Shehadul Islam, A. Aryasomayajula and P. R. Selvaganapathy, " A review on macroscale and microscale cell lysis methods," *Micromachines,* vol. 8, no. 3, p. 83, 2017.

[18] 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, 2021.

[19] H. Li, "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics,* vol. 34, no. 18, pp. 3094-3100, 2018.

[20] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup, "The sequence alignment/map format and SAMtools," *bioinformatics,* vol. 25, no. 16, pp. 2078-2079, 2009.

[21] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature biotechnology,* vol. 35, no. 4, pp. 316-319, 2017.

[22] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology,* vol. 15, pp. 1-12, 2014.

[23] J. Lu, F. P. Breitwieser, P. Thielen and S. L. Salzberg, "Bracken: estimating species abundance in metagenomics data," *PeerJ Computer Science,* vol. 3, p. e104, 2017.

[24] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. A. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seide and V. Spinu, "Welcome to the Tidyverse," *Journal of open source software,* vol. 4, no. 43, p. 1686, 2019.

[25] H. Wickham, T. Pedersen and D. Seidel, "scales: Scale Functions for Visualization. R package version 1.3.0," 2023. [Online]. Available: https://scales.r-lib.org. [Accessed 09 04 2025].

[26] Arnold and J., "ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 5.1.0.9000," 2024. [Online]. Available: https://jrnold.github.io/ggthemes/. [Accessed 09 04 2025].

[27] T. Seemann, "mlst Github," [Online]. Available: https://github.com/tseemann/mlst. [Accessed 31 11 2024].

[28] A. O. Afolayan, S. Ayala-Montaño, I. J. Akintayo, L. Duarte dos Santos and S. Reuter, "Assessing the performance of current strain resolution tools on long-read metagenomes," *bioRxiv,* Vols. 2024-11, 2024.

[29] M. Kolmogorov, D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Polevikov, T. P. L. Smith and P. A. Pevzner, "metaFlye: scalable long-read metagenome assembly using repeat graphs," *Nature methods,* vol. 17, no. 11, pp. 1103-1110, 2020.

[30] N. A. Bokulich and M. S. Robeson, "Bioinformatics challenges for profiling the microbiome in cancer: pitfalls and opportunities," *Trends in Microbiology,* 2024.

[31] A. Gihawi, Y. Ge, J. Lu, D. Puiu, A. Xu, C. S. Cooper, D. S. Brewer, M. Pertea and S. L. Salzberg, "Major data analysis errors invalidate cancer microbiome findings," *MBio,* vol. 14, no. 5, pp. e01607-23, 2023.

[32] R. Henry, C. Schang, S. Coutts, P. Kolotelo, T. Prosser, N. Crosbie, T. Grant, D. Cottam, P. O'Brien, A. Deletic and D. McCarthy, "Into the deep: evaluation of SourceTracker for assessment of faecal contamination of coastal waters," *Water research,* vol. 93, pp. 242-253, 2016.

[33] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman and N. Segata, "Shotgun metagenomics, from sampling to analysis," *Nature biotechnology,* vol. 35, no. 9, pp. 833-844, 2017.

[34] A. B. McIntyre, R. Ounit, E. Afshinnekoo, R. J. Prill, E. Hénaff, N. Alexander, S. S. Minot, D. Danko, J. Foox, S. Ahsanuddin, S. Tighe, N. A. Hasan, P. Subramanian, K. Moffat, S. Levy, S. Lonardi, N. Greenfield, R. Colwell and C. E. Mason, "omprehensive benchmarking and ensemble approaches for metagenomic classifiers," *Genome biology,* vol. 18, pp. 1-19, 2017.

[35] Y. Yue, H. Huang, Z. Qi, H. M. Dou, X. Y. Liu, T. F. Han, Y. Chen, X. J. Song, Y. H. Zhang and J. Tu, "Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets," *BMC bioinformatics,* vol. 21, pp. 1-15, 2020.

[36] M. Lataretu, S. Krautwurst, M. R. Huska, M. Marquet, A. Viehweger, S. D. Braun, C. Brandt and M. Hölzer, "Targeted decontamination of sequencing data with CLEAN," *bioRxiv,* pp. 2023-08, 2023.

[37] Y. Gao, H. Luo, H. Lyu, H. Yang, S. Yousuf, S. Huang and Y. X. Liu, "Benchmarking short-read metagenomics tools for removing host contamination," *GigaScience,* vol. 14, p. giaf004, 2024.

[38] B. Zou, J. Wang, Y. Ding, Z. Zhang, Y. Huang, X. Fang, K. C. Cheung, S. See and L. Zhang, "Deepurify: a multi-modal deep language model to remove contamination from metagenome-assembled genomes," *bioRxiv,* pp. 2023-09, 2023.

[39] R. Ounit and S. Lonardi, "Higher classification sensitivity of short metagenomic reads with CLARK-S," *Bioinformatics,* vol. 32, no. 24, pp. 3823-3825, 2016.

[40] D. T. McKnight, R. Huerlimann, D. S. Bower, L. Schwarzkopf, R. A. Alford and K. R. Zenger, "microDecon: a highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies," *Environmental DNA,* vol. 1, no. 1, pp. 14-25, 2019.

[41] Y. Liu, R. L. Elworth, M. D. Jochum, K. M. Aagaard and T. J. Treangen, "De novo identification of microbial contaminants in low microbial biomass microbiomes with Squeegee," *Nature Communications,* vol. 13, no. 1, p. 6799, 2022.

[42] R. Schmieder and R. Edwards, "Fast identification and removal of sequence contamination from genomic and metagenomic datasets," *PloS one,* vol. 6, no. 3, p. e17288, 2011.

[43] V. C. Piro and B. Y. Renard, "Contamination detection and microbiome exploration with GRIMER," *GigaScience,* vol. 12, p. giad017, 2023.

[44] K. K. Ko, K. R. Chng and N. Nagarajan, "Metagenomics-enabled microbial surveillance," *Nature Microbiology,* vol. 7, no. 4, pp. 486-496, 2022.

[45] L. Lange, G. Berg, T. Cernava, M. C. Champomier-Vergès, T. Charles, L. Cocolin, K. D'Hondt, T. Kostic, E. Maguin, T. Makhalanyane, A. Meisner, M. Ryan, G. S. Kiran, R. S. de Souza, Y. Sanz, M. Schloter, H. Smidt, S. Wakelin and A. Sessitsch, "Microbiome

ethics, guiding principles for microbiome research, use and knowledge management," *Environmental microbiome,* vol. 17, no. 1, p. 50, 2022.

[46] C. M. Cullen, K. K. Aneja, S. Beyhan, C. E. Cho, S. Woloszynek, M. Convertino, S. McCoy, Y. Zhang, M. Anderson, D. Alvarez-Ponce, E. Smirnova, L. Karstens, P. Dorrestein, H. Li, A. S. Gupta, K. Cheung, J. G. Powers, Z. Zhao and G. Rosen, "Emerging priorities for microbiome research," *Frontiers in microbiology,* vol. 11, p. 491374, 2020.

[47] K. A. Jolley and M. C. Maiden, "BIGSdb: scalable analysis of bacterial genome variation at the population level," *BMC bioinformatics,* vol. 11, pp. 1-11, 2010.

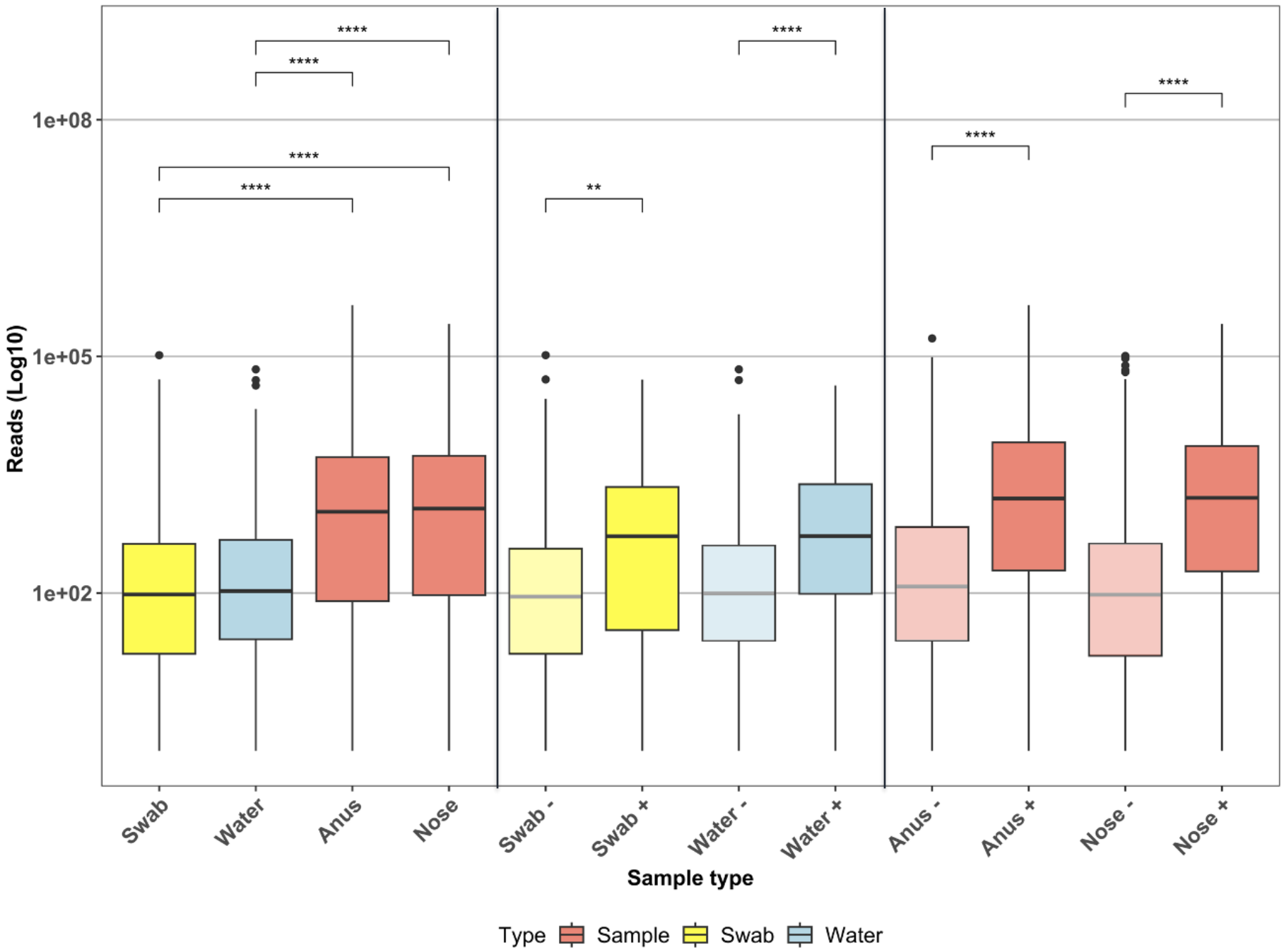

****
****
****
****
**
****
****
****
1e+08
1e+05
1e+02
Reads (Log10)
Swab
Water
Anus
Nose
Swab -
Swab +
Water -
Water +
Anus -
Anus +
Nose -
Nose +
Sample type
Type
Sample
Swab
Water

Fig. 1

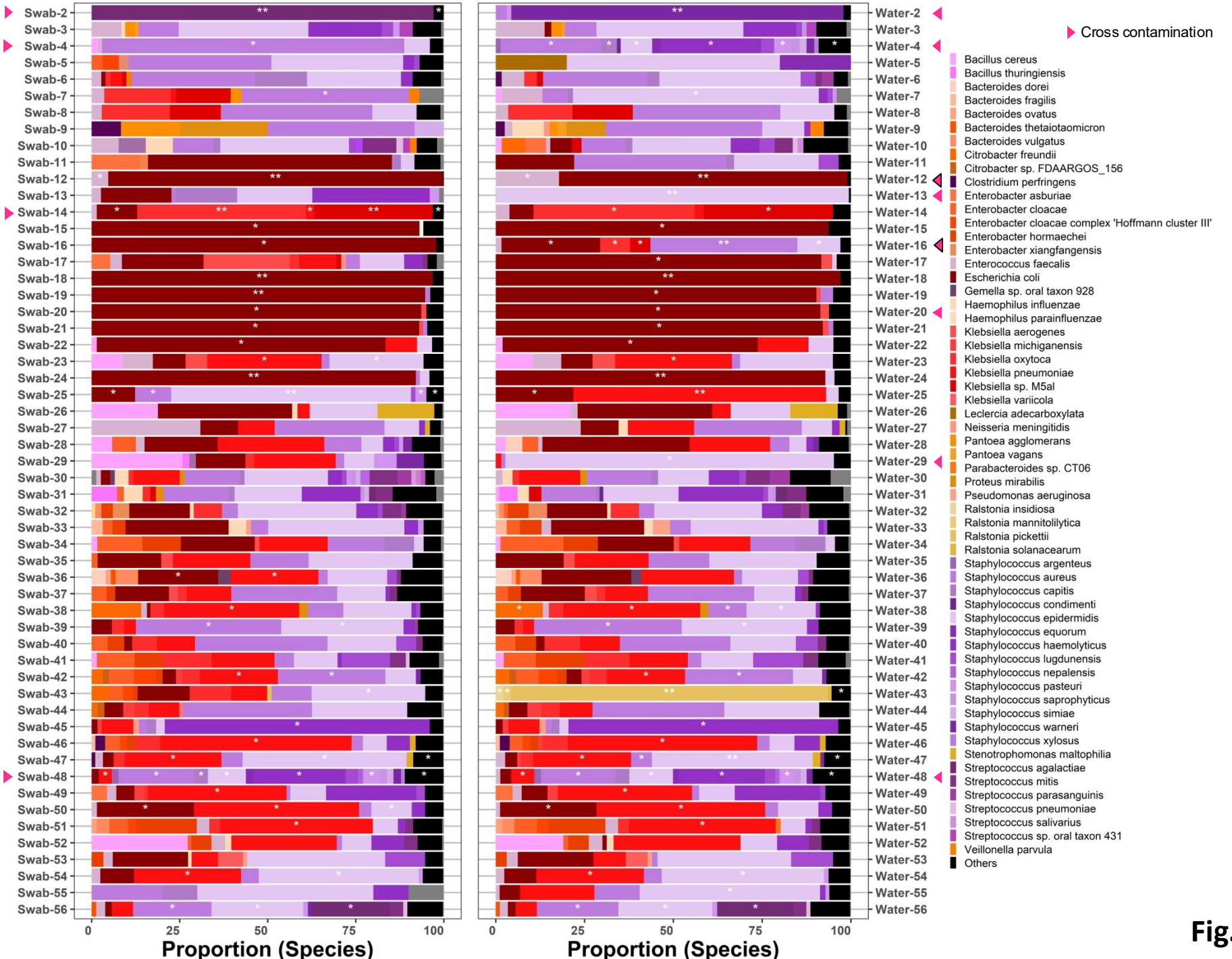

Swab-2
Swab-3
Swab-4
Swab-5
Swab-6
Swab-7
Swab-8
Swab-9
Swab-10
Swab-11
Swab-12
Swab-13
Swab-14
Swab-15
Swab-16
Swab-17
Swab-18
Swab-19
Swab-20
Swab-21
Swab-22
Swab-23
Swab-24
Swab-25
Swab-26
Swab-27
Swab-28
Swab-29
Swab-30
Swab-31
Swab-32
Swab-33
Swab-34
Swab-35
Swab-36
Swab-37
Swab-38
Swab-39
Swab-40
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
0
25
50
75
100
Proportion (Species)
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-27
Water-28
Water-29
Water-30
Water-31
Water-32
Water-33
Water-34
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
0
25
50
75
100
Proportion (Species)
Cross contamination
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 xiangfangensis
Enterococcus faecalis
Escherichia coli
Gemella sp. oral taxon 928
Haemophilus influenzae
Haemophilus parainfluenzae
Klebsiella aerogenes
Klebsiella michiganensis
Klebsiella oxytoca
Klebsiella pneumoniae
Klebsiella sp. M5al
Klebsiella variicola
Leclercia adecarboxylata
Neisseria meningitidis
Pantoea agglomerans
Pantoea vagans
Parabacteroides sp. CT06
Proteus mirabilis
Pseudomonas aeruginosa
Ralstonia insidiosa
Ralstonia mannitolilytica
Ralstonia pickettii
Ralstonia solanacearum
Staphylococcus argenteus
Staphylococcus aureus
Staphylococcus capitis
Staphylococcus condimenti
Staphylococcus epidermidis
Staphylococcus equorum
Staphylococcus haemolyticus
Staphylococcus lugdunensis
Staphylococcus nepalensis
Staphylococcus pasteuri
Staphylococcus saprophyticus
Staphylococcus simiae
Staphylococcus warneri
Staphylococcus xylosus
Stenotrophomonas maltophilia
Streptococcus agalactiae
Streptococcus mitis
Streptococcus parasanguinis
Streptococcus pneumoniae
Streptococcus salivarius
Streptococcus sp. oral taxon 431
Veillonella parvula
Others

Fig. 2

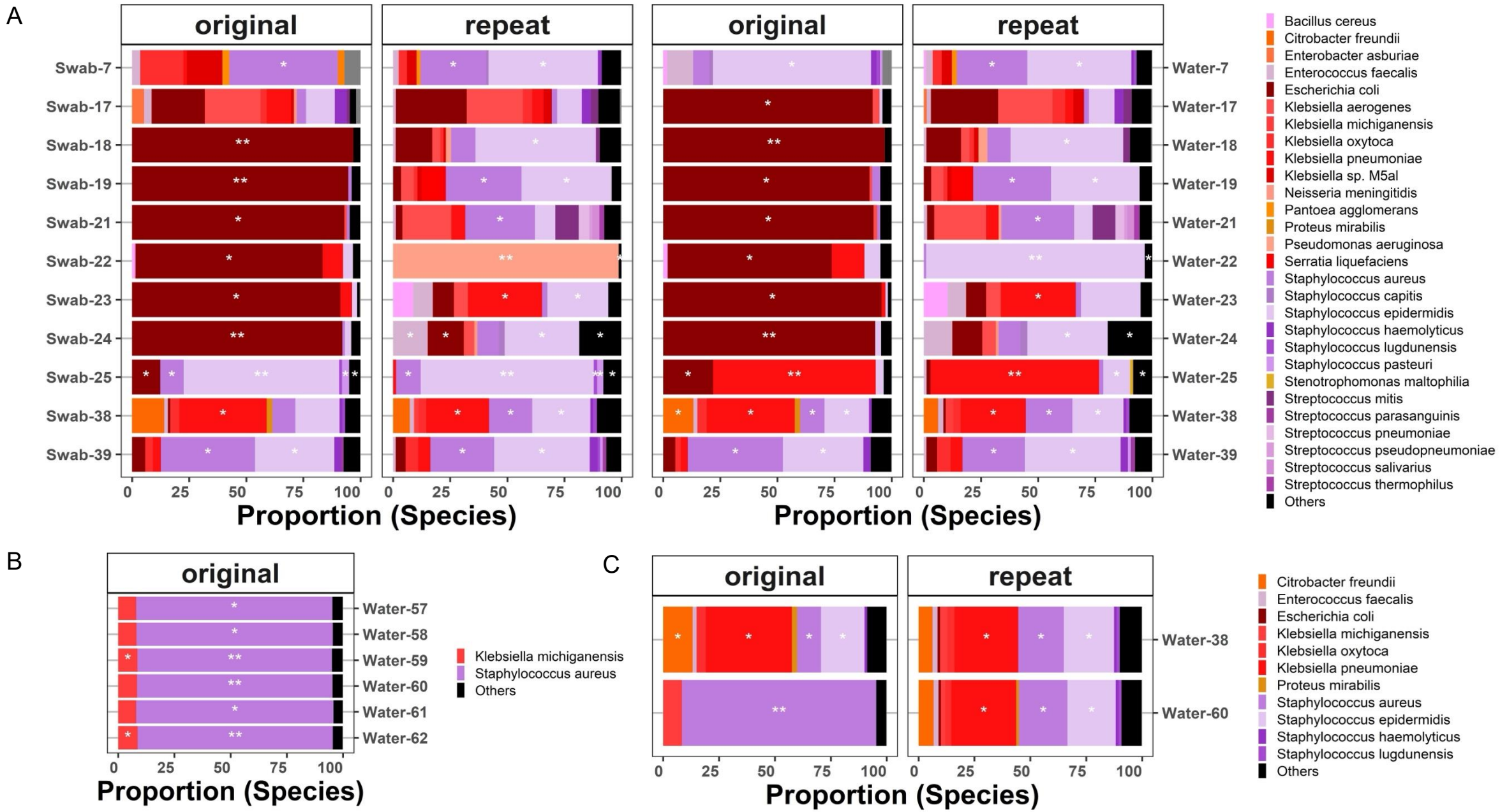

A
original
repeat
original
repeat
Swab-7
Swab-17
Swab-18
Swab-19
Swab-21
Swab-22
Swab-23
Swab-24
Swab-25
Swab-38
Swab-39
Water-7
Water-17
Water-18
Water-19
Water-21
Water-22
Water-23
Water-24
Water-25
Water-38
Water-39
0 25 50 75 100
Proportion (Species)
Bacillus cereus
Citrobacter freundii
Enterobacter asburiae
Enterococcus faecalis
Escherichia coli
Klebsiella aerogenes
Klebsiella michiganensis
Klebsiella oxytoca
Klebsiella pneumoniae
Klebsiella sp. M5al
Neisseria meningitidis
Pantoea agglomerans
Proteus mirabilis
Pseudomonas aeruginosa
Serratia liquefaciens
Staphylococcus aureus
Staphylococcus capitis
Staphylococcus epidermidis
Staphylococcus haemolyticus
Staphylococcus lugdunensis
Staphylococcus pasteuri
Stenotrophomonas maltophilia
Streptococcus mitis
Streptococcus parasanguinis
Streptococcus pneumoniae
Streptococcus pseudopneumoniae
Streptococcus salivarius
Streptococcus thermophilus
Others
B
original
Water-57
Water-58
Water-59
Water-60
Water-61
Water-62
Klebsiella michiganensis
Staphylococcus aureus
Others
0 25 50 75 100
Proportion (Species)
C
original
repeat
Water-38
Water-60
0 25 50 75 100
Proportion (Species)
Citrobacter freundii
Enterococcus faecalis
Escherichia coli
Klebsiella michiganensis
Klebsiella oxytoca
Klebsiella pneumoniae
Proteus mirabilis
Staphylococcus aureus
Staphylococcus epidermidis
Staphylococcus haemolyticus
Staphylococcus lugdunensis
Others

Fig. 3

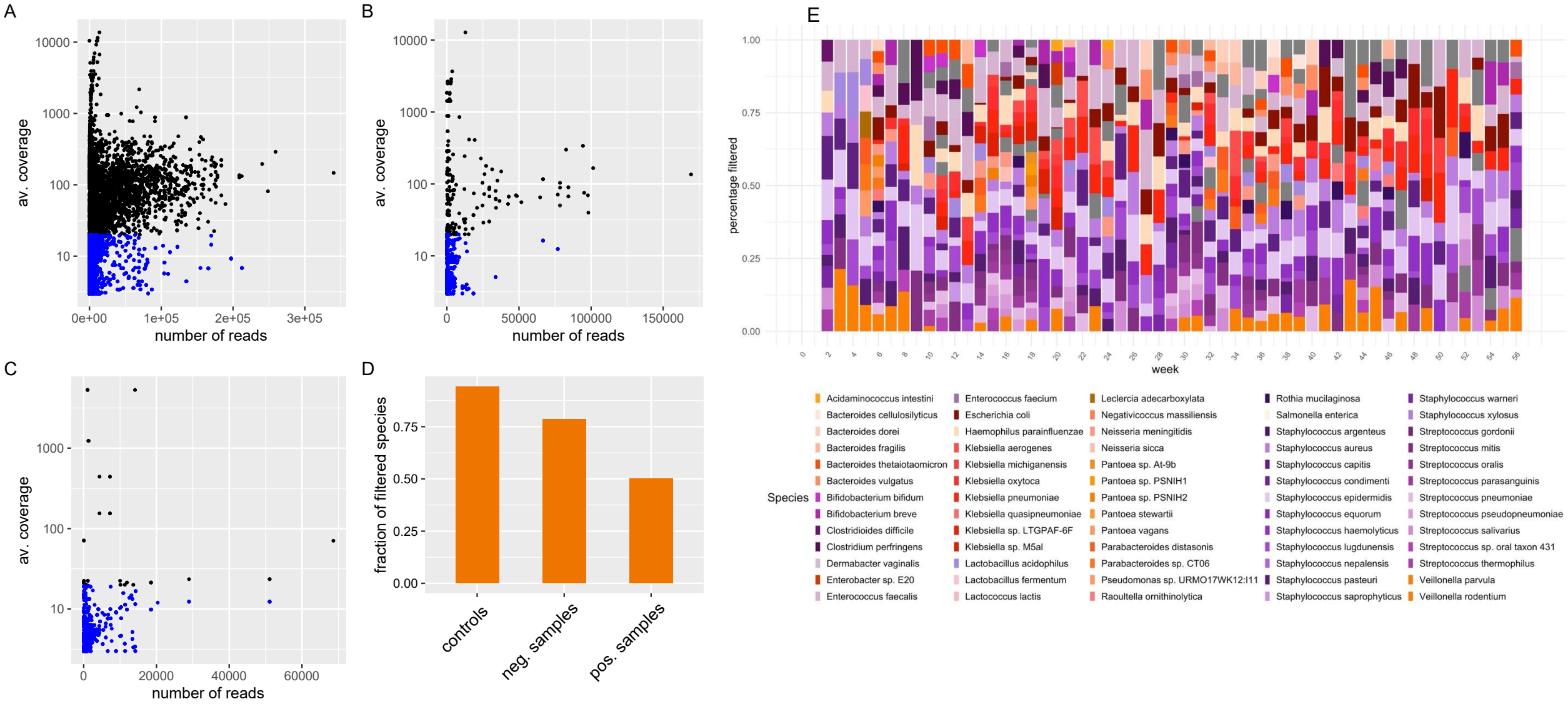

A
B
C
D
E
av. coverage
number of reads
fraction of filtered species
controls
neg. samples
pos. samples
percentage filtered
week
Species
Acidaminococcus intestini
Bacteroides cellulosilyticus
Bacteroides dorei
Bacteroides fragilis
Bacteroides thetaiotaomicron
Bacteroides vulgatus
Bifidobacterium bifidum
Bifidobacterium breve
Clostridioides difficile
Clostridium perfringens
Dermabacter vaginalis
Enterobacter sp. E20
Enterococcus faecalis
Enterococcus faecium
Escherichia coli
Haemophilus parainfluenzae
Klebsiella aerogenes
Klebsiella michiganensis
Klebsiella oxytoca
Klebsiella pneumoniae
Klebsiella quasipneumoniae
Klebsiella sp. LTGPAF-6F
Klebsiella sp. M5al
Lactobacillus acidophilus
Lactobacillus fermentum
Lactococcus lactis
Leclercia adecarboxylata
Negativicoccus massiliensis
Neisseria meningitidis
Neisseria sicca
Pantoea sp. At-9b
Pantoea sp. PSNIH1
Pantoea sp. PSNIH2
Pantoea stewartii
Pantoea vagans
Parabacteroides distasonis
Parabacteroides sp. CT06
Pseudomonas sp. URMO17WK12:I11
Raoultella ornithinolytica
Rothia mucilaginosa
Salmonella enterica
Staphylococcus argenteus
Staphylococcus aureus
Staphylococcus capitis
Staphylococcus condimenti
Staphylococcus epidermidis
Staphylococcus equorum
Staphylococcus haemolyticus
Staphylococcus lugdunensis
Staphylococcus nepalensis
Staphylococcus pasteuri
Staphylococcus saprophyticus
Staphylococcus warneri
Staphylococcus xylosus
Streptococcus gordonii
Streptococcus mitis
Streptococcus oralis
Streptococcus parasanguinis
Streptococcus pneumoniae
Streptococcus pseudopneumoniae
Streptococcus salivarius
Streptococcus sp. oral taxon 431
Streptococcus thermophilus
Veillonella parvula
Veillonella rodentium

Fig. 4

Stop & Go system
Genus
Yersinia-Stop
Veillonella-Stop
Veillonella-Go
Streptococcus-Stop
Streptococcus-Go
Stenotrophomonas-Stop
Staphylococcus-Stop
Staphylococcus-Go
Serratia-Stop
Serratia-Go
Salmonella-Stop
Salmonella-Go
Rothia-Stop
Rothia-Go
Raoultella-Go
Pseudomonas-Stop
Proteus-Stop
Proteus-Go
Parabacteroides-Stop
Parabacteroides-Go
Pantoea-Stop
Pantoea-Go
Neisseria-Stop
Neisseria-Go
Negativicoccus-Go
Moraxella-Stop
Micrococcus-Stop
Leclercia-Stop
Lactobacillus-Stop
Lactobacillus-Go
Klebsiella-Stop
Klebsiella-Go
Haemophilus-Stop
Haemophilus-Go
Gemella-Stop
Fusobacterium-Stop
Escherichia-Stop
Escherichia-Go
Enterococcus-Stop
Enterococcus-Go
Enterobacter-Stop
Enterobacter-Go
Corynebacterium-Stop
Clostridium-Stop
Clostridium-Go
Citrobacter-Stop
Citrobacter-Go
Bifidobacterium-Go
Bacteroides-Stop
Bacteroides-Go
Bacillus-Stop
Aggregatibacter-Go
Acinetobacter-Stop
Acinetobacter-Go
Acidaminococcus-Go
Week
Stop
25
50
75
100
Go
75
50
25

Fig. 5

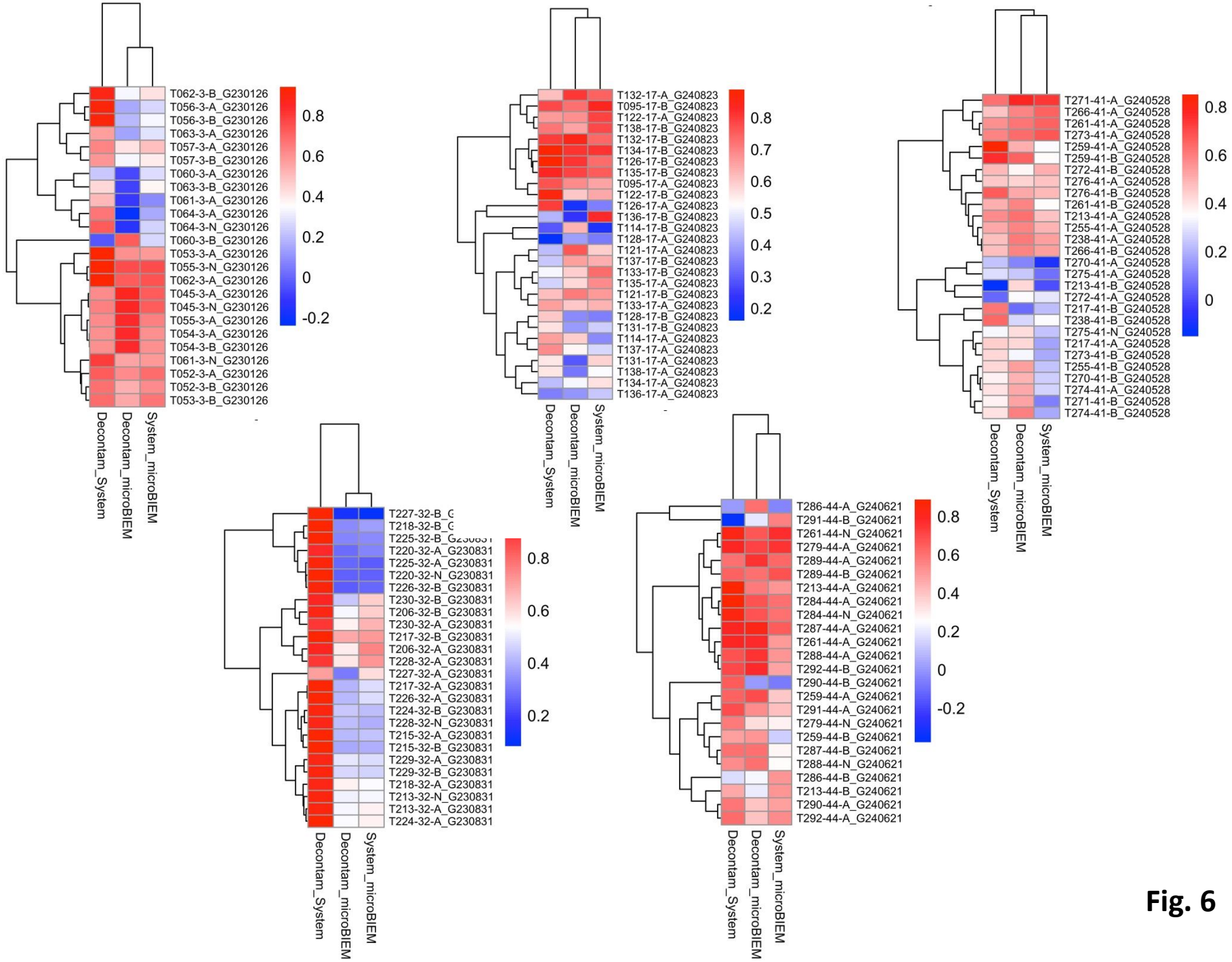

T062-3-B_G230126
T056-3-A_G230126
T056-3-B_G230126
T063-3-A_G230126
T057-3-A_G230126
T057-3-B_G230126
T060-3-A_G230126
T063-3-B_G230126
T061-3-A_G230126
T064-3-A_G230126
T064-3-N_G230126
T060-3-B_G230126
T053-3-A_G230126
T055-3-N_G230126
T062-3-A_G230126
T045-3-A_G230126
T045-3-N_G230126
T055-3-A_G230126
T054-3-A_G230126
T054-3-B_G230126
T061-3-N_G230126
T052-3-A_G230126
T052-3-B_G230126
T053-3-B_G230126
0.8
0.6
0.4
0.2
0
-0.2
Decontam_System
Decontam_microBIEM
System_microBIEM
T132-17-A_G240823
T095-17-B_G240823
T122-17-A_G240823
T138-17-B_G240823
T132-17-B_G240823
T134-17-B_G240823
T126-17-B_G240823
T135-17-B_G240823
T095-17-A_G240823
T122-17-B_G240823
T126-17-A_G240823
T136-17-B_G240823
T114-17-B_G240823
T128-17-A_G240823
T121-17-A_G240823
T137-17-B_G240823
T133-17-B_G240823
T135-17-A_G240823
T121-17-B_G240823
T133-17-A_G240823
T128-17-B_G240823
T131-17-B_G240823
T114-17-A_G240823
T137-17-A_G240823
T131-17-A_G240823
T138-17-A_G240823
T134-17-A_G240823
T136-17-A_G240823
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Decontam_System
Decontam_microBIEM
System_microBIEM
T271-41-A_G240528
T266-41-A_G240528
T261-41-A_G240528
T273-41-A_G240528
T259-41-A_G240528
T259-41-B_G240528
T272-41-B_G240528
T276-41-A_G240528
T276-41-B_G240528
T261-41-B_G240528
T213-41-A_G240528
T255-41-A_G240528
T238-41-A_G240528
T266-41-B_G240528
T270-41-A_G240528
T275-41-A_G240528
T213-41-B_G240528
T272-41-A_G240528
T217-41-B_G240528
T238-41-B_G240528
T275-41-N_G240528
T217-41-A_G240528
T273-41-B_G240528
T255-41-B_G240528
T270-41-B_G240528
T274-41-A_G240528
T271-41-B_G240528
T274-41-B_G240528
0.8
0.6
0.4
0.2
0
Decontam_System
Decontam_microBIEM
System_microBIEM
T227-32-B_G
T218-32-B_G
T225-32-B_G230831
T220-32-A_G230831
T225-32-A_G230831
T220-32-N_G230831
T226-32-B_G230831
T230-32-B_G230831
T206-32-B_G230831
T230-32-A_G230831
T217-32-B_G230831
T206-32-A_G230831
T228-32-A_G230831
T227-32-A_G230831
T217-32-A_G230831
T226-32-A_G230831
T224-32-B_G230831
T228-32-N_G230831
T215-32-A_G230831
T215-32-B_G230831
T229-32-A_G230831
T229-32-B_G230831
T218-32-A_G230831
T213-32-N_G230831
T213-32-A_G230831
T224-32-A_G230831
0.8
0.6
0.4
0.2
Decontam_System
Decontam_microBIEM
System_microBIEM
T286-44-A_G240621
T291-44-B_G240621
T261-44-N_G240621
T279-44-A_G240621
T289-44-A_G240621
T289-44-B_G240621
T213-44-A_G240621
T284-44-A_G240621
T284-44-N_G240621
T287-44-A_G240621
T261-44-A_G240621
T288-44-A_G240621
T292-44-B_G240621
T290-44-B_G240621
T259-44-A_G240621
T291-44-A_G240621
T279-44-N_G240621
T259-44-B_G240621
T287-44-B_G240621
T288-44-N_G240621
T286-44-B_G240621
T213-44-B_G240621
T290-44-A_G240621
T292-44-A_G240621
0.8
0.6
0.4
0.2
0
-0.2
Decontam_System
Decontam_microBIEM
System_microBIEM

Fig. 6

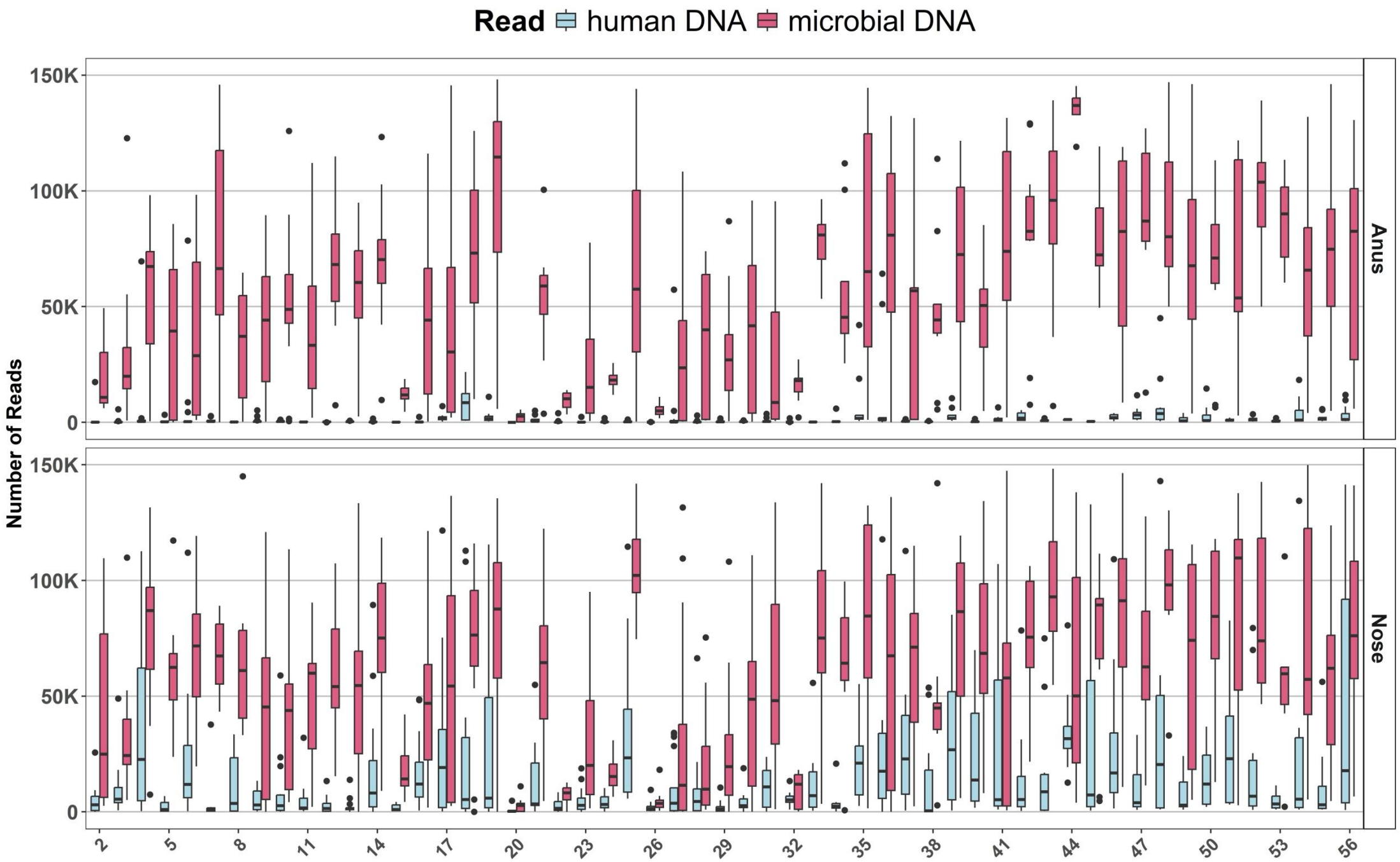

Read human DNA microbial DNA
Number of Reads
150K
100K
50K
0
Anus
Nose
2
5
8
11
14
17
20
23
26
29
32
35
38
41
44
47
50
53
56

Fig. S1

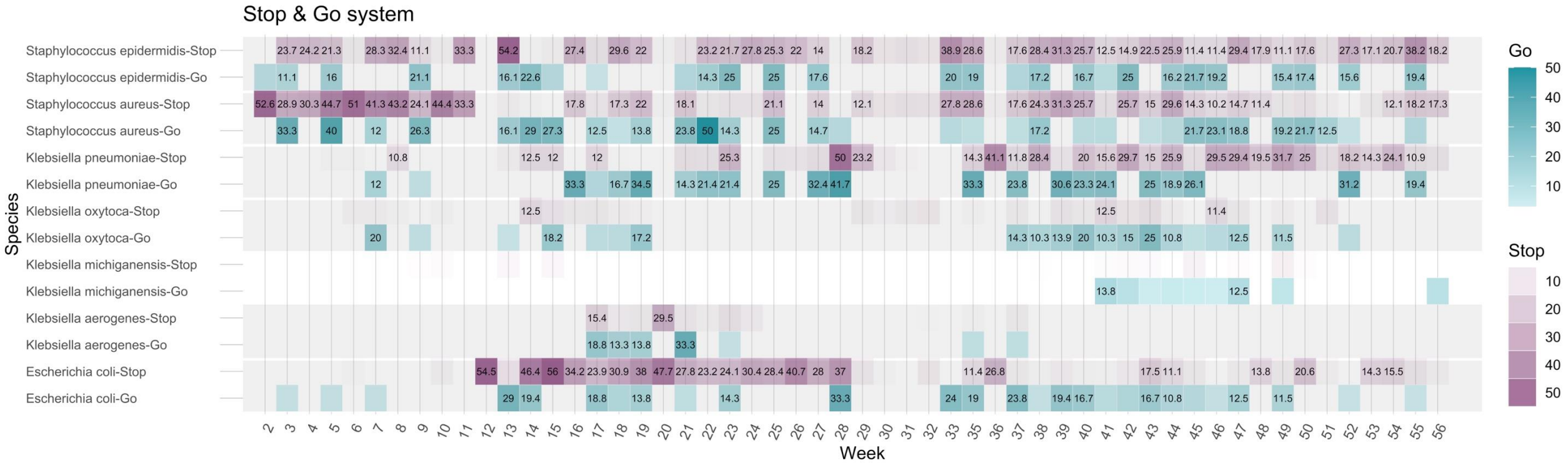

Stop & Go system
Species
Staphylococcus epidermidis-Stop
Staphylococcus epidermidis-Go
Staphylococcus aureus-Stop
Staphylococcus aureus-Go
Klebsiella pneumoniae-Stop
Klebsiella pneumoniae-Go
Klebsiella oxytoca-Stop
Klebsiella oxytoca-Go
Klebsiella michiganensis-Stop
Klebsiella michiganensis-Go
Klebsiella aerogenes-Stop
Klebsiella aerogenes-Go
Escherichia coli-Stop
Escherichia coli-Go
Week
Go
Stop

Fig. S2

A.

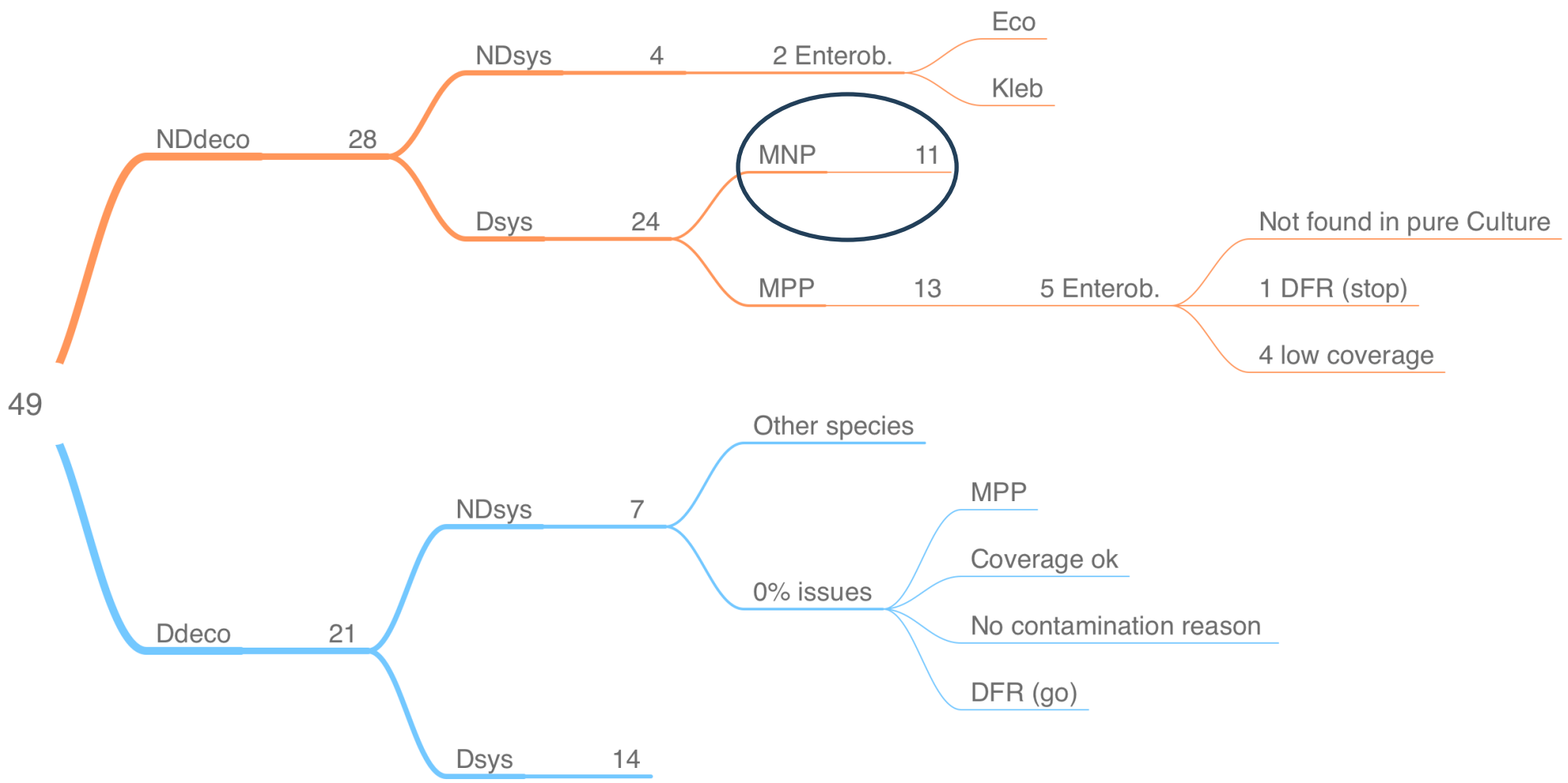

49
NDdeco
28
NDsys
4
2 Enterob.
Eco
Kleb
Dsys
24
MNP
11
MPP
13
5 Enterob.
Not found in pure Culture
1 DFR (stop)
4 low coverage
Ddeco
21
NDsys
7
Other species
0% issues
MPP
Coverage ok
No contamination reason
DFR (go)
Dsys
14

Fig. S3

B.

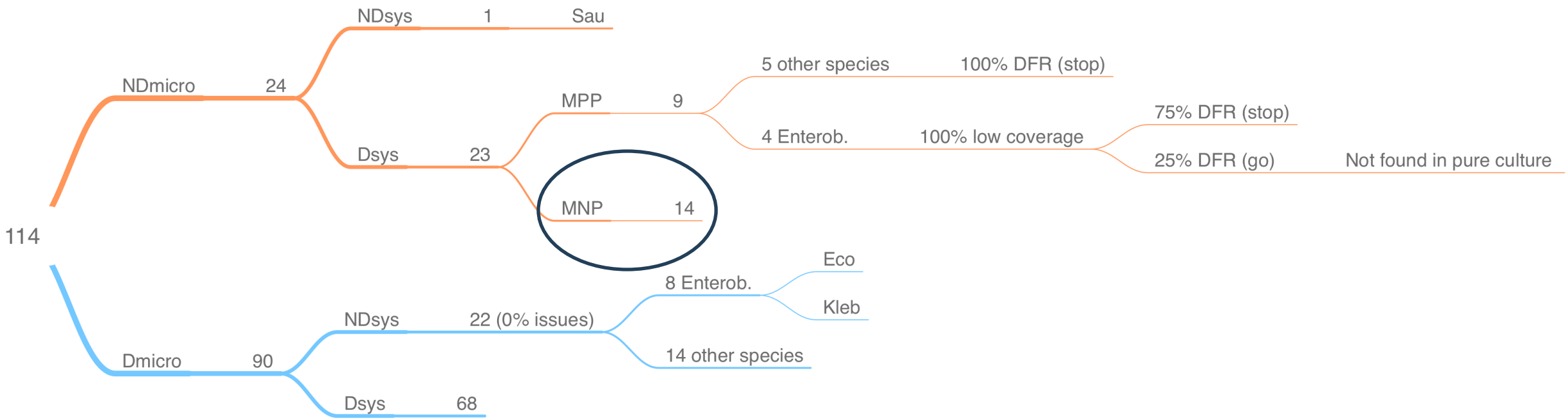
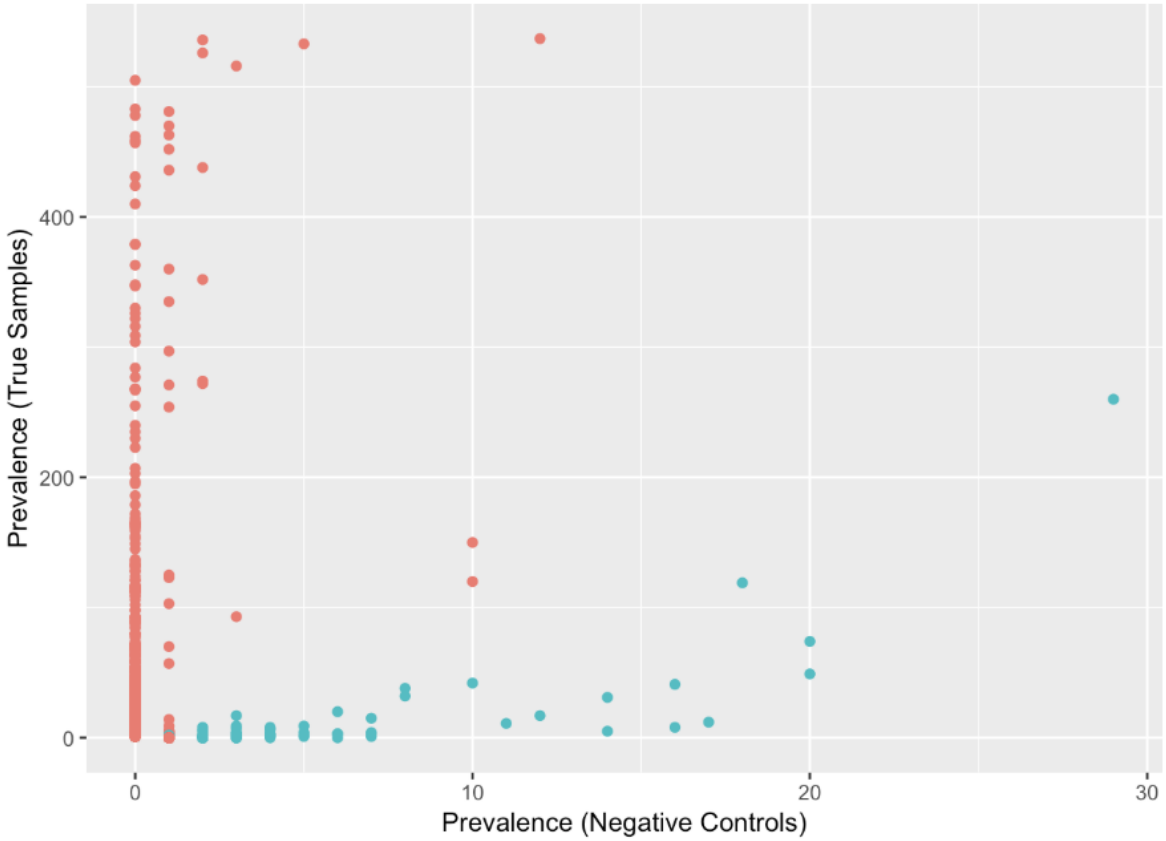

114
NDmicro
24
NDsys
1
Sau
Dsys
23
MPP
9
5 other species
100% DFR (stop)
4 Enterob.
100% low coverage
75% DFR (stop)
25% DFR (go)
Not found in pure culture
MNP
14
Dmicro
90
NDsys
22 (0% issues)
8 Enterob.
Eco
Kleb
14 other species
Dsys
68

Fig. S3

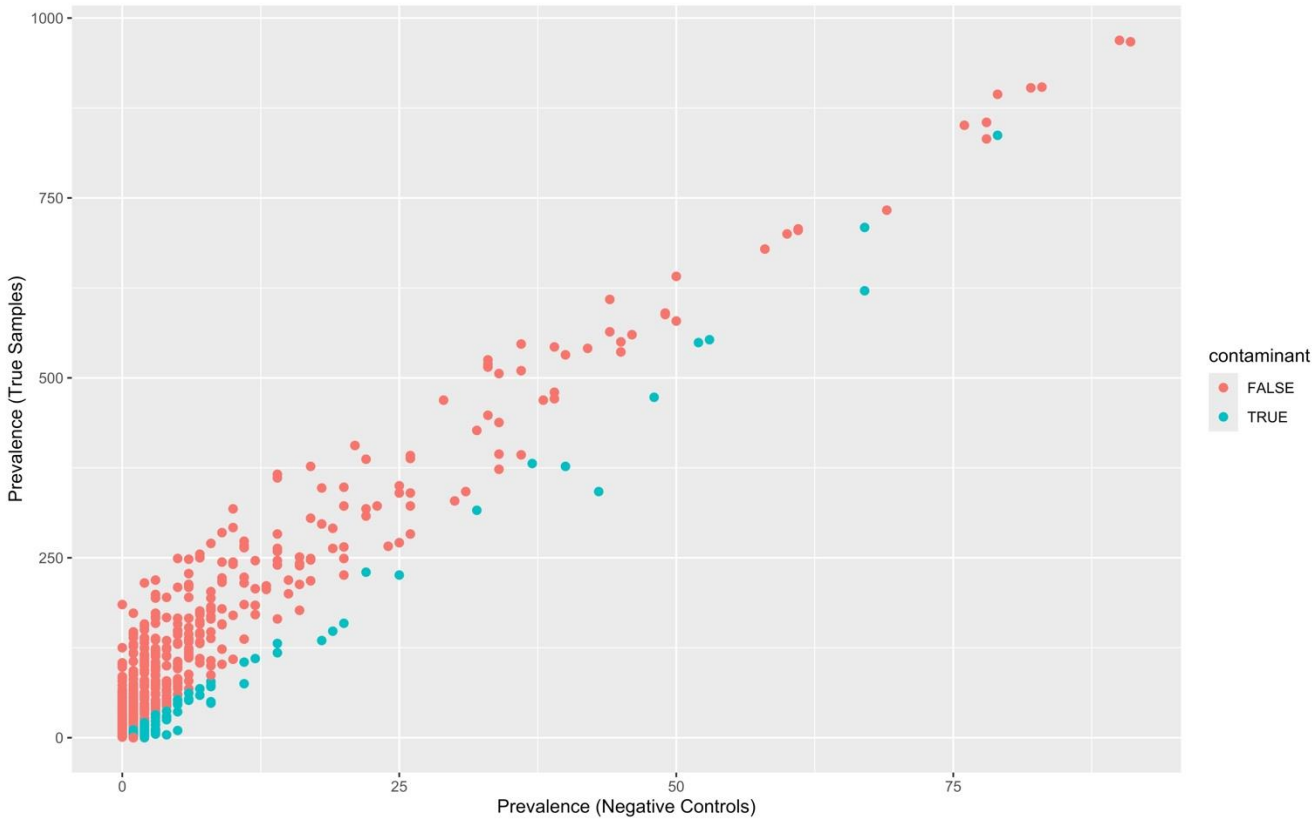
A

PS dataset (default – threshold 0.1)



B

Our dataset (threshold 0.5)



Excluded weeks: 2, 4, 7, 12, 13, 14, 15, 16, 20, 26, 29, 48

Fig. S4