Water Research 210 (2022) 117993

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Toward shotgun metagenomic approaches for microbial source tracking sewage spills based on laboratory mesocosms

Blake G. Lindner [a], Brittany Suttner [a], Kevin J. Zhu [a], Roth E. Conrad [b], Luis M. Rodriguez-R [a,c], Janet K. Hatt [a], Joe Brown [a,1], Konstantinos T. Konstantinidis [a,*]

[a] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[b] Ocean Science and Engineering, Georgia Institute of Technology, 311 Ferst Drive, ES&T Building, Room 3321, Atlanta, GA 30332, USA
[c] Department of Microbiology and Digital Science Center (DiSC), University of Innsbruck, Innsbruck, Tyrol 6020, Austria

ARTICLE INFO

Keywords:
Source tracking
Metagenomics
Sewage collection systems
Wastewater
Microbial ecology
Water quality

ABSTRACT

Little is known about the genomic diversity of the microbial communities associated with raw municipal wastewater (sewage), including whether microbial populations specific to sewage exist and how such populations could be used to improve source attribution and apportioning in contaminated waters. Herein, we used the influent of three wastewater treatment plants in Atlanta, Georgia (USA) to perturb laboratory freshwater mesocosms, simulating sewage contamination events, and followed these mesocosms with shotgun metagenomics over a 7-day observational period. We describe 15 abundant non-redundant bacterial metagenome-assembled genomes (MAGs) ubiquitous within all sewage inocula yet absent from the unperturbed freshwater control at our analytical limit of detection. Tracking the dynamics of the populations represented by these MAGs revealed varied decay kinetics, depending on (inferred) phenotypes, e.g., anaerobes decayed faster than aerobes under the well-aerated incubation conditions. Notably, a portion of these populations showed decay patterns similar to those of common markers, *Enterococcus* and HF183. Despite the apparent decay of these populations, the abundance of β-lactamase encoding genes remained high throughout incubation relative to the control. Lastly, we constructed genomic libraries representing several different fecal sources and outline a bioinformatic approach which leverages these libraries for identifying and apportioning contamination signal among multiple probable sources using shotgun metagenomic data.

1. Introduction

Wastewater collection systems (or simply, collection systems) represent an important engineering control for the collection of human feces, commercial or industrial wastewaters, and sometimes stormwater, particularly in certain urban settings. The operation and maintenance of collection systems pose unique challenges, often due to their size, complexity, and capital costs (Salman et al., 2012; Berendes et al., 2018; McLellan et al., 2018). Population growth and distribution changes – especially growing urbanization trends – highlight the importance of maintaining and expanding efficient collection systems for an increasing fraction of the global population (ten Veldhuis et al., 2010). Severe weather, pipe blockages, aging, and other issues of system failure can lead to the accidental release of untreated wastewater (sewage) from collection systems into waterways or floodwaters (Salman et al., 2012; Berendes et al., 2018; McLellan et al., 2018; Olds et al., 2018). As sewage is a significant reservoir of both chemical and biological pollutants, its release into the environment poses serious environmental and human health risks, including potential exposure to human pathogens (Ashbolt et al., 2010; Fouz et al., 2020; Medina et al., 2020; Eisenberg et al., 2016) and possible dissemination of antimicrobial resistance genes (ARGs) among microbial populations (Su et al., 2020; Kessler 2011; Lira et al., 2020).

Microbial source tracking (MST) refers to a collection of forensic tools developed to identify the presence and source of contamination among multiple probable fecal sources, including sewage (Harwood et al., 2014). In large part, the technical approaches behind MST methods have been developed in response to both the difficulty of

* Corresponding author.
E-mail address: kostas@ce.gatech.edu (K.T. Konstantinidis).
[1] Present address: Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, North, Carolina, NC 27599, United States

https://doi.org/10.1016/j.watres.2021.117993
Received 4 October 2021; Received in revised form 17 December 2021; Accepted 18 December 2021
Available online 25 December 2021

assaying for the diverse array of relevant human pathogens as well as the practical need to keep methods relatively rapid and inexpensive. Existing approaches have relied on indicator organisms to imply the presence of fecal pollution and sometimes as proxies for the presence of human pathogens in contaminated waters. Specifically, fecal indicator bacteria (FIB) include an aggregation of bacterial populations considered representatives of microbial communities inhabiting the guts of warm-blooded animals. Widely used indicator organisms include *Escherichia coli* and *Enterococcus* spp. More recently, MST genetic markers from distinct bacterial lineages have been used that leverage known host specificity of distinct populations for source attribution (Bernhard et al., 2000). Some markers (e.g., the HF183 primer targeting a human-associated *Bacteroides* clade) have found effective use in environmental management strategies as the basis for inferring the amount of sewage present and thereby, a potential array of pathogen concentrations for iterative risk assessment simulations (Boehm et al., 2015). Yet, the use of FIB and MST gene markers has had challenges: most notably, that the concentration of most markers are rarely found to co-vary with pathogen concentrations, marker concentrations fluctuate with sewage age and the capability of FIB to adapt to environmental conditions can all combine to confound results interpretation (Harwood et al., 2014; Korajkic et al., 2018; Ahmed et al., 2016, 2019; Devane et al., 2020).

In recent years, targeted metabarcoding methods have examined sewage and sewage-contaminated waters via the 16S rRNA gene or the internal transcribed spacer (ITS) for prokaryotes and fungi, respectively (Ahmed et al., 2016; Unno et al., 2018; McLellan and Eren, 2014; Assress et al., 2019). These studies have revealed a distinct sewage "microbiome" dominated by taxa that proliferate in collection systems, sometimes far beyond the abundance of human gut associated populations (Newton et al., 2015; McLellan and Roguet, 2019; McLellan et al., 2010). However, these single-gene assays offer limited resolution to distinguish between environmental or non-environmental strains of the same species due to conservation of the rRNA gene or the ITS region. Likewise, these methods do not provide information about the gene content associated with important populations (e.g., emergent pathogens, ARGs present) or resolve finer community-wide compositional shifts (Ahmed et al., 2016; Poretsky et al., 2014). Therefore, rRNA gene-based approaches are limited with respect to quantifying health risks associated with the detection of biomarkers or guide the development of more holistic environmental management criteria (e.g., site specific criteria).

Whole genome shotgun sequencing (or metagenomics), which recovers fragments of the genomes in a sample, have revealed that bacteria and archaea predominantly form sequence-discrete populations with intra-population genomic sequence relatedness typically ranging from approximately 95% to 100% average nucleotide identity (ANI) depending on the population considered – a level that the 16S rRNA gene cannot often assess (Caro-Quintero and Konstantinidis, 2012; Rodriguez-R et al., 2021). Metagenomic approaches offer unique advantages for environmental health monitoring tasks including: (1) extensive gene content information of abundant populations, (2) precise ecological estimates of relative abundance at the species level and (3) examination of intra-species diversity (Segata 2018). Despite its potential for circumventing some of the challenges facing existing MST and metabarcoding methods, whole genome shotgun sequencing has not been utilized in monitoring municipal sewage pollution. To date, metagenomic applications have focused on understanding the microbiology of biological wastewater treatment, treated effluents and their receiving waters, or viral populations (Lira et al., 2020; Cai and Zhang, 2013; Bibby and Peccia, 2013). In part, this is because it remains unclear how to best merge the methods and bioinformatics behind metagenomic practices with existing MST and environmental monitoring paradigms (Hong et al., 2020). Widespread application of this technology in the field requires that several outstanding issues be resolved, including the detection limits of metagenomic analyses, whether whole and/or metagenome-assembled genomes (MAGs) can serve as source-specific fecal contamination markers and how metagenomic approaches can infer the relative contribution of various fecal inputs (referred to hereafter as "source apportioning").

Here, we offer a metagenomic perspective on sewage-related bacterial populations and explore their relationships with culture and PCR-based markers during a simulated failure of a collection system (i.e., spill). Specifically, we simulated sewage contamination events in lake water obtained from a local drinking water and recreational use reservoir within dialysis bag laboratory mesocosms that were incubated in the dark for one week. Shotgun metagenomic sequencing was performed to search for potential sewage-specific biomarkers, test the effectiveness of genome collections for fecal source attribution and apportioning, and directly screen for both pathogens and antimicrobial resistance genes. We further support these aims by developing and testing a theoretical analytical limit of detection which can help guide the future application and interpretation of metagenomics to these issues.

2. Materials and methods

2.1. Sampling and mesocosm operation

Samples were collected in sterile glass 1 L bottles from the primary influent of three WWTPs located in the Atlanta Metropolitan region of Georgia (USA) to serve as representatives of sewage across three different sewersheds. Each sewershed was comprised of collection systems with separate stormwater and wastewater conveyance (i.e., separate sewers). Approximately 50 L of surface water from Lake Lanier, Georgia was also collected concurrently. Hereafter, these sample groups are referred to as sewersheds A, B, and C. All sewage and water samples were immediately transported to the lab and stored in darkness at 4 °C until mesocosm setup, which occurred within 24 h. For mesocosm setup, 40 L tanks were filled with lake water and a pump installed for aeration. Experimental dialysis bags were prepared with 110 mL 10% (v/v) sewage and lake water mixture and control bags were filled with 110 mL uninoculated lake water and closed on both ends using polypropylene Spectra/Por clamps (Spectrum Laboratories). Both experimental (n = 12 × 3 sewersheds = 36 bags) and control (n = 12 bags) dialysis bags were then added to the tank. A small headspace of air was left in each bag when sealing with clamps so that they could float freely in the tank. Dialysis bag pore sizes (6–8 kDa molecular weight cutoff) permit the transport of small molecules and ions, but bacterial and viral particles are contained within the bags. Mesocosms were kept in darkness at 22 °C throughout the duration of the experiment. Sampling occurred at 1, 4, and 7 days by retrieving experimental and control bags from the mesocosm for destructive processing.

2.2. Culturing, DNA purification, qPCR, and shotgun sequencing

EPA Method 1600 (USEPA, 2009) was followed for enumerating volumetric *Enterococcus* CFUs. Three replicates of each sample were diluted 10-fold and then plated in duplicate. All dilutions yielding measurements within an acceptable range for counting were averaged to estimate CFUs/100 mL for a sample. Mesocosm sampling, DNA extraction and subsequent qPCR analysis occurred as described previously in Suttner et al. (Suttner et al., 2021). Briefly, water samples were passed through 0.45 μm pore size polycarbonate (PC) membranes and stored at −80 °C in 2 mL screw cap bead tubes until processed (within 1–3 months). DNA was extracted from PC membranes using the Qiagen PowerFecal kit following the manufacturer's instructions with only one exception: mechanical cell lysis was performed by bead beating in two 1-minute intervals using the Biospec Mini-Beadbeater-24 with icing between intervals. These DNA extractions were used as template for qPCR with the HF183/BFDRev assay (Wade et al., 2010) and a universal 16S rRNA gene qPCR assay (GenBac16S) to quantify 16S rRNA gene copies across samples (Ritalahti et al., 2006). Metagenomic sequencing

was performed using the Illumina Nextera XT kit with library average insert size determined on an Agilent 2100 instrument using a HS DNA kit and library concentrations determined using the Qubit 1X dsDNA assay. Samples were then pooled and sequenced on the HiSeq 2500 instrument as described previously (Johnston et al., 2019).

All qPCR reactions were run using an Applied Biosystems 7500 Fast thermocycler and the cycling parameters were as follows: 2 min at 50 °C, 10 min at 95 °C, and 40 cycles of 15 s at 95 °C and 60 s at 60 °C. Assay reactions used 2 μL of template DNA in 20 μL qPCR reactions with the TaqMan Universal PCR Master Mix (Applied Biosystems). The primer and probe concentrations were 0.25 μM for HF183 assay and 0.3 μM for the Bac16S assay. Template DNAs were run diluted 5-fold (to remove the effect of PCR inhibitors) based on the expected marker concentration and quality of each sample. Further details on qPCR reaction set up and standard plasmids for absolute quantification are provided in Suttner et al. (2021) and reiterated within Supporting Information (**SI, Table S1**). To test for extraneous DNA and potential contamination from sample handling, 50 mL of sterile PBS was also filtered onto PC membranes and processed following the same DNA extraction at every sampling time point as described above.

2.3. Sequence data analysis

Short reads were quality trimmed and Nextera adapters removed with Trimmomatic 0.39 (Bolger et al., 2014). Quality trimming was performed to remove poor quality bases along both ends of sequences and subsequent removal of any sequences below 50 bp in length. *k*-mer based operation of Nonpareil 3.304 (-T kmer) was used to estimate the fraction of alpha diversity covered by the sequencing effort of each metagenome (Rodriguez-R et al., 2018). Beta diversity across trimmed short reads was assessed with the default settings of simka 1.5.1 based on Bray-Curtis dissimilarity values and visualized by principal coordinate analysis (PCoA) (Benoit et al., 2016). Kraken2 was used to assign taxonomy and estimate simple relative abundance against a custom library, including bacteria, archaea, viruses, protozoa, human, and fungal reference genomes at the rank of class (Wood et al., 2019). Trimmed short reads were assembled individually with IDBA (UD) 1.1.3 and SPAdes ("–meta") 3.14.0 using *k*-mer sizes between 20 and 127 (Peng et al., 2012; Prjibelski et al., 2020). Contigs shorter than 3 Kbp were removed prior to population genome binning, which was performed with MaxBin 2.2.7 and MetaBAT 2.12.1 (Wu et al., 2016; Kang et al., 2019). Additionally, in a parallel workflow, trimmed short reads were normalized via the BBNorm function of the BBtools suite (version 38) to bring depths between 10 and 30X sequencing depth and then subsequently assembled and binned as described above (Bushnell, 2014). All resulting metagenome-assembled genomes (MAGs) from both regular and depth-normalized short read assemblies were dereplicated using MiGA 0.7.24.0 via the derep_wf function (Rodriguez-R et al., 2018). Groups of MAGs sharing ANI $\geq$ 95% were clustered into species-like populations (hereafter, "populations") with representative MAGs for each population selected by highest completeness and lowest redundancy. Populations with no representative MAG having a MiGA quality score above 30% and/or redundancies below 5% were excluded from further analysis. Both Traitar 1.1.2 and MicrobeAnnotator were used with default settings to infer potential phenotypes and annotate draft genomes, respectively (Weimann et al., 2016; Ruiz-Perez et al., 2021). Lastly, MAGs were screened for cross-reactivity using the FastANI tool to search for other genomes with ANI $\geq$ 95% across a suite of reference databases (Jain et al., 2018).

From the PATRIC database, version 3.6.9, 1097 pathogenic bacterial genome accession IDs were recovered by querying for host name "Human, homo sapiens" and "good" quality. This included both genomes tagged as "Reference" ($n = 28$) and "Representative" ($n = 1069$) (Davis et al., 2020). Of these, 1076 genomes were recoverable from NCBI for use in this study. Abundance estimates of pathogen genomes were assessed by competitive short read mapping with Magic-BLAST 1.4.0 (-splice F) (Boratyn et al., 2019). Resulting alignments were filtered using minimum cut-off of 70 bp alignment length, 95% query coverage by alignment and 95% identity to avoid spurious matches. Additionally, for virulence gene detection, only experimentally verified nucleotide entries in the Virulence Factor Database (Liu et al., 2019) were used.

Evaluating MAG relative abundance across the time series was accomplished similarly using Magic-BLAST 1.4.0, where MAGs were concatenated into a single library to which reads were competitively mapped. Additionally, DIAMOND 2.0.1 (blastx –ultra-sensitive) was used to search short reads against the reference gene sequences of pre-compiled 150 bp β-lactamase ROCker models to reliably identify short reads belonging to β-lactamase encoding genes (Buchfink et al., 2015; Zhang et al., 2020). Reads mapping to these reference sequences were selected for best bit-score alignment and subsequently filtered by ROCker v1.5.2 as described previously (Orellana et al., 2017).

2.4. Detection and quantification of metagenomic features

For a reference genome, MAG, or gene to be considered detected in a sample, at least 10% of the target sequence was required to be covered by reads (i.e., breadth of coverage: hereafter, C), as proposed previously for robust detection of targets in metagenomic datasets (Castro et al., 2018). Or, as written, the analytical limit of detection (LOD) used here:

$$Analytical\ LOD:\ C \geq 0.1 \tag{1}$$

The LOD was automatically implemented by calculating sequencing depth and breadth similarly to Rodriguez-R et al. (2020) for estimating "Truncated Average Depth" at 80% (hereafter, the function TAD80). Python scripts used for this approach are available online at: https://github.com/rotheconrad/00_in-situ_GeneCoverage. In short, the TAD80 function estimates sequencing depth by first sorting genomic positions according to their sequencing depth and then removing the upper 10% and lower 10% of positions before averaging the sequencing depth along the remaining 80% of positions. Since truncation of targets with breadth of coverage near the detection limits (e.g., C $\approx$ 0.1) could introduce artificially lower values, a quantification threshold was also necessary to avoid systemic underestimation of abundance for targets near LOD. From Lander and Waterman (1988), breadth of coverage (C) is related to sequencing depth (ρ) by the following:

$$C = 1 - e^{-\rho} \tag{2}$$

Thus, for the analytical LOD defined above, the expected sequencing depth (ρ) is simply -ln(0.9) for targets at detectable limits. We formalize a quantification threshold which measures whether a target is quantifiable following application of the truncation function (TAD80) with:

$$Quantification\ Threshold: TAD80(\rho) \geq -\ln(0.9) \tag{3}$$

For simplicity in our metagenomic results, we describe those targets which satisfied the LOD condition but were below the quantification threshold as targets that were "detected but not quantifiable" (DNQ).

To convert relative abundance of detected target genomes to absolute abundances (e.g., cells/mL), the following approach was used. Single copy gene coverage or genome equivalents (GEQ) and average genome size (AGS) of metagenomes were evaluated using MicrobeCensus 1.1.0 (Nayfach and Pollard, 2015). The 16S rRNA gene-carrying reads were identified and extracted using sortmeRNA 4.2.0 and the average 16S rRNA gene coverage was estimated as the sum of extracted read lengths divided by 1540 bp, the average length of the bacterial 16S rRNA gene (Kopylova et al., 2012; Wang et al., 2007). Average 16S rRNA gene copy number (16S ACN) for each metagenome was determined by the ratio between 16S rRNA sequencing depth (ρ_{16S}) and GEQ:

$$16S\ rRNA\ ACN = \frac{\rho_{16s}}{GEQ} \tag{4}$$

The copy number of the 16S rRNA gene per mL as quantified by qPCR was divided by the 16S rRNA ACN to obtain an estimate for the number

of cells in each sample, assuming that one prokaryotic genome was approximately equivalent to one prokaryotic cell:

$$Estimated\ Prokaryotic\ Cell\ Density\ \left(\frac{cells}{mL}\right) = \frac{16S\ rRNA\ \left(\frac{copies}{mL}\right)}{16S\ rRNA\ ACN} \quad (5)$$

These measures were taken to help control for bias in relative abundance estimation due to changes in overall microbial load (cells per volume) and 16S rRNA gene ACN variation throughout the experiment (Lin and Peddada, 2020; Morton et al., 2019). Finally, absolute abundances were estimated by multiplying a population's genome equivalents by the estimate for the number of cells in a sample. This was accomplished using the following equation for a given population via the truncated average sequencing depth [TAD80(ρ)], GEQ and total estimated prokaryotic cell density:

$$Est.\ Pop.\ Cell\ Density\ \left(\frac{cells}{mL}\right) = \frac{TAD80(\rho)}{GEQ} * Est.\ Prok.\ Cell\ Density\ \left(\frac{cells}{mL}\right) \quad (6)$$

Further, an extension of our definitions of LOD was used in tandem with cell density estimations for theorizing the smallest abundance detectable as a function of GEQ and cell density via:

$$Detectable\ Pop.\ Size\ \left(\frac{cells}{mL}\right) \geq \frac{-\ln(0.9)}{GEQ} * Est.\ Prok.\ Cell\ Density\ \left(\frac{cells}{mL}\right) \quad (7)$$

2.5. Curation of source-specific genome collections

It was necessary to curate a collection of source-specific genomes in order to support our efforts to develop metagenomic based source attribution and apportioning approaches. In short, we collected reference genomes, MAGs, and isolate genomes from several large-scale studies of host microbiomes. These datasets included genomes gathered from the fecal microbiomes of humans ($n = 4644$ genomes), pigs ($n = 1667$ genomes), and chickens ($n = 5675$) and the rumen microbiome of cows ($n = 2124$ genomes) (Almeida et al., 2021; Stewart et al., 2019; Gilroy et al., 2021; Chen et al., 2021). MAGs produced in this study were also included as representatives of sewage sources. Detailed methods for the curation and dereplication of these collections are summarized in Supporting Information and visualized in SI Fig. 6S. Lastly, these libraries are hosted online for public use and download at http://enve-omics.ce.gatech.edu/data/mst_library.

2.6. Data availability

Sewage and mesocosm short reads as well as sewage-associated MAGs can be accessed through NCBI within BioProject PRJNA691978.

3. Results

3.1. Culture and qPCR data

Both fecal indicators (*Enterococcus* and HF183) were in the same order of magnitude across the sewage samples gathered as inoculum for the mesocosms. Sewage from sewersheds A and B contained counts with averages of 3.7E+04 and 3.1E+04 Enterococci CFUs/100 mL and 2.4E+06 and 3.6E+06 HF183 copies/mL, respectively. Within sewershed C, counts were lower having 1.3E+04 Enterococci CFUs/100 mL and 1.5E+06 HF183 copies/mL. Similarly, quantification of the 16S rRNA gene copy number within the inoculum indicated that overall, microbial loads were lower in sewershed C than sewersheds A and B at the time of sampling (SI Fig. S1). Monitoring Enterococci and HF183 qPCR markers across the mesocosm timeseries revealed that the markers decreased throughout the experiment in all replicates but were still detectable at day 7 and remained higher than the established or recommended water quality criteria for recreational use waters (i.e., 36 CFUs/100 mL and 41 HF183 copies/mL) (USEPA, 2015; Boehm et al., 2018). Only the HF183 marker within sewershed C mesocosm decreased below detection on Day 7 (Fig. 1). Neither marker was detected in the (un-inoculated) freshwater serving as control at any time point during mesocosm operation.

3.2. Estimated microbial load

The estimated prokaryotic cell density of the inoculum varied based on quantification of the 16S rRNA gene: 1.1E+09, 2.0E+09, and 1.8E+08 cells/mL were estimated for sewersheds A, B and C, respectively. Following dilution and mixing of the inoculum into the mesocosms, day 0 estimates for cell densities were 2.0E+07, 1.7E+08, and 2.5E+07 cells/mL. Thereafter, cell density in both sewershed A and sewershed C mesocosm increased considerably in the first 24 h to 1.8E+08 and 6.9E+07 estimated cells/mL (a 924% and 275% increase) while sewershed B decreased to 1.5E+08 cells/mL. Subsequent time points revealed steady decreases in cell densities approaching the control cell density at day 7 of 7.9E+05 cells/mL (**SI Table S2**).

3.3. Metagenomic coverage and compositional shifts

Between 1.5 Gbp to 3.5 Gbp of data per sample remained following read quality trimming and adapter removal, which corresponded to a range of 9 to 27 million reads. Sequencing effort covered between 36 and 67% of expected nucleotide diversity (N_d) across all samples based on the Nonpareil algorithm, which estimates sequence coverage based on the degree of redundancy among the metagenomic reads available for each dataset (Rodriguez-R et al., 2018). This level of coverage is adequate for comparing the abundance of features (e.g., genomes, genes) across samples (Rodriguez-R and Konstantinidis, 2014). N_d estimations of the inoculum and control samples were similar, and day 0 values closely followed that of their respective sources. A decrease in N_d occurred within the first 24 h for all three biological replicates; lower diversities were observed in day 1 samples compared to those for the inoculum, day 0 samples and the control. The sewershed B series increased in diversity for the remaining days while both sewersheds A and C fluctuate thereafter (**SI, Table S2**).

Observations of beta diversity revealed that the earlier timeseries samples (day 0 and day 1) remained quite similar to the inoculum. By day 4, considerable shifts in community composition were observed driving the sewage contaminated waters closer to the control (**SI Fig. S2**). *k*-mer mapping to characterize these community-wide shifts using Kraken2 at the class level showed the depletion of *Bacteroidia, Epsilonproteobacteria,* and *Clostridia* following inoculation. None of these classes were detectable in the control samples. An increase of *Gammaproteobacteria* abundance occurred within the first 24 h across all replicates after which this class gradually decreased in abundance with time. Additionally, increases in *Alphaproteobacteria* and *Cytophagia* occurred in later time points (day 4 and day 7), far beyond levels observed in the control, suggesting that the later timepoint samples had not yet fully recovered from perturbation. Class level relative metagenome-based abundances, qPCR, culture, and cell density estimation results are summarized on Fig. 1.

3.4. Sewage-associated population genome binning and taxonomic identification

Seven hundred twenty MAGs were recovered from inoculum and timeseries sample assemblies. The 720 MAGs were dereplicated at the ANI ≥ 95% level and the highest quality MAG per resulting ANI group was selected, generating a single representative MAG for 49 sequence discrete populations (hereafter, simply "populations"). Competitive read mapping to the representative MAG of these populations revealed two

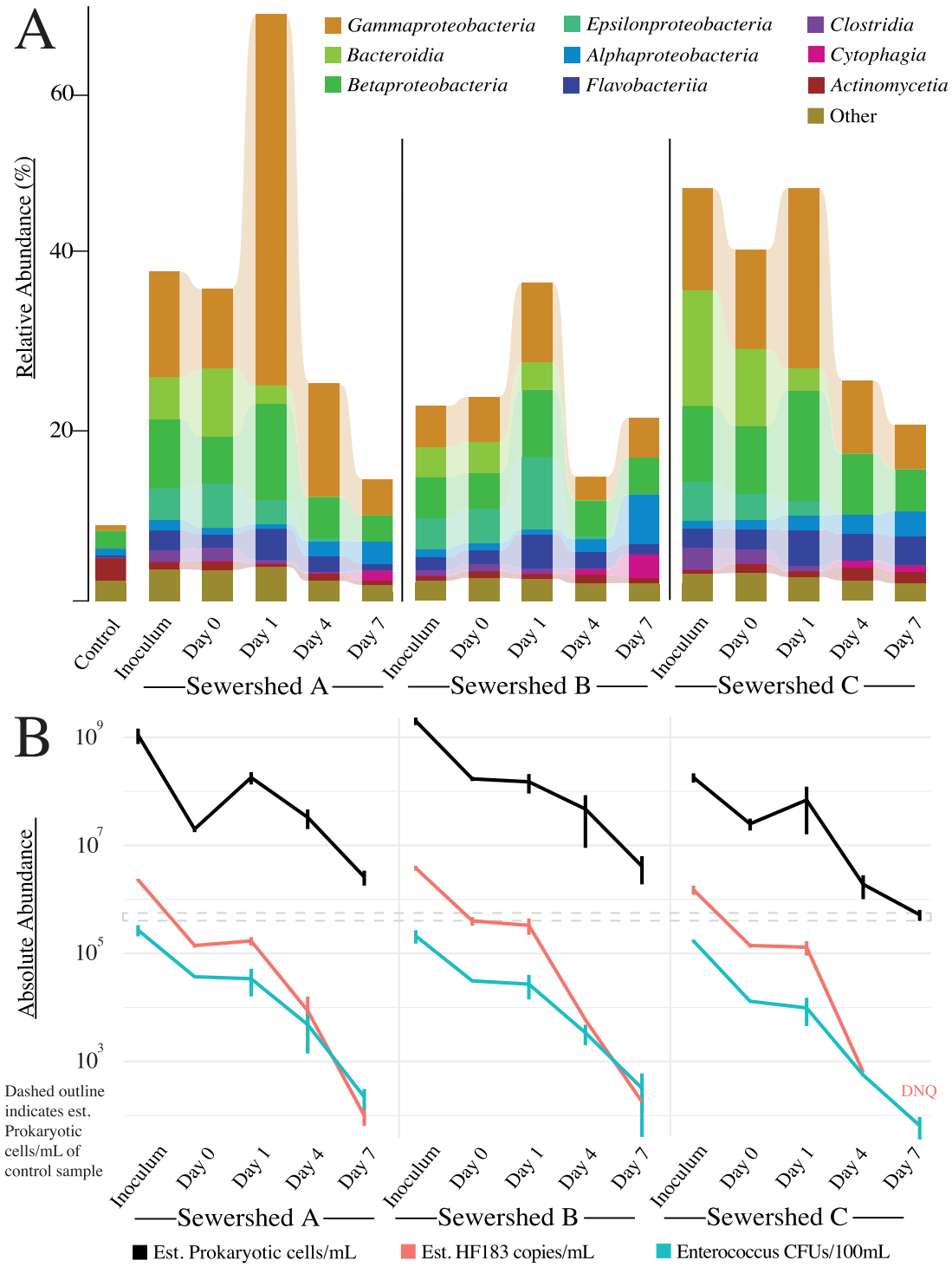

Fig. 1. Panel A: Class level abundances across control, inoculum and timeseries for sewersheds A, B and C based on kmer classification by Kraken2 against a custom-built database of reference genomes. Total height of bars represents the percentage of kmers confidently classified to the corresponding taxon (Figure key). The maximum and minimum percentages of kmers confidently classified were 69.0% from sewershed A day 1 and 8.9% from the control, respectively. **Panel B**: Estimated cell density, estimated HF183 copy concentration and Enterococci colony forming units (CFU) for the same samples. The dashed lines indicate the estimated cell density range for the control sample. HF183 was detected but not quantifiable (DNQ) for sewershed C on day 7.

groupings delineated by their presence or absence in the inoculum. Of the total 49, 33 populations were detected within sewage inoculum samples with varying degrees of prevalence across replicates. We selected a subset of 15 of these 33 populations that were above the quantification threshold in each inoculum sample, which we refer to as "sewage-associated populations". This selection process was motivated twofold: First, to focus only on core populations shared between the inoculum recovered from each sewershed examined herein. Second, as an effort to exclude potentially noisy, nonspecific, or transient populations from further analysis. The sewage-associated populations and their representative MAGs are summarized in Table 1. Additionally, we validated our analytical detection and quantification limits using mock data of known composition to ensure these criteria were suitable for identifying sewage-associated populations (**SI Table S3.A**) Sczyrba

Table 1
Summary of representative MAGs recovered in this study representing sewage-associated populations.

Taxonomic Summary											
Population	Confident Taxonomy (p<0.05)		Best match in MiGA TypeMat Database	Similarity (%)	Metric	Completeness (%)	Redundancy (%)	Length (Mbp)	N50 (bp)	CDS	GC (%)
01	Genus	*Arcobacter*	*Arcobacter cryaerophilus GCA 002,992,955*	92.8	ANI	87.7	1.9	1.38	7214	1519	28.77
03	Genus	*Acinetobacter*	*Acinetobacter johnsonii NZ CP065666*	96.5	ANI	78.3	0	1.99	6506	2157	41.9
04	Genus	*Aeromonas*	*Aeromonas caviae GCA 000,819,785*	93.5	ANI	56.6	0.9	2.99	6625	3099	61.78
13	Class	*Bacteroidia*	*Paludibacter propionicigenes WB4 NC 014,734*	55.1	AAI	51.9	1.9	0.8	4680	764	39.63
15	Species	*A. caviae*	*Aeromonas caviae GCA 00,820,265*	98.0	ANI	40.6	0	1.57	4876	1684	61.79
18	Genus	*Cloacibacterium*	*Cloacibacterium rupense GCA 014,645,495*	88.2	ANI	61.3	3.8	1.58	5371	1596	33.27
19	Family	*Campylobacteraceae*	*Arcobacter suis CECT 7833 NZ CP032100*	72.1	AAI	49.1	0	0.95	5098	1130	28.6
28	Order	*Neisseriales*	*Rivicola pingtungensis GCA 003,201,855*	67.3	AAI	75.5	0	1.13	5350	1176	56.67
29	Genus	*Moraxella*	*Moraxella osloensis GCA 001,679,175*	95.4	ANI	75.5	0	1.83	9146	1726	44.48
30	Species	*A. temperans*	*Acidovorax temperans GCA 006,716,905*	97.3	ANI	91.5	0.9	2.8	8597	2816	63.59
33	Genus	*Flavobacterium*	*Flavobacterium succinicans LMG 10,402 GCA 000,611,675*	87.3	ANI	88.7	2.8	2.81	10,562	2699	35.43
43	Species	*P. copri*	*Prevotella copri DSM 18,205 GCA 009,495,405*	97.1	ANI	52.8	0	2.36	11,303	1981	46.62
44	Species	*B. vulgatus*	*Bacteroides vulgatus ATCC 8482 NC 009,614*	99.0	ANI	49.1	0	2.67	5144	2496	41.9
47	Family	*Aeromonadaceae*	*Tolumonas auensis DSM 9187 NC 012,691*	83.5	ANI	98.1	1.9	2.67	16,590	2612	47.97
49	Species	*R. pingtungensis*	*Rivicola pingtungensis GCA 003,201,855*	97.5	ANI	46.2	0.9	2.03	8236	2031	62.89

Footnote: "Metric" refers to whether average nucleotide (ANI) or average amino acid identity (AAI) was used to calculate similarity. "Completeness" indicates what percentage of single-copy marker genes appear in a MAG. "Redundancy" (or "Contamination") indicates the frequency at which multple copies of those same single-copy genes appear in a MAG. "N50″ represents the contig length at which contigs covering 50% of the MAG are greater than or equal to its value. "CDS" represents the number of predicted coding sequences in a MAG.

et al., 2017). We found our approach, as described in Materials and Methods (Eqs. (1) and (2), robust for reducing quantification error and detected targets as expected according to sequencing effort and target genome size, except on very limited occasions when close relatives were present in the sample at relative abundances many times greater than the target. (**SI Table S3B,C**).

Our collection of ubiquitous sewage-associated populations in sewersheds A, B, and C represented, respectively, 9.5%, 5.7%, and 13.3% of the total reads in inoculum metagenomes and 15.9%, 8.8%, and 19.6% of GEQ (genome equivalents). Estimated absolute abundances of these populations varied across the samples, from a maximum of 4.4E+07 cells/mL (Pop.01, sewershed B) to a minimum of 2.3E+05 cells/mL (Pop.04, sewershed C). Within the inoculum, the median and mean absolute abundances of an individual sewage-associated population was 5.3E+06 and 8.4E+06 cells/mL, respectively. Overall, sewershed C had substantially lower population densities due to the difference in total microbial load compared to sewersheds A and B, as noted above. Consistently, the sewage-associated populations presented here capture a larger portion of the metagenomic samples associated with sewershed C (compared to A or B), further indicating that the sewershed C samples may have simply had more dilute microbial load at the time of sampling. Overall, these results reveal that this collection of populations consistently represent highly abundant members of the sewage microbiome across these biological replicates and possibly a substantial part of the total sewage microbial community.

Comparison of the corresponding representative MAG sequences against type material in the MiGA "TypeMat" database (Rodriguez-R et al., 2018a) revealed several entries with close matches to previously described taxa at the species level (e.g., >95% ANI) including *Aeromonas caviae* (Pop.15), *Acidovorax temperans* (Pop.30), *Prevotella copri* (Pop.43), *Bacteroides vulgatus* (Pop.44), and *Rivicola pingtungensis* (Pop.49). Of the remaining, six populations matched known genus-level representatives, potentially representing a novel species of the matching genera. Two populations matched members of a known family, one to members of a known order, and one to members of a known class (Table 1). The population with the most distant match in the database (Pop.13, matching class *Bacteroidia*) with 55.1% average amino acid identity (AAI) to *Paludibacter propioncigenes*.

Collections of bacterial isolate genomes and/or MAGs from freshwater (Rodriguez-R et al., 2020), activated sludge (Ye et al., 2020), anaerobic digestors (Campanaro et al., 2020), the human gut environments (Almeida et al., 2021), and the broad general-purpose GEMs catalog (Nayfach et al., 2021), were examined to assess specificity between these 15 sewage-associated populations and other microbiomes. Of these sewage-associated populations, some (n = 11) may belong to species with members also inhabiting non-sewage microbiomes such as biological wastewater treatment processes or the human gut (**SI Table S4**). Importantly, only a single population, *Moraxella* (Pop.29), was found via these database searches to match (95.1–95.0% ANI, borderline of universal species cutoff) genomes recovered from aquatic environments (both marine and freshwater) (Rodriguez-R et al., 2018b). This finding suggests Population 29 could be less effective as an entry in a sewage-specific genomic library utilized for MST approaches if other *Moraxella* are in high abundance within unperturbed environments.

3.5. Sewage-associated population decay and putative phenotyping

Overall, all populations experienced rapid decline in estimated cell densities across the timeseries with most populations below detection limits following day 4. *Acinetobacter* sp., *Cloacibacterium sp*., *Acidovorax temperans*, and *Flavobacterium sp*. (Pop.03, Pop.18, Pop.30 and Pop.33, respectively) were detectable in at least one biological replicate at day 7 but most of these observations were below quantification. Signal from sewershed A had the greatest persistence; of the four mesocosms with quantifiable levels of a sewage-associated population by day 7, three belonged to the series of sewershed A. Notably, *Acidovorax temperans* (Pop.30) was the only population detected at day 7 in all three sewersheds (Fig. 2).

All populations remaining detectable at day 7 were putatively phenotyped as aerobic or facultatively anaerobic by Traitar analysis except for *Cloacibacterium* sp. (Pop.18), which could not be confidently classified. Nonetheless, *Cloacibacterium* sp. belongs to a genus of facultative anaerobes (*Cloacibacterium*), suggesting that it likely is a facultative population and that the representative MAG did not contain the necessary genes for confident phenotyping due to incompleteness. No population – regardless of (predicted) preference for oxygen – showed an increased estimated cell density outside the first 24 h of the incubation. All sewage-associated populations were likely gram negative, rod or oval-shaped bacteria as predicted by Traitar (SI Fig. S4).

3.6. Human markers and sewage-associated populations

Our results suggested that several of the sewage-associated populations are possibly linked to the human gut microbiome (**SI Table S4**). Based on whole genome comparisons (via ANI), Pop.43 and Pop.44 were assigned to *Bacteroidales* lineages that likely represent different clades than those represented by HF183. This was concluded based on either analysis of the 16S rRNA genes carried by these populations' representative MAG (HF183 is a 16S rRNA gene-based marker) or, if a 16S rRNA gene was not binned with the MAG, the 16S rRNA genes carried on the closest matching cultured relative showing at least 95% ANI to the representative MAG (See Table 1). In either case, HF183 was not a match for Pop.43 or Pop.44 which is consistent with the notion that HF183 typically belongs to *B. dorei* (*Phocaeicola dorei*) and its closest relatives. Modeling the linear relationship between either HF183 or *Enterococcus* concentrations against the estimated cell densities of the sewage-associated populations revealed divergent results for both markers. Specifically, HF183 had excellent correlations against some populations (i.e., anaerobic Pop.43 and Pop.44, and aerobic Pop.30 and Pop.28) but highly variable correlations overall (R^2 between 0.35 to 0.97) while *Enterococcus* had worse correlations but with a tighter range (R^2 between 0.5 to 0.8) (Fig. 3). As noted above, not all the sewage-associated populations highlighted as potentially co-habiting the human gut co-varied in abundance as well with HF183 concentrations. For example, correlations with HF183 concentrations were moderate with the presumed aerobes of Pop.03 ($R^2 = 0.69$) and Pop.29 ($R^2 = 0.75$) but poor for the facultative anaerobic Pop.15 ($R^2 = 0.35$).

3.7. Source attribution and apportioning assessment

Source specific genomic libraries were collected and curated as described above and in the Supporting Information. These libraries contain genomes representing populations which are likely restricted to a particular contamination source. Short reads from the metagenomes collected across the incubation were mapped to these source specific libraries via Magic-BLAST and normalized to both genome length and GEQ as described above. The results of this exercise provide an estimation for the percentage of prokaryotic cells likely originating from a particular contamination source (Fig. 4A). No source category was detected in the control sample. Further, human and sewage signals dominated the timeseries across each sewershed – though these signals showed rapid decline following day 4. The pig, cow, and chicken source categories were either not detected or were consistently $\leq$0.1% GEQ.

3.8. Pathogen and virulence genes assessment

To assess the ability of the metagenomic approach to provide insights into the health risk associated with bacterial pathogens introduced by sewage contamination during mesocosm operation, we recruited metagenomic short reads to 1076 pathogenic bacterial genomes recovered from the PATRIC webserver (**Supplement, Table S5**). Results revealed that 63, 38, and 129 pathogen genomes from sewersheds A, B, and C, respectively, within the inoculum had sequencing depths at or above our established LOD after read mapping (**Supplement, Table S6**). In contrast, immediately following inoculation on day 0 many reference genomes were no longer detectable, with a total of 61, 25, and 20 pathogenic genomes detected from sewersheds A, B, and C, respectively. Obviously, for many of these organisms, pathogenicity is a function of exact genotype (e.g., the *E. coli* pathotypes) and the methods used herein were developed for species-level detection and not optimized for distinguishing between closely related genotypes of the same species at low abundances (Castro et al., 2018).

Therefore, due to the low relative abundances of these pathogens that we observed and the need to assess the actual genetic content present within these populations, we examined the relative abundance of experimentally verified genes within the Bacterial Virulence Factor Database (VFDB) as proxies for key bacterial pathogens (Fig. 4B). The virulence signal within inoculum metagenomes primarily comprised those belonging to *Aeromonas, Klebsiella*, and *Shigella* pathogenic genera, consistent with the whole-genome detection results above. Sewage from both sewershed A and C appeared to have greater virulence factor signals compared to sewage from sewershed B, which had drastically lower detected levels of *Aeromonas* VFs (virulence factors) and no detection of *Klebsiella, Shigella* or *Escherichia* VFs. Within the sewershed A and C timeseries, average virulence abundance was lower on day 0 than in the inoculum but quickly reached a maximum in 24 h before substantially decreasing by day 4 and being below detection by day 7. The change was primarily due to a substantial increase in the abundance of *Aeromonas hydrophila* VFs. This trend was consistent among genes *hlyA* (hemolysin), *aerA* (aerolysin) and *act* (*Aeromonas* enterotoxin) – essential cytotoxins for *Aeromonas* spp. pathogenicity – across the timeseries. Alignment of these three cytotoxin genes to the MAG representing Pop. 15 revealed that it likely carries a gene encoding for *hlyA* but *aerA* and *act* were either not binned with the draft genome or truly not carried by this population. Upon further inquiry, the closest matching entry on NCBI's Genome database was *Aeromonas caviae* NZ_AP022214 (ANI = 98.0%), which represents a strain isolated from a Japanese wastewater treatment plant that has not been implicated in disease or designated as an obligate pathogen. Hence, to what extent the MAG identified represents a pathogenic or opportunistic pathogenic population remains somewhat speculative.

3.9. β-lactam resistance gene assessment

Several classes representing the breadth of β-lactamase-encoding gene diversity were present in the metagenomes from all samples. The uninoculated lake water (control) sample showed very low abundance of β-lactamase encoding genes across each class (sum of classes was 0.078 total β-lactamase encoding genes/genome equivalent) – though a subset of metallo-β-lactamase encoding genes (MBLS3) was noticeably pronounced (0.06 gene copies/genome equivalent). In the inoculum samples, total observed β-lactamase signal was much greater in sewersheds A and C (1.07 and 1.14 total gene copies /genome equivalent, respectively) compared to sewershed B (0.51 total β-lactamase encoding genes/genome equivalent), but the relative contribution of each class was consistent, with genes encoding for BlaA, BlaC and OXA dominating. In contrast, by day 4 and to a greater extent by day 7, the

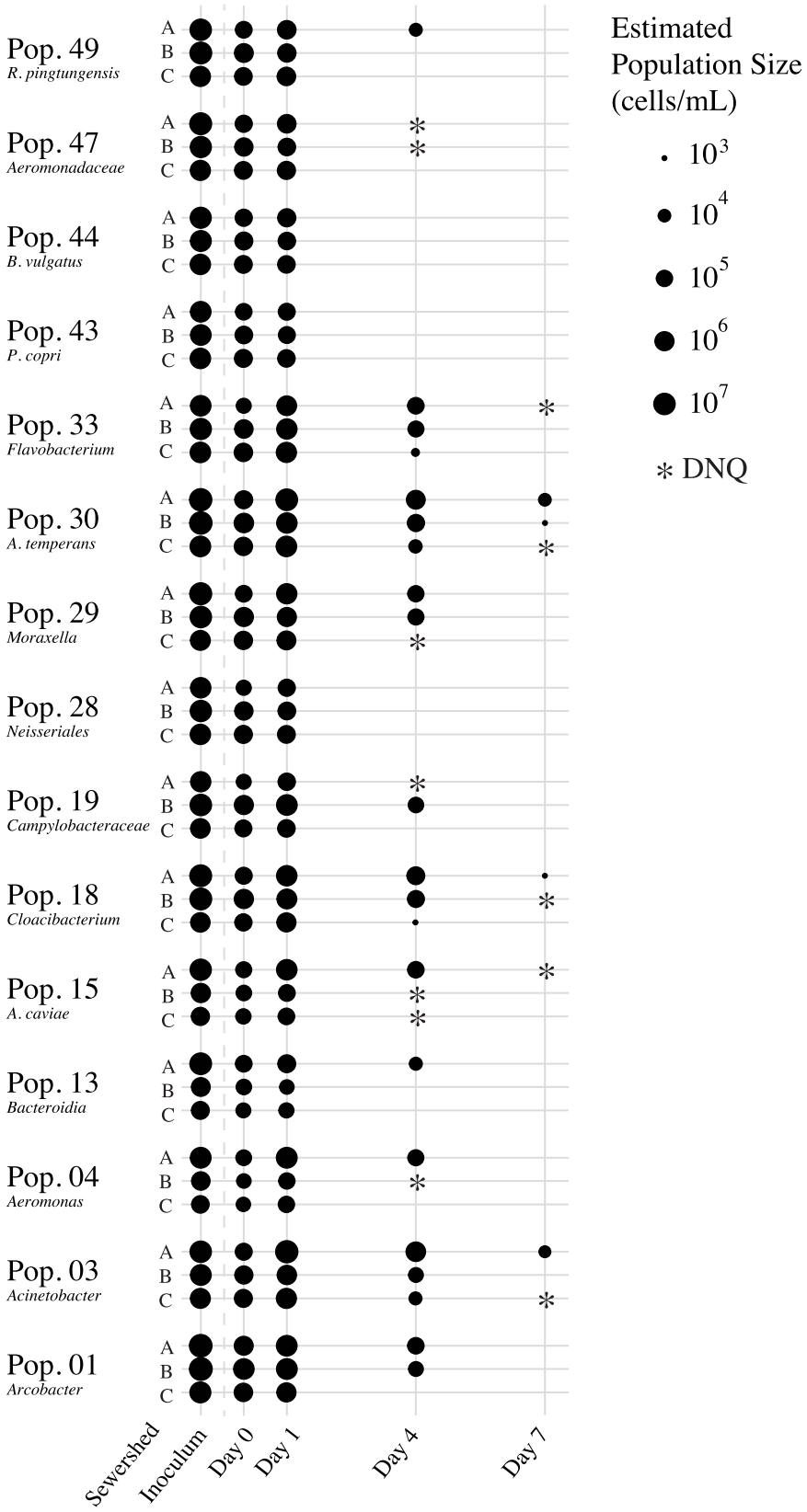

Fig. 2. Estimated cell densities of sewage-associated populations across inoculum and timeseries samples. Cell densities (absolute abundances) were estimated as described in the Materials and Methods section (via Eq. (6)). Populations that were detectable (via Eq. (2)) but that "did not quantify" (DNQ) above our quantification threshold (via Eq. (3)) are labelled with an asterisk.

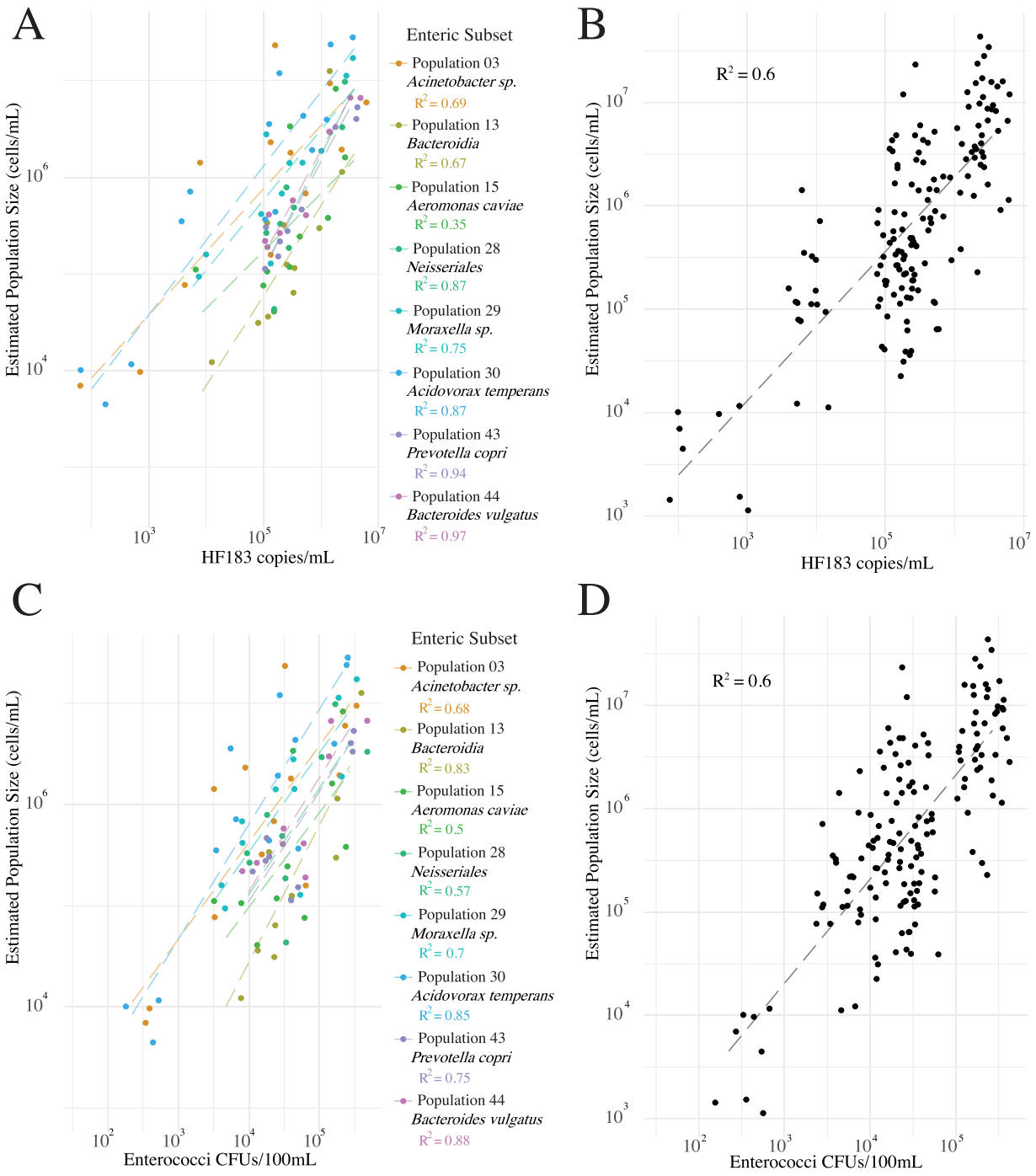

Fig. 3. Log-log scatter plots of estimated population densities across inoculum and timeseries samples against HF183 and Enterococci concentrations. Lines of best fit are shown dashed with their associated coefficients. Panel A: HF183 copy number versus the concentration of sewage-associated populations likely to also be enteric ($n = 8$). Panel B: HF183 copy number versus the concentration of all sewage-associated populations ($n = 15$). Panel C: Enterococci concentration versus the concentration of sewage-associated populations likely to also be enteric ($n = 8$). Panel D: Enterococci concentration versus the concentration of all sewage-associated populations ($n = 15$).

frequency of genes encoding for BlaA, BlaC and OXA decreased consistently while those encoding for MBLs increased (Fig. 4C). Along with a shift in prominence of these β-lactamase gene classes, both sewersheds A and C showed steep decreases in the relative number of β-lactamase encoding genes/genome equivalent between day 0 and day 7. Sewershed C showed the same shifts in prominence between classes, yet total signal remained consistent with 0.55 and 0.54 total β-lactamase gene copies/genome equivalent on day 0 and 7, respectively.

4. Discussion

4.1. Sewershed microbial diversity

Collection systems represent a key component of modern sanitation infrastructure. Despite the importance of sewage as a reservoir for human pathogens, antimicrobial resistance genes and the recent widespread utilization of wastewater-based epidemiology, the sewage microbiome remains relatively understudied at the whole genome level.

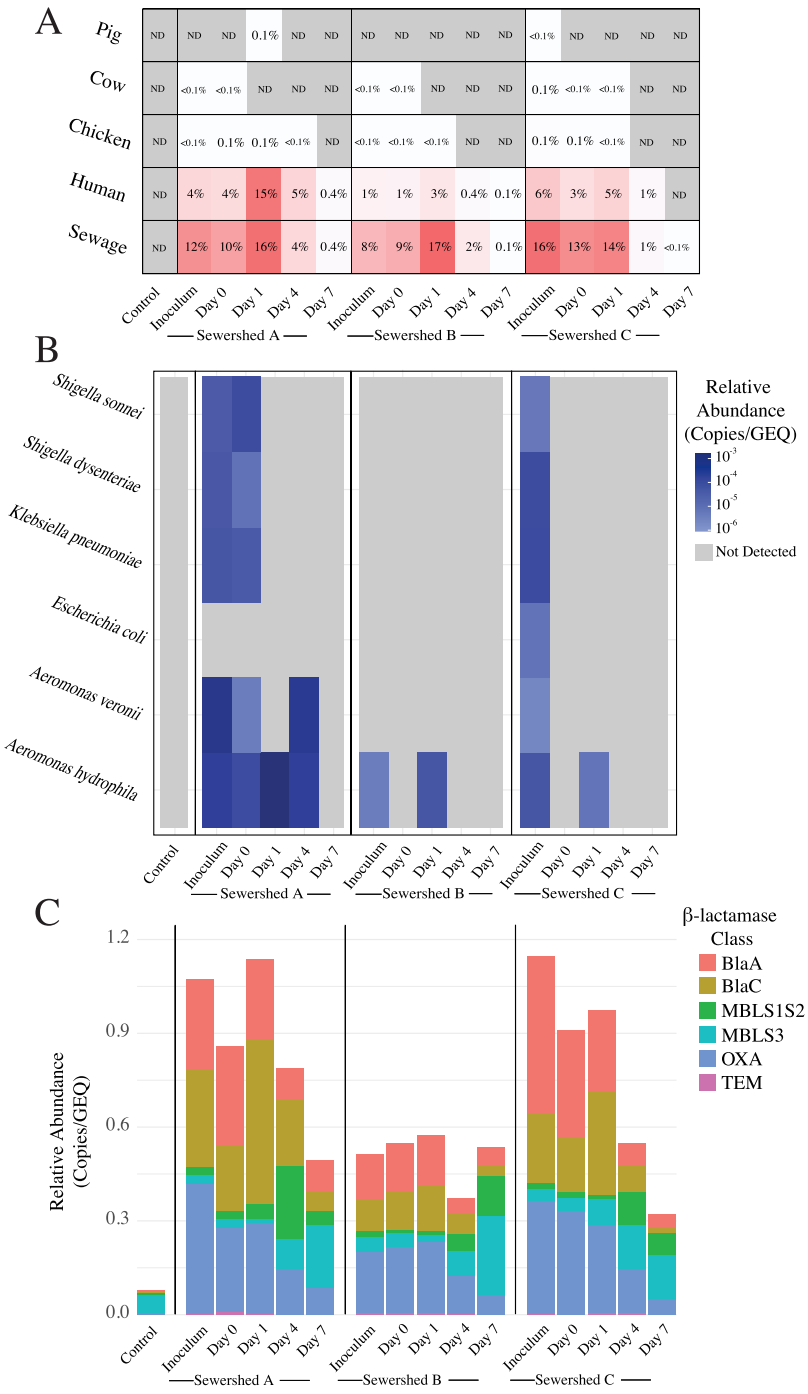

Fig. 4. Abundance patterns of source specific genomic libraries, virulence factors, and β-lactamase encoding genes across inoculum and timeseries metagenomes. All normalization was performed against genome equivalents (GEQ). **Panel A**: Source attribution and apportioning results based on reads mapped against MAGs curated for different fecal sources. Percentages represent an estimation for the fraction of the prokaryotic cells which can be confidently attributed as belonging to one of the fecal sources. **Panel B**: Virulence factor (VF) gene abundance dynamics based on short reads mapping against experimentally verified VF reference nucleotide sequences. **Panel B**: β -lactamase encoding gene (BLA) abundance dynamics across inoculum, timeseries and control metagenome based on searches of reads against reference ARG sequences and ROCker model filtering of the resulting matches. Relative abundance is calculated by normalizing the average sequencing depth of each gene to GEQ after ROCker filtering and then summing across BLA classes.

Our results indicated that the sewage samples we collected from three separate collection systems were dominated by what have been aptly named microbial "weeds" in literature and which we have observed as belonging to several sewage-associated populations that appear quite prolific (Assress et al., 2019; Newton et al., 2015) (Fig. 1A). Others have reported that several of these populations are also present at high relative abundances within sewersheds spanning another urban landscape (VandeWalle et al., 2012).

These sewage-associated populations showed different preference for oxygen, appearing to span strict anaerobic, facultative, and aerobic metabolic phenotypes. Notably, the signal associated with these populations in the metagenomic datasets decayed non-uniformly during mesocosm operation, though the most persistent populations were aerotolerant, acetate-utilizing populations which contained genes related to aromatic degradation and/or nitrogen metabolism. Depending on additional inquiry, it may be possible to leverage the ratio between abundances of anaerobic and aerobic (or facultatively anaerobic) sewage-associated populations in future work for inferring the date of pollution events linked to sewage contamination. For all 15 populations described here, their linear relationship with HF183 and Enterococci had a combined R^2 of 0.6 (Fig. 2), revealing overall consistent results for different markers under the conditions tested here. However, these correlations were drawn from the limited number of mesocosm incubations and in situ population dynamics are likely to differ according to varying environmental and biological factors which were not controlled for herein (Ahmed et al., 2019).

4.2. Source attribution and apportioning with source specific genomic libraries

Populations specific to municipal sewage likely exist and represent a subset of the microbiome of collection systems which – if better catalogued – may be useful for identifying and quantifying sewage pollution in natural ecosystems (SI Fig. S6). We demonstrated, through a proof-of-concept workflow, the capacity for read mapping of metagenomic datasets to curated source specific genomic libraries to perform simultaneous source attribution and apportioning. This approach yields a relatively easy-to-interpret metric representing the approximate percentage of prokaryotic cells within a sample that belong to a contamination source (Fig. 4A). Importantly, our approach represents a novel development given that current approaches utilizing sequence data for MST problems are not designed to distinguish between multiple fecal sources (McGhee et al., 2020) or cannot directly assess source apportioning between multiple sources (Roguet et al., 2020).

4.3. β-lactamase encoding genes surveillance

Additionally, we leveraged our metagenomes to survey for β-lactamase encoding genes across the inoculum and timeseries. The abundance of β-lactamases across the inoculum samples was substantially higher (7–15 times) compared to the control (Fig. 4**C**). This result was consistent with the literature regarding heightened ARG abundance within collection systems (Li et al., 2021). Specifically, others have reported substantial abundances of β -lactamase OXA genes on both *Campylobacteraceae* and *Aeromonadaceae* clades in sewage (Hultman et al., 2018). Indeed, the abundance of reads belonging to β-lactamase encoding genes, especially of the OXA-encoding class, were the most abundant in the inoculum and early time points where these sewage-associated clades (e.g., Pop.01, Pop.19) persisted in the lake water. Overall, these results indicated that sewage contamination imparted a substantial and lasting increase to the abundance of genes encoding β-lactamases even after 7 days following the contamination event (Fig. 4C). More work is needed to elucidate the genomic context of this increased β-lactamase encoding gene abundance (e.g., whether they belong to or have been transferred to organisms capable of driving clinically relevant cases of antimicrobial resistance). Nonetheless, our results allow for a quantitative view of the abundance of these genes relative to the natural environment, which could be quite relevant for assessing associated health risks as part of future work.

4.4. Shotgun sequencing and monitoring environmental waters

Importantly, although sewershed A and B showed what appears to be similar concentrations of human input according to HF183 concentrations within the inoculum (SI Fig. S1), the pathogen detection results revealed via the sequence data were quite varied (Fig. 4B, SI Table S6). Results from both read mapping to bacterial pathogen genomes and the experimentally verified VFDB collection were consistent in suggesting that bacterial virulence may have been more elevated in the sewershed A inoculum compared to sewershed B. This contrast between sewersheds with equal human marker concentrations yet apparently unequal bacterial pathogen load illustrates how shotgun sequence data can facilitate perspectives on the actual co-variance of marker and pathogen. Yet these insights clearly depend on sufficient sequencing effort and/or relatively high pathogen concentrations to avoid the possibility of false negative results.

In particular, the estimated smallest detectable population size associated with our analysis and sequencing effort ranged between approximately 2E+05 to 1E+02 cells/mL based on qPCR-based cell count normalization and the sequencing effort applied (Materials and Methods, **SI Table S2**). Approaches for estimating analytical LOD within metagenomic based analysis remain rare within the literature, especially as it relates to work done in the environment as opposed to clinical settings (Wendl et al., 2013; Ebinger et al., 2021). Yet, the concept of detection and quantification limits in metagenomics is a major challenge to its thorough incorporation into environmental monitoring approaches because 1) it is necessary to track biomarkers or pathogens down to quite low relative abundance in the field (e.g., at frequencies <1E-09 target basepairs/total basepairs), and 2) leveraging extraordinary sequencing effort is currently expensive and not practical when limitations of expertise and computational resources exist. Our approach provides the means to establish theoretical analytical LOD for metagenomic analyses based on sequencing effort which is useful for determining and interpreting the meaning of "non-detects".

Additionally, using average genome size (AGS) and total cell density estimates within the inoculum, we estimate that approximately 3.5Tbp of sequencing effort is necessary for detecting a population with concentration of 1E+02 cells/mL within the high microbial loading conditions such as those observed in the inoculum (sewage). In contrast, following the decline in cell density and increase in AGS across the timeseries, the estimated sequencing effort required to detect a population of 1E+02 cells/mL drops to 10Gbp in day 7 conditions (which had far smaller microbial loads). Therefore, our approach and results reported here for sequencing effort estimation may be helpful for informing the planning and execution of future environmental monitoring work utilizing metagenomic approaches (**SI, Table S7**). Though, crucial to note is the fact that our approaches for analytical LOD, and sequencing effort estimation assumes unbiased sequencing and does not consider sampling or processing recoveries – where the latter limitation is obviously broadly applicable to all molecular methods. Total detection limits, in the context of analytical limits as well as both sequencing bias and sampling/processing recoveries, will be important caveats to consider for future metagenomic workflows aiming to surveil pathogens in sewage collection systems and their releases into the natural and built environment (Hull et al., 2019).

Our efforts have shown how metagenomic datasets can provide insights on multiple questions critical to environmental monitoring and water quality: pathogen detection, source attribution and apportioning, and ARG persistence in the environment. In our view, confident and direct detection of pathogens within metagenomic datasets will remain primarily a logistical challenge due to the large amount of sequencing effort required to reliably detect bacterial pathogens at concentrations that are very low yet still quite relevant for safeguarding public health. For example, we have shown how via metagenomics one could track a broad range of population sizes – about five orders of magnitude (from about 1E+01 to 1E+02 cells/mL) – but that reliable detection depends on both sequencing effort and microbial load.

Thus, when performed alone, metagenomic approaches are unlikely to be the most prudent technology for routine monitoring and directly informing health risks associated with sewage contamination, especially when pathogen or virulence genes are at these relatively low abundances (e.g., below 1E+02 features/mL). This issue is also compounded by the large contribution of non-bacterial pathogens (e.g., viruses and protozoa) to illness risk in contaminated waters. In contrast, metagenomic approaches are increasingly poised to resolve questions related to source attribution and apportioning by improving our understanding (and the size of public databases) of the genomes maintained by source-specific microbial populations.

4.5. Limitations

Our dataset is of limited size and scope considering that, on a global scale, we examined sewage from collection systems in essentially equivalent geographies. The assortment of sewage-associated populations described here, although ubiquitous across the sewersheds we sampled, likely maintain differing prevalence across time or space. Furthermore, many draft genomes we produced are not complete, so further work will be needed to establish more practical views on both the geographic range of these populations and their genomic content

and diversity. Yet, we see advancing our knowledge of sewage-associated populations as a potential contribution towards newly developing forensic approaches that help monitor, manage, and repair essential infrastructure (Gonzalez et al., 2020). For example, we observed several highly abundant populations with a range restricted to only one or two of the three sewersheds. Going forward, it will be important to gage whether populations (or genotypes within a population) exist that are specific to individual sewersheds. Further inquiry in this direction may also lead to strategies for resolving source attribution problems when multiple collection systems with differing catchment compositions are all possible sources of contamination in the same water environment.

Our reporting for source apportioning (Fig. 4A) reports %GEQ belonging to each genome library. This metric represents an estimation of the fraction of prokaryote cells which we are confident belong to a particular source library. Yet, the values reported herein should not be interpreted as representing the fraction of total fecal material belonging to a particular source. Additionally, some signal is reported as belonging to off-target libraries (e.g., chicken) despite our efforts to eliminate cross-reactive genomic entries based on ANI comparisons *a priori*. We believe this signal likely belongs to genomes of populations with close relatives within either the background matrix (e.g., freshwater) or sewage microbial communities which have yet to be cataloged. Thus, as more genomic datasets from these environments becomes available it will be important to update these source-specific libraries to ensure better performance and less cross-reactivity.

5. Conclusions

- We tracked the microbial dynamics of a simulated sewage spill in freshwater mesocosms for 7 days using shotgun metagenomes, culture, and qPCR to better establish how shotgun metagenomics can assist with water quality monitoring efforts.
- Metagenomic analysis revealed that genes related to bacterial virulence and antimicrobial resistance were substantially enriched by the addition of sewage compared to the pristine control but became markedly depleted by the 4th day.
- Genome reconstruction and comparison to available public databases suggest that collection systems likely harbor their own (specific) microbial populations which are largely distinct from those in other environments – including the human gut.
- Genomes from publicly available datasets – including those recovered by this study – were compiled and analyzed to provide a set of source-specific and non-redundant genomic libraries.
- A reproducible bioinformatic workflow was developed, harnessing a well-defined limit of detection and the source-specific genome libraries developed herein, to perform source attribution and apportioning of fecal signal in metagenomic datasets recovered from the water environment.
- Direct detection of pathogenic bacteria remains challenging due to the large amount of sequencing effort necessary to confidently detect rare features in a community.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Cobb County Water System, Gwinnett County Department of Water Resources, and the City of Atlanta Department of Watershed Management for assistance with this work. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA. This work was supported by the US National Science Foundation, award numbers 1511825 (to J.B and K.T. K) and 1831582 (K.T.K.), US Environmental Protection Agency grant #84020301 (K.T.K.), and the US National Science Foundation Graduate Research Fellowship under grant number DGE-1650044 (to B.S.). The funding agencies had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2021.117993.

References

Ahmed, W., Hughes, B., Harwood, V.J., 2016. Current status of marker genes of bacteroides and related taxa for identifying sewage pollution in environmental waters. Water 8 (6), 231. https://doi.org/10.3390/w8060231 (Basel).

Ahmed, W., Zhang, Q., Kozak, S., Beale, D., Gyawali, P., Sadowsky, M.J., Simpson, S., 2019. Comparative decay of sewage-associated marker genes in beach water and sediment in a subtropical region. Water Res. 149, 511–521. https://doi.org/10.1016/j.watres.2018.10.088.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., Segata, N., Kyrpides, N.C., Finn, R.D., 2021. A Unified catalog of 204,938 reference genomes from the human gut microbiome. Nat. Biotechnol. 39 (1), 105–114. https://doi.org/10.1038/s41587-020-0603-3.

Ashbolt, N.J., Schoen, M.E., Soller, J.A., Roser, D.J., 2010. Predicting pathogen risks to aid beach management: the real value of quantitative microbial risk assessment (QMRA). Water Res. 44 (16), 4692–4703. https://doi.org/10.1016/j.watres.2010.06.048.

Assress, H.A., Selvarajan, R., Nyoni, H., Ntushelo, K., Mamba, B.B., Msagati, T.A.M, 2019. Diversity, co-occurrence and implications of fungal communities in wastewater treatment plants. Sci. Rep. 9 (1), 14056. https://doi.org/10.1038/s41598-019-50624-z.

Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., Lemaitre, C., 2016. Multiple comparative metagenomics using multiset k -mer counting. PeerJ Comput. Sci. 2, e94. https://doi.org/10.7717/peerj-cs.94.

Berendes, D.M., Yang, P.J., Lai, A., Hu, D., Brown, J., 2018. Estimation of global recoverable human and animal faecal biomass. Nat. Sustain. 1 (11), 679–685. https://doi.org/10.1038/s41893-018-0167-0.

Bernhard, A.E., Field, K.G., 2000. A PCR assay to discriminate human and ruminant feces on the basis of host differences in *Bacteroides-Prevotella* genes encoding 16S RRNA. Appl. Environ. Microbiol. 66 (10), 4571–4574. https://doi.org/10.1128/AEM.66.10.4571-4574.2000.

Bibby, K., Peccia, J., 2013. Identification of viral pathogen diversity in sewage sludge by metagenome analysis. Environ. Sci. Technol. 47 (4), 1945–1951. https://doi.org/10.1021/es305181x.

Boehm, A.B., Graham, K.E., Jennings, W.C., 2018. Can we swim yet? Systematic review, meta-analysis, and risk assessment of aging sewage in surface waters. Environ. Sci. Technol. 52 (17), 9634–9645. https://doi.org/10.1021/acs.est.8b01948.

Boehm, A.B., Soller, J.A., Shanks, O.C., 2015. Human-associated fecal quantitative polymerase chain reaction measurements and simulated risk of gastrointestinal illness in recreational waters contaminated with raw sewage. Environ. Sci. Technol. Lett. 2 (10), 270–275. https://doi.org/10.1021/acs.estlett.5b00219.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Boratyn, G.M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., Madden, T.L., 2019. Magic-BLAST, an accurate RNA-Seq aligner for long and short reads. BMC Bioinform. 20 (1), 405. https://doi.org/10.1186/s12859-019-2996-x.

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12 (1), 59–60. https://doi.org/10.1038/nmeth.3176.

Bushnell, B., 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner; LBNL-7065E. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).

Cai, L., Zhang, T., 2013. Detecting human bacterial pathogens in wastewater treatment plants by a high-throughput shotgun sequencing technique. Environ. Sci. Technol. 47 (10), 5433–5441. https://doi.org/10.1021/es400275r.

Campanaro, S., Treu, L., Rodriguez-R, L.M., Kovalovszki, A., Ziels, R.M., Maus, I., Zhu, X., Kougias, P.G., Basile, A., Luo, G., Schlüter, A., Konstantinidis, K.T., Angelidaki, I., 2020. New insights from the biogas microbiome by comprehensive genome-resolved metagenomics of nearly 1600 species originating from multiple anaerobic digesters. Biotechnol. Biofuels 13 (1), 25. https://doi.org/10.1186/s13068-020-01679-y.

Caro-Quintero, A., Konstantinidis, K.T., 2012. Bacterial species may exist, metagenomics reveal. Environ. Microbiol. 14 (2), 347–355. https://doi.org/10.1111/j.1462-2920.2011.02668.x.

Castro, J.C., Rodriguez-R, L.M., Harvey, W.T., Weigand, M.R., Hatt, J.K., Carter, M.Q., Konstantinidis, K.T., 2018. ImGLAD: accurate detection and quantification of target organisms in metagenomes. PeerJ 6. https://doi.org/10.7717/peerj.5882.

Chen, C., Zhou, Y., Fu, H., Xiong, X., Fang, S., Jiang, H., Wu, J., Yang, H., Gao, J., Huang, L., 2021. Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. Nat. Commun. 12 https://doi.org/10.1038/s41467-021-21295-0.

Davis, J.J., Wattam, A.R., Aziz, R.K., Brettin, T., Butler, R., Butler, R.M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E.M., Gabbard, J.L., Gerdes, S., Guard, A., Kenyon, R.W., Machi, D., Mao, C., Murphy-Olson, D., Nguyen, M., Nordberg, E.K., Olsen, G.J., Olson, R.D., Overbeek, J.C., Overbeek, R., Parrello, B., Pusch, G.D., Shukla, M., Thomas, C., VanOeffelen, M., Vonstein, V., Warren, A.S., Xia, F., Xie, D., Yoo, H., Stevens, R., 2020. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. Nucleic Acids Res. 48 (D1), D606–D612. https://doi.org/10.1093/nar/gkz943.

Devane, M.L., Moriarty, E., Weaver, L., Cookson, A., Gilpin, B., 2020. Fecal indicator bacteria from environmental sources; strategies for identification to improve water quality monitoring. Water Res. 185, 116204 https://doi.org/10.1016/j.watres.2020.116204.

Ebinger, A., Fischer, S., Höper, D., 2021. A theoretical and generalized approach for the assessment of the sample-specific limit of detection for clinical metagenomics. Comput. Struct. Biotechnol. J. 19, 732–742. https://doi.org/10.1016/j.csbj.2020.12.040.

Eisenberg, J.N.S., Bartram, J., Wade, T.J., 2016. The water quality in Rio highlights the global public health concern over untreated sewage. Environ. Health Perspect. 124 (10), A180–A181. https://doi.org/10.1289/EHP662.

Fouz, N., Pangesti, K.N.A., Yasir, M., Al-Malki, A.L., Azhar, E.I., Hill-Cawthorne, G.A., Abd El Ghany, M., 2020. The contribution of wastewater to the transmission of antimicrobial resistance in the environment: implications of mass gathering settings. Trop. Med. Infect. Dis. 5 (1) https://doi.org/10.3390/tropicalmed5010033.

Gilroy, R., Ravi, A., Getino, M., Pursley, I., Horton, D.L., Alikhan, N.F., Baker, D., Gharbi, K., Hall, N., Watson, M., Adriaenssens, E.M., Foster-Nyarko, E., Jarju, S., Secka, A., Antonio, M., Oren, A., Chaudhuri, R.R., La Ragione, R., Hildebrand, F., Pallen, M.J., 2021. Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture. PeerJ 9, e10941. https://doi.org/10.7717/peerj.10941.

Gonzalez, D., Keeling, D., Thompson, H., Larson, A., Denby, J., Curtis, K., Yetka, K., Rondini, M., Yeargan, E., Egerton, T., Barker, D., Gonzalez, R., 2020. Collection system investigation microbial source tracking (CSI-MST): applying molecular markers to identify sewer infrastructure failures. J. Microbiol. Methods 178, 106068. https://doi.org/10.1016/j.mimet.2020.106068.

Harwood, V.J., Staley, C., Badgley, B.D., Borges, K., Korajkic, A., 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. FEMS Microbiol. Rev. 38 (1), 1–40. https://doi.org/10.1111/1574-6976.12031.

Hong, P.Y., Mantilla-Calderon, D., Wang, C., 2020. Metagenomics as a tool to monitor reclaimed-water quality. Appl. Environ. Microbiol. 86 (16), e00720–e00724. https://doi.org/10.1128/AEM.00724-20. /aem/86/16/AEM.00724-20.atom.

Hull, N.M., Ling, F., Pinto, A.J., Albertsen, M., Jang, H.G., Hong, P.Y., Konstantinidis, K.T., LeChevallier, M., Colwell, R.R., Liu, W.T., 2019. Drinking water microbiome project: is it time? Trends Microbiol. 27 (8), 670–677. https://doi.org/10.1016/j.tim.2019.03.011.

Hultman, J., Tamminen, M., Pärnänen, K., Cairns, J., Karkman, A., Virta, M., 2018. Host range of antibiotic resistance genes in wastewater treatment plant influent and effluent. FEMS Microbiol. Ecol. 94 (fiy038) https://doi.org/10.1093/femsec/fiy038.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ani analysis of 90 K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 9 (1), 5114. https://doi.org/10.1038/s41467-018-07641-9.

Johnston, E.R., Kim, M., Hatt, J.K., Phillips, J.R., Yao, Q., Song, Y., Hazen, T.C., Mayes, M.A., Konstantinidis, K.T., 2019. Phosphate addition increases tropical forest soil respiration primarily by deconstraining microbial population growth. Soil Biol. Biochem. 130, 43–54. https://doi.org/10.1016/j.soilbio.2018.11.026.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7. https://doi.org/10.7717/peerj.7359.

Kessler, R., 2011. Stormwater strategies: cities prepare aging infrastructure for climate change. Environ. Health Perspect. 119 (12), a514–a519. https://doi.org/10.1289/ehp.119-a514.

Kopylova, E., Noé, L., Touzet, H., 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28 (24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611.

Korajkic, A., McMinn, B.R., Harwood, V.J., 2018. Relationships between microbial indicators and pathogens in recreational water settings. Int. J. Environ. Res. Public Health 15 (12). https://doi.org/10.3390/ijerph15122842.

Lander, E.S., Waterman, M.S., 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2 (3), 231–239. https://doi.org/10.1016/0888-7543(88)90007-9.

Li, L., Nesme, J., Quintela-Baluja, M., Balboa, S., Hashsham, S., Williams, M.R., Yu, Z., Sørensen, S.J., Graham, D.W., Romalde, J.L., Dechesne, A., Smets, B.F., 2021. Extended-spectrum β-lactamase and carbapenemase genes are substantially and sequentially reduced during conveyance and treatment of urban sewage. Environ. Sci. Technol. 55 (9), 5939–5949. https://doi.org/10.1021/acs.est.0c08548.

Lin, H., Peddada, S.D., 2020. Analysis of compositions of microbiomes with bias correction. Nat. Commun. 11 (1), 3514. https://doi.org/10.1038/s41467-020-17041-7.

Lira, F., Vaz-Moreira, I., Tamames, J., Manaia, C.M., Martínez, J.L., 2020. Metagenomic analysis of an urban resistome before and after wastewater treatment. Sci. Rep. 10 (1), 8174. https://doi.org/10.1038/s41598-020-65031-y.

Liu, B., Zheng, D., Jin, Q., Chen, L., Yang, J., 2019. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 47 (D1), D687–D692. https://doi.org/10.1093/nar/gky1080.

McGhee, J.J., Rawson, N., Bailey, B.A., Fernandez-Guerra, A., Sisk-Hackworth, L., Kelley, S.T., 2020. Meta-sourcetracker: application of Bayesian source tracking to shotgun metagenomics. PeerJ 8, e8783. https://doi.org/10.7717/peerj.8783.

McLellan, S.L., Eren, A.M., 2014. Discovering new indicators of fecal pollution. Trends Microbiol. 22 (12), 697–706. https://doi.org/10.1016/j.tim.2014.08.002.

McLellan, S.L., Huse, S.M., Mueller-Spitz, S.R., Andreishcheva, E.N., Sogin, M.L., 2010. Diversity and population structure of sewage derived microorganisms in wastewater treatment plant influent. Environ. Microbiol. 12 (2), 378–392. https://doi.org/10.1111/j.1462-2920.2009.02075.x.

McLellan, S.L., Sauer, E.P., Corsi, S.R., Bootsma, M.J., Boehm, A.B., Spencer, S.K., Borchardt, M.A., 2018. Sewage loading and microbial risk in urban waters of the great lakes. Elem. Sci. Anthr. 6 (46) https://doi.org/10.1525/elementa.301.

McLellan, S.L., Roguet, A., 2019. The unexpected habitat in sewer pipes for the propagation of microbial communities and their imprint on urban waters. Curr. Opin. Biotechnol. 57, 34–41. https://doi.org/10.1016/j.copbio.2018.12.010.

Medina, W.R.M., Eramo, A., Tu, M., Fahrenfeld, N.L., 2020. Sewer biofilm microbiome and antibiotic resistance genes as function of pipe material, source of microbes, and disinfection: field and laboratory studies. Environ. Sci.: Water Res. Technol. 6 (8), 2122–2137. https://doi.org/10.1039/D0EW00265H.

Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K., Knight, R., 2019. Establishing microbial composition measurement standards with reference frames. Nat. Commun. 10 https://doi.org/10.1038/s41467-019-10656-5.

Nayfach, S., Pollard, K.S., 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol. 16 (1), 51. https://doi.org/10.1186/s13059-015-0611-7.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T.B.K., Nielsen, T., Kirton, E., Faria, J.P., Edirisinghe, J.N., Henry, C.S., Jungbluth, S.P., Chivian, D., Dehal, P., Wood-Charlson, E.M., Arkin, A.P., Tringe, S.G., Visel, A., Woyke, T., Mouncey, N.J., Ivanova, N.N., Kyrpides, N.C., Eloe-Fadrosh, E.A., 2021. A genomic catalog of earth's microbiomes. Nat. Biotechnol. 39 (4), 499–509. https://doi.org/10.1038/s41587-020-0718-6.

Newton, R.J., McLellan, S.L., Dila, D.K., Vineis, J.H., Morrison, H.G., Eren, A.M., Sogin, M.L., 2015. Sewage reflects the microbiomes of human populations. MBio 6 (2). https://doi.org/10.1128/mBio.02574-14.

Olds, H.T., Corsi, S.R., Dila, D.K., Halmo, K.M., Bootsma, M.J., McLellan, S.L., 2018. High levels of sewage contamination released from urban areas after storm events: a quantitative survey with sewage specific bacterial indicators. PLoS Med. 15 (7), e1002614 https://doi.org/10.1371/journal.pmed.1002614.

Orellana, L.H., Rodriguez-R, L.M., Konstantinidis, K.T., 2017. ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. Nucleic. Acids. Res. 45 (3), e14. https://doi.org/10.1093/nar/gkw900.

Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28 (11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174.

Poretsky, R., Rodriguez-R, L.M., Luo, C., Tsementzi, D., Konstantinidis, K.T., 2014. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One 9 (4), e93827. https://doi.org/10.1371/journal.pone.0093827.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., Korobeynikov, A., 2020. Using SPAdes de novo assembler. Curr. Protoc. Bioinform. 70 (1) https://doi.org/10.1002/cpbi.102.

Ritalahti, K.M., Amos, B.K., Sung, Y., Wu, Q., Koenigsberg, S.S., Löffler, F.E., 2006. Quantitative PCR targeting 16S rRNA and reductive dehalogenase genes simultaneously monitors multiple *Dehalococcoides* strains. AEM 72 (4), 2765–2774. https://doi.org/10.1128/AEM.72.4.2765-2774.2006.

Rodriguez-R, L.M., Gunturu, S., Harvey, W.T., Rosselló-Mora, R., Tiedje, J.M., Cole, J.R., Konstantinidis, K.T., 2018a. The microbial genomes atlas (MiGA) webserver: taxonomic and gene diversity analysis of archaea and bacteria at the whole genome level. Nucleic Acids Res. 46, W282–W288. https://doi.org/10.1093/nar/gky467. Web Server issue.

Rodriguez-R, L.M., Gunturu, S., Tiedje, J.M., Cole, J.R., Konstantinidis, K.T., 2018b. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. mSystems 3 (3), e00018–e00039. https://doi.org/10.1128/mSystems.00039-18 msystems/3/3/msys.00039-18.atom.

Rodriguez-R, L.M., Jain, C., Conrad, R.E., Aluru, S., Konstantinidis, K.T., 2021. Reply to: "Re-evaluating the evidence for a universal genetic boundary among microbial species. Nat. Commun. 12 (1), 4060. https://doi.org/10.1038/s41467-021-24129-1.

Rodriguez-R, L.M., Konstantinidis, K.T., 2014. Estimating coverage in metagenomic data sets and why it matters. ISME J. 8 (11), 2349–2351. https://doi.org/10.1038/ismej.2014.76.

Rodriguez-R, L.M., Tsementzi, D., Luo, C., Konstantinidis, K.T., 2020. Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences. Environ. Microbiol. 22 (8), 3394–3412. https://doi.org/10.1111/1462-2920.15112.

Roguet, A., Esen, Ö.C., Eren, A.M., Newton, R.J., McLellan, S.L., 2020. FORENSIC: an online platform for fecal source identification. mSystems 5 (2). https://doi.org/10.1128/mSystems.00869-19.

Ruiz-Perez, C.A., Conrad, R.E., Konstantinidis, K.T., 2021. MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. BMC Bioinform. 22 (1), 11. https://doi.org/10.1186/s12859-020-03940-5.

Salman, B., Salem, O., 2012. Modeling failure of wastewater collection lines using various section-level regression models. J. Infrastruct. Syst. 18 (2), 146–154. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000075.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T.S., Shapiro, N., Blood, P.D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M.Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiūtė, M., Hansen, L.H., Sørensen, S.J., Chia, B.K.H., Denis, B., Froula, J.L., Wang, Z., Egan, R., Don Kang, D., Cook, J.J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.W., Singer, S.W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M.D., Lingner, T., Lin, H.H., Liao, Y.C., Silva, G.G.Z., Cuevas, D. A., Edwards, R.A., Saha, S., Piro, V.C., Renard, B.Y., Pop, M., Klenk, H.P., Göker, M., Kyrpides, N.C., Woyke, T., Vorholt, J.A., Schulze-Lefert, P., Rubin, E.M., Darling, A. E., Rattei, T., McHardy, A.C., 2017. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. Nat. Methods 14 (11), 1063–1071. https://doi.org/10.1038/nmeth.4458.

Segata, N., 2018. On the road to strain-resolved comparative metagenomics. mSystems 3 (2), e00117–e00190. https://doi.org/10.1128/mSystems.00190-17 msystems/3/2/msys.00190-17.atom.

Stewart, R.D., Auffret, M.D., Warr, A., Walker, A.W., Roehe, R., Watson, M., 2019. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat. Biotechnol. 37 (8), 953–961. https://doi.org/10.1038/s41587-019-0202-3.

Su, X., Liu, T., Beheshti, M., Prigiobbe, V., 2020. Relationship between infiltration, sewer rehabilitation, and groundwater flooding in coastal urban areas. Environ. Sci. Pollut. Res. 27 (13), 14288–14298. https://doi.org/10.1007/s11356-019-06513-z.

Suttner, B., Lindner, B.G., Kim, M., Conrad, R.E., Rodriguez, L.M., Orellana, L.H., Johnston, E.R., Hatt, J.K., Zhu, K.J., Brown, J., Konstantinidis, K.T., 2021. Metagenome-based comparisons of decay rates and host-specificity of fecal microbial communities for improved microbial source tracking. bioRxiv. https://doi.org/10.1101/2021.06.17.448865, 2021.06.17.448865.

ten Veldhuis, J.A.E., Clemens, F.H.L.R., Sterk, G., Berends, B.R., 2010. Microbial risks associated with exposure to pathogens in contaminated urban flood water. Water Res. 44 (9), 2910–2918. https://doi.org/10.1016/j.watres.2010.02.009.

Unno, T., Staley, C., Brown, C.M., Han, D., Sadowsky, M.J., Hur, H.G., 2018. Fecal Pollution: new trends and challenges in microbial source tracking using next-generation sequencing: progress and challenges in MST. Environ. Microbiol. 20 (9), 3132–3140. https://doi.org/10.1111/1462-2920.14281.

USEPA, 2009. Method 1600: Enterococci in Water by Membrane Filtration Using *Membrane-Enterococcus* Indoxyl-β-D-Glucoside Agar (MEI). United States Environmental Protection Agency.

USEPA, 2015. Recreational Water Quality Criteria. United States Environmental Protection Agency October. USEPA.

VandeWalle, J.L., Goetz, G.W., Huse, S.M., Morrison, H.G., Sogin, M.L., Hoffmann, R.G., Yan, K., McLellan, S.L., 2012. *Acinetobacter, Aeromonas and Trichococcus* populations dominate the microbial community within urban sewer infrastructure: dominant microbial populations of sewer infrastructure. Environ. Microbiol. 14 (9), 2538–2552. https://doi.org/10.1111/j.1462-2920.2012.02757.x.

Wade, T.J., Sams, E., Brenner, K.P., Haugland, R., Chern, E., Beach, M., Wymer, L., Rankin, C.C., Love, D., Li, Q., Noble, R., Dufour, A.P., 2010. Rapidly measured indicators of recreational water quality and swimming-associated illness at marine beaches: a prospective cohort study. Environ. Health 9 (1), 66. https://doi.org/10.1186/1476-069X-9-66.

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73 (16), 5261–5267. https://doi.org/10.1128/AEM.00062-07.

Weimann, A., Mooren, K., Frank, J., Pope, P.B., Bremges, A., McHardy, A.C., 2016. From genomes to phenotypes: traitar, the microbial trait analyzer. mSystems 1 (6). https://doi.org/10.1128/mSystems.00101-16.

Wendl, M.C., Kota, K., Weinstock, G.M., Mitreva, M., 2013. Coverage theories for metagenomic DNA sequencing based on a generalization of stevens' theorem. J. Math. Biol. 67 (5), 1141–1161. https://doi.org/10.1007/s00285-012-0586-x.

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20 (1), 257. https://doi.org/10.1186/s13059-019-1891-0.

Wu, Y.W., Simmons, B.A., Singer, S.W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32 (4), 605–607. https://doi.org/10.1093/bioinformatics/btv638.

Ye, L., Mei, R., Liu, W.T., Ren, H., Zhang, X.X., 2020. Machine learning-aided analyzes of thousands of draft genomes reveal specific features of activated sludge processes. Microbiome 8 (1), 16. https://doi.org/10.1186/s40168-020-0794-3.

Zhang, S.Y.; Suttner, B.; Rodriguez-R, L.; Orellana, L.; Rowell, J.; Webb, H.; Williams-Newkirk, A.; Huang, A.; Konstantinidis, K. Rocker models for reliable detection and typing of short read sequences carrying β-lactamases; preprint; Research Square, 2020. 10.21203/rs.3.rs-113339/v1.