bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Novel municipal sewage-associated bacterial genomes and their potential in source tracking

Blake G. Lindner[1], Brittany Suttner[1], Roth E. Conrad[2], Luis M. Rodriguez-R[1,3], Janet K. Hatt[1], Kevin J. Zhu[1], Joe Brown[1a], and Konstantinos T. Konstantinidis[1*]

[1] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[2] Ocean Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[3] Department of Microbiology and Digital Science Center (DiSC), University of Innsbruck, 6020 Innsbruck, Tyrol, Austria

Present address:

[a] 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

* To whom correspondence should be addressed.

Konstantinos T. Konstantinidis,

311 Ferst Drive, ES&T Building, Room 3321,

Georgia Institute of Technology.

Atlanta, GA, 30332.

Telephone: 404-639-4292

Email: kostas@ce.gatech.edu

Abstract

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Little is known about the genomic diversity of raw municipal wastewater (sewage) microbial communities, including to what extent sewage-specific populations exist and how they can be used to improve source attribution and partitioning in sewage-contaminated waters. Herein, we used the influent of three wastewater treatment plants in Atlanta, Georgia (USA) as inoculum in multiple controlled laboratory mesocosms to simulate sewage contamination events and followed these perturbed freshwater microbial communities with metagenomics over a 7-day observational period. We describe 15 abundant non-redundant bacterial metagenome-assembled genomes (MAGs) ubiquitous within all sewage inoculum yet absent from the unperturbed freshwater control at our analytical limit of detection. Tracking the dynamics of populations represented by these MAGs revealed varied decay kinetics, depending on (inferred) phenotypes, e.g., anaerobes decayed faster under the well-aerated incubation conditions. Notably, a portion of these populations show decay patterns similar to common markers, *Enterococcus* and HF183. Comparisons against MAGs from different sources such as human and animal feces, revealed low cross-reactivity, indicating how genomic collections could be used to sensitively identify sewage contamination and partition signal among multiple sources. Overall, our results indicate the usefulness of metagenomic approaches for assessing sewage contamination in waterbodies and provides needed methodologies for doing so.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(1-3). 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 (4). 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 (1-3,5). 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 (6-9) and possible dissemination of antimicrobial resistance genes (ARGs) among microbial populations (10-12).

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 (13). In large part, the technical approaches behind MST methods have been developed in response to both the difficulty of assaying for the diverse array of relevant human pathogens as well as the practical need to keep MST 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 fecal 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 (14). Some markers (e.g., HF183 assay 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

thereby, a potential array of pathogen concentrations for iterative risk assessment simulations (15). 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 (13,16-18).

Within 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 (17,19-21). 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 (22-24). However, these single-gene assays offer limited resolution to distinguish between environmental or non-environmental strains of the same species due to the high sequence 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) or resolve finer community-wide compositional shifts (17,25). 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 ~95% to ~100% average nucleotide identity (or ANI) depending on the population considered (26). Metagenomic approaches offer unique advantages for environmental health monitoring tasks including: 1) extensive gene content information of abundant populations, 2) precise

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ecological estimates of relative abundance at the species level and 3) examination of intra-species diversity (27). Despite its potential for circumventing some of the challenges facing existing MST and metabarcoding methods, whole genome shotgun sequencing has not been fully utilized in monitoring municipal sewage pollution. To date, most applications have focused on understanding the microbiology of biological wastewater treatment, treated effluents and their receiving waters, or viral populations (12,28,29). 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 (30). 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 partitioning").

Here, we offer a genome-centric view of sewage-related bacterial populations and explore their relationships with culture and PCR-based markers during a simulated failure of a collection system (e.g., an overflow event). Specifically, we simulated sewage contamination events in lake water obtained from a local drinking and recreational use reservoir, Lake Lanier (GA, USA), within dialysis bag mesocosms that were incubated in darkness 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 partitioning, and directly screen for both pathogens and antimicrobial resistance genes. Lastly, we propose a theoretical analytical limit of detection for metagenomics that could help guide the future application and interpretation of whole genome shotgun sequencing to these issues.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

Sample collection, mesocosm setup, and sample processing

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 hours. 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 x 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.

Mesocosm sampling, DNA extraction and subsequent qPCR analysis occurred as described previously in Suttner et al (35). Briefly, water samples were passed through 0.45 µm poly-carbonate (PC) membranes and stored at -80 °C in 2 mL screw cap bead tubes until

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

processed (within 1-3 months). EPA Method 1600 (31) was followed for enumerating volumetric *Enterococcus* CFUs. DNA was extracted from PC membranes using the Qiagen PowerFecal kit and 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 for qPCR with the HF183/BFDRev assay (32) and a universal 16S rRNA gene qPCR assay (GenBac16S) to quantify 16S rRNA gene copies across samples (34). 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 1 X dsDNA assay. Samples were then pooled and sequenced on the HiSeq 2500 instrument as described previously (33).

All qPCR reactions were run using an Applied Biosystems 7500Fast thermocycler and the cycling parameters were as follows: 2 min at 50 °C, 10 min at 95 °C, and 40 cycles of 15 sec at 95 °C and 60 sec 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 GenBac16S assay. Template DNAs were run undiluted or diluted 5-fold (to remove the effect of PCR inhibitors) depending 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. (35) and reiterated within the supplement here (**Supplement 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Bioinformatics sequence processing and population genome binning.

Short reads were quality trimmed and Nextera adapters removed with Trimmomatic 0.39 (36). 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 (37). 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) (38). 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 (39). 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 (40,41). Contigs shorter than 3Kbp were removed prior to population genome binning, which was performed with MaxBin 2.2.7 and MetaBAT 2.12.1 (42,43). Additionally, in a parallel workflow, trimmed short reads were normalized via the BBNorm function of the BBtools suite (version 38) to bring read depths between 10-30X sequencing depth and then subsequently assembled and binned as described above (44). 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 (45). 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

annotate draft genomes, respectively (46,47). Lastly, MAGs were screened for cross-reactivity using the FastANI tool to search for other genomes with ANI ≥ 95% across a suite of reference databases (48).

Annotation of sequence data

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) (49). Of these, 1076 genomes were recoverable from NCBI for use in this study. Abundance estimates of pathogen genomes were assessed by short read mapping with Magic-BLAST 1.4.0 (-splice F) (50). 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 (51) were used. Evaluating MAG 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 (52,53). Reads mapping to these reference sequences were selected for best bit-score alignment and subsequently filtered by ROCker v1.5.2 as described previously (54).

Estimation of limit of detection, and relative or absolute abundance

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 (55). Or, as written, the analytical limit of detection (LOD) used here:

Eqn 1: $Analytical\ LOD$: $C \geq 0.1$

The LOD was automatically implemented by calculating sequencing depth and breadth similarly to Rodriguez-R et al (56) 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 ≈ 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 (57), breadth of coverage (C) is related to sequencing depth (ρ) by the following:

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

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:

Eqn 3: $Quantification\ Threshold$: $TAD80(\rho) \geq -\ln(0.9)$

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 coverage 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 (58). 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 (59,60). 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:

Eqn 4: $16S\ rRNA\ ACN = \frac{\rho_{16S}}{GEQ}$

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:

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

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 (61,62). 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_i)], GEQ and total estimated prokaryotic cell density:

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

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:

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

Results:

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/100mL and 2.4E+06 and 3.6E+06 HF183 copies/mL, respectively. Within sewershed C, counts were lower having 1.3E+04 Enterococci CFUs/100mL 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 (**Supplement Figure 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/100mL and 41 HF183 copies/mL). Only the HF183 marker within sewershed C mesocosm decreased below detection on Day 7 (**Figure 1**). Neither marker was detected in the (un-inoculated) freshwater sample serving as control at any time point during mesocosm operation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Estimated Microbial Load

Prokaryotic cell density of the inoculum varied per mesocosm based on quantification of the 16S rRNA gene (see Methods for details): 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 hours to 1.8E+08 and 6.9E+07 estimated cells/mL (a 924% and 275% increase) while sewershed B decreased to (an estimated) 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 estimated cells/mL (**Supplement Table S2**).

Metagenomic-based Coverage and Compositional Shifts of the Mesocosms Over Time

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 to 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 (36). This level of coverage is adequate for comparing the abundance of features (e.g., genomes, genes) across samples (63). 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 hours 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

295 increased in diversity for the remaining days while both sewersheds A and C vacillate thereafter
296 (**Supplement Table S2**).

297 Observations of beta diversity revealed that the earlier timeseries samples (day 0 and day
298 1) remained quite similar to the inoculum. By day 4, considerable shifts in community
299 composition were observed driving the sewage contaminated waters closer to the control
300 (**Supplement Figure S2**). *k*-mer mapping to characterize these community-wide shifts using
301 Kraken2 at the class level showed the depletion of *Bacteroidia*, *Epsilonproteobacteria*, and
302 *Clostridia* following inoculation. None of these classes were detectable in the control samples.
303 An increase of *Gammaproteobacteria* abundance occurred within the first 24 hours across all
304 replicates after which this class gradually decreased in abundance with time. Additionally,
305 increases in *Alphaproteobacteria* and *Cytophagia* occurred in later time points (day 4 and day 7),
306 far beyond the increase observed in the control, suggesting that the later timepoint samples had
307 not yet fully recovered from perturbation. Class level relative metagenome-based abundances,
308 qPCR, culture, and cell density estimation results are summarized on **Figure 1**.
309
310 Sewage-associated Population Genome Binning
311 Seven hundred twenty MAGs were recovered from inoculum and timeseries sample
312 assemblies. The 720 MAGs were dereplicated at the ANI ≥ 95% level, resulting in 49 MAGs
313 representing sequence discrete populations (hereafter, simply "populations"). Competitive read
314 mapping to the representative MAG of these populations revealed two groupings delineated by
315 their presence or absence in the inoculum. Of the total 49, 33 populations were detected within
316 sewage inoculum samples with varying degrees of prevalence across replicates. We selected a
317 subset of 15 of these 33 populations that were above the quantification threshold in each

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 (**Supplement Table S3.A**) (64). We found our approach, as described in Methods (Eqn. 1 and 2), was robust for reducing quantification error and detected targets of known relative abundance as expected according to sequencing effort and target genome size, except on very limited occasions when close relatives were present in the sample at frequencies many times greater than the target genome. (**Supplement Table S3.B,C**).

Our collection of ubiquitous sewage-associated populations in sewersheds A, B, and C inoculum metagenomes were represented by respectively 9.5%, 5.7%, and 13.3% reads and 15.9%, 8.8%, and 19.6% of GEQ. 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,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

these results reveal that this collection of populations consistently represent highly abundant members of the sewage microbiome across biological replicates and a substantial part of the total sewage microbial community.

Comparison of the corresponding representative MAG sequences against type material in the MiGA "TypeMat" database (45) 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 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 (56), activated sludge (65), anaerobic digestors (66), the human gut environments (67), and the broad general-purpose GEMs catalog (68), 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 (**Supplement 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) (47). This finding suggests Population 29

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

363 could be less effective as an entry in a sewage-specific genomic library utilized for MST
364 approaches if other *Moraxella* are in high abundance within the pristine environment.
365
366 Sewage-associated Population Decay and Putative Phenotyping
367 Overall, all populations experienced rapid decline in estimated cell densities across the
368 timeseries with most populations below detection limits following day 4. *Acinetobacter sp.*,
369 *Cloacibacterium sp*., *Acidovorax temperans*, and *Flavobacterium sp*. (Pop.03, Pop.18, Pop.30
370 and Pop.33, respectively) were detectable in at least one biological replicate at day 7 but most of
371 these observations were below quantification. Signal from sewershed A had the greatest
372 persistence; of the four mesocosms with quantifiable levels of a sewage-associated population by
373 day 7, three belonged to the series of sewershed A. Notably, *Acidovorax temperans* (Pop.30) was
374 the only population detected at day 7 in all three sewersheds (**Figure 2**).
375 All populations remaining detectable at day 7 were putatively phenotyped as aerobic or
376 facultatively anaerobic by Traitar analysis except for *Cloacibacterium sp.* (Pop.18), which could
377 not be confidently classified. Nonetheless, *Cloacibacterium sp.* belongs to a genus of facultative
378 anaerobes (*Cloacibacterium*), suggesting that it likely is a facultative population and that the
379 representative MAG did not contain the necessary genes for confident phenotyping due to
380 incompleteness. No population – regardless of (predicted) preference for oxygen – showed an
381 increased estimated cell density outside the first 24 hours of the incubation. All sewage-
382 associated populations were likely gram negative, rod or oval-shaped bacteria as predicted by
383 Traitar (**Supplement Figure S4**).
384 MicrobeAnnotator indicated *Acinetobacter sp.* (Pop.03) and *Acidovorax temperans*
385 (Pop.30) contained modules for aromatic carbon degradation which were rare genomic features

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

among the representative MAGs. *Acinetobacter sp.* (Pop.03) was reported to contain complete benzoate degradation and catechol ortho-cleavage modules, while *Acidovorax temperans* (Pop.30) contained complete catechol ortho-cleavage and incomplete catechol meta-cleavage modules but none for benzoate degradation (**Supplement Figure S5**).

Human Markers vs Sewage-associated Populations

Our results suggested that several of the sewage-associated populations are possibly linked to the human gut microbiome (**Supplement Table S4**). Based on AAI values, Pop.43 and Pop.44 were assigned to *Bacteroidales* lineages but likely represent a different lineage than that represented by HF183 based on the 16S rRNA genes carried on these populations' closest complete genome matches from a cultured representative (See **Table 1**). Modelling 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. 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) (**Figure 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 Pop.15 ($R^2 = 0.35$).

Genome-based Source Tracking

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We aimed to demonstrate how read mapping metagenomic data to genome collections can perform differential source attribution of the fecal contamination within our single input mesocosm experiments. Our goal was to construct libraries of genomes representing the microbial pollutants expected to belong to a specific fecal source. To accomplish this, we downloaded MAGs, isolate genomes, and other reference genomes from several largescale studies of host microbiomes to create a library for determining source attribution with whole-genome sequences (67, 69-71) The collected data included genomic entries representing the fecal microbiomes of humans (n=4644), pigs (n=1667), and chickens (n=5675) and the rumen microbiome of cows (n=2124). Using MiGA, we dereplicated entries from each collection to obtain representative genomes with the highest quality for groups of genomic entries with ANI ≥ 95% to each other to reduce both the size of the dataset and redundant entries representing highly similar populations within a collection of host-associated genomes. This approach was the same as described in our methods for processing and selecting the MAGs described above and shown on **Table 1**. Using these dereplicated genomes and including the 15 representative MAGs we produced herein, we searched for and removed any instances of ANI ≥ 95% matches across these collections of host-associated genomes to control for potential cross-reactivity. One exception for which we did not remove matching genomes was between matching human and sewage genomes since they represent the same fecal source and would not complicate interpretation of results (**Supplement Table S4**). Our library construction and curation efforts are summarized in **Figure 4**.

Next, we aimed to use our host-specific source libraries to perform source attribution and partitioning as if our mesocosm data represented metagenomes recovered from a waterbody contaminated by a single unknown source. Thus, we performed competitive read mapping of the

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

metagenomic data to the finalized non-redundant genomic library using Magic-BLAST and custom scripts for TAD80 calculation as described above for tracking the sewage-associated populations individually. Resulting TAD80 values were summed within each source category and normalized to GEQ to allow interpretation of these results as the percentage of contribution from each fecal source in the form of % GEQ (e.g., percentage of prokaryotic genomes) belonging to a source (**Figure 5, A**). No source category was detected in the control samples and signal from our collection of sewage MAGs dominated the timeseries across all sewersheds but rapidly disappeared after day 4. The combined human signal followed a similar pattern as the sewage though usually at about 10% less GEQ. The pig, cow, and chicken source categories were usually not detected or were consistently <0.1% GEQ. Hence, this approach provides the means to assess contamination at the metagenomic read level, circumventing a substantial computational burden to assemble, bin and obtain de-replicated MAGs.

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 (see Methods, Eqn. 1) (**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. It should be mentioned, for many of these

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

organisms, pathogenicity is a function of 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 (55).

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 (**Figure 5, B**). 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 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 hours before substantially decreasing by day 4 and being below detection by day 7. The increase was primarily due to substantial increase in the abundance of *Aeromonas hydrophila* VFs. This trend was consistent among genes encoding for *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

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

pathogen. Hence, to what extent the MAG identified represents a pathogenic or opportunities pathogenic population remains somewhat speculative.

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) were 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 frequency of genes encoding for BlaA, BlaC and OXA decreased consistently while those encoding for MBLs increased (**Figure 5**, C). 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.

Discussion:

Sewershed Microbial Diversity

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Collection systems represent a key component of modern sanitation infrastructure. Despite the importance of sewage as a reservoir for human pathogens and antimicrobial resistance genes, the sewage microbiome remains relatively understudied at the whole genome level. Our results indicated that the sewage samples we collected from three separate collection systems across the Atlanta Metropolitan region were dominated by what have been aptly named microbial "weeds" in literature and which we have observed as belonging to several sewage-associated populations which appear quite prolific (21,22). Others have reported many of these populations are also present at high relative abundances within sewersheds spanning another urban landscape (72). Our resulting analysis expanded our understanding of these bacterial populations and their fate during a simulated contamination event by recovering representative draft genomes (MAGs) for these ubiquitous populations and tracking their abundances over time with controls in place for microbial load fluctuations. Specifically, we found primary sewage-associated populations to represent clades within the classes *Gammaproteobacteria* (e.g., *Acinetobacter*, *Aeromonas*), *Betaproteobacteria* (e.g., *Acidovorax*, *Rivicola*) and *Epsilonproteobacteria* (e.g., *Arcobacter*) (**Figure 1,A**).

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

described here, their linear relationship with HF183 and Enterococci had a combined R^2 of 0.6 (**Figure 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.

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 a more practical view of both the 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 these essential infrastructures (73). For example, we observed several highly abundant populations with a range restricted to only one or two of the three sewersheds. It would be important to gauge whether populations (or genotypes within a population) exist that are specific to individual sewersheds, and how the physicochemical characteristics (e.g., municipal vs. industrial waste, flow rates) of different waste streams might drive the formation of these distinctions. Further inquiry in this direction may also lead to strategies for resolving source attribution inquiries when multiple collection systems with differing catchment compositions are all possible sources of contamination in the same water environment.

Source Attribution and Partitioning with Host-Specific Genomic Libraries

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Several of the abundant sewage-associated populations identified above appear closely related (likely at the species level) to members of human or chicken fecal microbiomes (**Figure 4**). Regardless, it appears populations specific to municipal sewage likely exist and represent a contingent of the sewage microbiome which – if better catalogued – may be useful for identifying and quantifying sewage pollution in natural ecosystems independent of human-associated markers. We have demonstrated, through a proof-of-concept workflow, the capacity for read mapping metagenomic datasets to host-specific genomic libraries for performing both source attribution and partitioning (**Figure 5, A**). Although source attribution has been well-developed in existing metagenomic approaches, no metagenomic methods have been developed which are also capable of simultaneous source partitioning. Here, we have performed source partitioning for each entry within a host-specific genomic library to a metagenome's GEQ, which yields a relatively easy-to-interpret metric describing the percentage of genomes within a metagenome that belong to a given host or source specific library. We see this as a promising avenue for metagenomic-based MST methods and believe the approach could eventually be utilized in the field pending further testing and refinement by mixed input experiments.

β-lactamase Encoding Genes Surveillance

Additionally, we leveraged our metagenomes to survey for β-lactamase 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 (**Figure 5**). This result was consistent with both our expectations and the literature regarding heightened ARG abundance within collection systems (74). Specifically, others have reported substantial abundances of β -lactamase OXA genes on both *Campylobacteraceae* and *Aeromonadaceae* clades in sewage (75). Indeed, the abundance of reads belonging to β-lactamase encoding genes, especially of the OXA-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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. 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 where the freshwater used in the mesocosm incubations originated, which is relevant for assessing the associated public health risk.

Pathogen and Virulence Gene Surveillance

Importantly, although Sewershed A and B showed what appears to be similar concentrations of human input according to HF183 concentrations within the inoculum (**Supplement Figure S1**), the pathogen detection results revealed via the sequence data were quite varied (**Figure 4B**, **Supplement Table S6**). Results from both read mapping to bacterial pathogen genomes and the experimentally verified VFDB collection were consistent in suggesting that bacterial virulence was 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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 (Methods, **Supplement 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 (76,77). 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 (i.e., <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".

Using AGS and total cell density estimates within the inoculum, we estimate approximately 3.5Tb of sequencing effort is necessary for detecting a population with concentration of 1E+02 cells/mL within the high microbial loading conditions observed in the inoculum. 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 10Gb in day 7 conditions. 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 (**Supplement 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

615 recoveries – where the latter limitation is obviously broadly applicable to all molecular methods.
616 Total detection limits, in the context of analytical limits as well as both sequencing bias and
617 sampling/processing recoveries, will be important caveats to consider for future metagenomic
618 workflows aiming to surveil pathogens in sewage collection systems or their releases into the
619 environment (78).

620 While we do not envision that PCR and culturing will be replaced by metagenomics for
621 routine monitoring because the former techniques are cheaper, easier to analyze, and have a
622 greater dynamic range of detection – our results do show how metagenomics can provide unique
623 insights into sewage pollution events such as differential pathogen content of different sewage or
624 possibly distinguishing between different sewersheds potentially contributing to contamination.
625 Further, we have shown how metagenomics could track a broad range of population sizes –
626 about six orders of magnitude (from about 1E+01 to 1E+07 cells/mL) – which is adequate for
627 certain applications and/or high-volume pollution events.

628

629 Conclusions

630 Our efforts have shown how metagenomic datasets can provide insights on multiple
631 questions critical to environmental monitoring and water quality: pathogen detection, source
632 attribution and partitioning, and ARG persistence in the environment. In our view, confident and
633 direct detection of pathogens within metagenomic datasets will remain primarily a logistical
634 challenge due to the large amount of sequencing effort required to reliably detect bacterial
635 pathogens at concentrations that are very low yet still quite relevant to protecting public health.
636 Thus, when performed alone, metagenomic approaches are unlikely to be the most prudent
637 technology for routine monitoring and directly informing health risks associated with sewage

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

638 contamination, especially when pathogen or virulence genes are at these relatively low
639 abundances (e.g., below 1E+02 features/mL). This issue is also compounded by the large
640 contribution of non-bacterial pathogens (e.g., viruses and protozoa) to illness risk in
641 contaminated waters. In contrast, metagenomic approaches are increasingly poised to resolve
642 questions related to source attribution and partitioning by improving our understanding (and the
643 size of our databases) of the genomes maintained by source-specific microbial populations.
644

645 **Acknowledgments**:
646 The authors would like to thank the Cobb County Water System, Gwinnett County Department
647 of Water Resources, and the City of Atlanta Department of Watershed Management for
648 assistance with this work. This research was supported in part through research
649 cyberinfrastructure resources and services provided by the Partnership for an Advanced
650 Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.
651 This work was supported by the US National Science Foundation, award numbers 1511825 (to
652 J.B and K.T.K) and 1831582 (K.T.K.) and the US National Science Foundation Graduate
653 Research Fellowship under grant number DGE-1650044 (to B.S.). The funding agencies had no
654 role in the study design, data collection and analysis, decision to publish, or preparation of the
655 manuscript.
656

657 **Conflict of interest:** The authors declare no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

(1) Salman, B.; Salem, O. Modeling Failure of Wastewater Collection Lines Using Various Section-Level Regression Models. *Journal of Infrastructure Systems* **2012**, *18* (2), 146–154. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000075.

(2) Berendes, D. M.; Yang, P. J.; Lai, A.; Hu, D.; Brown, J. Estimation of Global Recoverable Human and Animal Faecal Biomass. *Nature Sustainability* **2018**, *1* (11), 679–685. https://doi.org/10.1038/s41893-018-0167-0.

(3) McLellan, S. L.; Sauer, E. P.; Corsi, S. R.; Bootsma, M. J.; Boehm, A. B.; Spencer, S. K.; Borchardt, M. A. Sewage Loading and Microbial Risk in Urban Waters of the Great Lakes. *Elementa: Science of the Anthropocene* **2018**, *6* (46). https://doi.org/10.1525/elementa.301.

(4) ten Veldhuis, J. A. E.; Clemens, F. H. L. R.; Sterk, G.; Berends, B. R. Microbial Risks Associated with Exposure to Pathogens in Contaminated Urban Flood Water. *Water Research* **2010**, *44* (9), 2910–2918. https://doi.org/10.1016/j.watres.2010.02.009.

(5) Olds, H. T.; Corsi, S. R.; Dila, D. K.; Halmo, K. M.; Bootsma, M. J.; McLellan, S. L. High Levels of Sewage Contamination Released from Urban Areas after Storm Events: A Quantitative Survey with Sewage Specific Bacterial Indicators. *PLOS Medicine* **2018**, *15* (7), e1002614. https://doi.org/10.1371/journal.pmed.1002614.

(6) Ashbolt, N. J.; Schoen, M. E.; Soller, J. A.; Roser, D. J. Predicting Pathogen Risks to Aid Beach Management: The Real Value of Quantitative Microbial Risk Assessment (QMRA). *Water Research* **2010**, *44* (16), 4692–4703. https://doi.org/10.1016/j.watres.2010.06.048.

(7) Fouz, N.; Pangesti, K. N. A.; Yasir, M.; Al-Malki, A. L.; Azhar, E. I.; Hill-Cawthorne, G. A.; Abd El Ghany, M. The Contribution of Wastewater to the Transmission of Antimicrobial Resistance in

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the Environment: Implications of Mass Gathering Settings. *Trop Med Infect Dis* **2020**, *5* (1). https://doi.org/10.3390/tropicalmed5010033.

(8) Medina, W. R. M.; Eramo, A.; Tu, M.; Fahrenfeld, N. L. 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.* **2020**, *6* (8), 2122–2137. https://doi.org/10.1039/D0EW00265H.

(9) Eisenberg, J. N. S.; Bartram, J.; Wade, T. J. The Water Quality in Rio Highlights the Global Public Health Concern Over Untreated Sewage. *Environ Health Perspect* **2016**, *124* (10), A180–A181. https://doi.org/10.1289/EHP662.

(10) Su, X.; Liu, T.; Beheshti, M.; Prigiobbe, V. Relationship between Infiltration, Sewer Rehabilitation, and Groundwater Flooding in Coastal Urban Areas. *Environ Sci Pollut Res* **2020**, *27* (13), 14288–14298. https://doi.org/10.1007/s11356-019-06513-z.

(11) Kessler, R. Stormwater Strategies: Cities Prepare Aging Infrastructure for Climate Change. *Environ Health Perspect* **2011**, *119* (12), a514–a519. https://doi.org/10.1289/ehp.119-a514.

(12) Lira, F.; Vaz-Moreira, I.; Tamames, J.; Manaia, C. M.; Martínez, J. L. Metagenomic Analysis of an Urban Resistome before and after Wastewater Treatment. *Scientific Reports* **2020**, *10* (1), 8174. https://doi.org/10.1038/s41598-020-65031-y.

(13) Harwood, V. J.; Staley, C.; Badgley, B. D.; Borges, K.; Korajkic, A. Microbial Source Tracking Markers for Detection of Fecal Contamination in Environmental Waters: Relationships between Pathogens and Human Health Outcomes. *FEMS Microbiol Rev* **2014**, *38* (1), 1–40. https://doi.org/10.1111/1574-6976.12031.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(14) Bernhard, A. E.; Field, K. G. 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.* **2000**, *66* (10), 4571–4574. https://doi.org/10.1128/AEM.66.10.4571-4574.2000.

(15) Boehm, A. B.; Soller, J. A.; Shanks, O. C. 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.* **2015**, *2* (10), 270–275. https://doi.org/10.1021/acs.estlett.5b00219.

(16) Korajkic, A.; McMinn, B. R.; Harwood, V. J. Relationships between Microbial Indicators and Pathogens in Recreational Water Settings. *Int J Environ Res Public Health* **2018**, *15* (12). https://doi.org/10.3390/ijerph15122842.

(17) Ahmed, W.; Hughes, B.; Harwood, V. J. Current Status of Marker Genes of Bacteroides and Related Taxa for Identifying Sewage Pollution in Environmental Waters. *Water* **2016**, *8* (6), 231. https://doi.org/10.3390/w8060231.

(18) Fecal Indicator Bacteria from Environmental Sources; Strategies for Identification to Improve Water Quality Monitoring. Water Research 2020, 185, 116204. https://doi.org/10.1016/j.watres.2020.116204.

(19) Unno, T.; Staley, C.; Brown, C. M.; Han, D.; Sadowsky, M. J.; Hur, H.-G. Fecal Pollution: New Trends and Challenges in Microbial Source Tracking Using next-Generation Sequencing: Progress and Challenges in MST. *Environ Microbiol* **2018**, *20* (9), 3132–3140. https://doi.org/10.1111/1462-2920.14281.

(20) McLellan, S. L.; Eren, A. M. Discovering New Indicators of Fecal Pollution. *Trends in Microbiology* **2014**, *22* (12), 697–706. https://doi.org/10.1016/j.tim.2014.08.002.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(21) Assress, H. A.; Selvarajan, R.; Nyoni, H.; Ntushelo, K.; Mamba, B. B.; Msagati, T. A. M. Diversity, Co-Occurrence and Implications of Fungal Communities in Wastewater Treatment Plants. *Scientific Reports* **2019**, *9* (1), 14056. https://doi.org/10.1038/s41598-019-50624-z.

(22) Newton, R. J.; McLellan, S. L.; Dila, D. K.; Vineis, J. H.; Morrison, H. G.; Eren, A. M.; Sogin, M. L. Sewage Reflects the Microbiomes of Human Populations. *mBio* **2015**, *6* (2). https://doi.org/10.1128/mBio.02574-14.

(23) McLellan, S.L.; Roguet, A. The Unexpected Habitat in Sewer Pipes for the Propagation of Microbial Communities and Their Imprint on Urban Waters. *Current Opinion in Biotechnology* **2019**, *57*, 34–41. https://doi.org/10.1016/j.copbio.2018.12.010.

(24) McLellan, S. L.; Huse, S. M.; Mueller-Spitz, S. R.; Andreishcheva, E. N.; Sogin, M. L. Diversity and Population Structure of Sewage Derived Microorganisms in Wastewater Treatment Plant Influent. *Environ Microbiol* **2010**, *12* (2), 378–392. https://doi.org/10.1111/j.1462-2920.2009.02075.x.

(25) Poretsky, R.; Rodriguez-R, L. M.; Luo, C.; Tsementzi, D.; Konstantinidis, K. T. Strengths and Limitations of 16S RRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLOS ONE* **2014**, *9* (4), e93827. https://doi.org/10.1371/journal.pone.0093827.

(26) Caro-Quintero, A.; Konstantinidis, K. T. Bacterial Species May Exist, Metagenomics Reveal. Environmental Microbiology 2012, 14 (2), 347–355. https://doi.org/https://doi.org/10.1111/j.1462-2920.2011.02668.x.

(27) Segata, N. On the Road to Strain-Resolved Comparative Metagenomics. *mSystems* **2018**, *3* (2), e00190-17, /msystems/3/2/msys.00190-17.atom. https://doi.org/10.1128/mSystems.00190-17.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(28) Cai, L.; Zhang, T. Detecting Human Bacterial Pathogens in Wastewater Treatment Plants by a High-Throughput Shotgun Sequencing Technique. *Environ. Sci. Technol.* **2013**, *47* (10), 5433–5441. https://doi.org/10.1021/es400275r.

(29) Bibby, K.; Peccia, J. Identification of Viral Pathogen Diversity in Sewage Sludge by Metagenome Analysis. *Environ. Sci. Technol.* **2013**, *47* (4), 1945–1951. https://doi.org/10.1021/es305181x.

(30) Hong, P.-Y.; Mantilla-Calderon, D.; Wang, C. Metagenomics as a Tool To Monitor Reclaimed-Water Quality. *Appl Environ Microbiol* **2020**, *86* (16), e00724-20, /aem/86/16/AEM.00724-20.atom. https://doi.org/10.1128/AEM.00724-20.

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

(32) 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. Rapidly Measured Indicators of Recreational Water Quality and Swimming-Associated Illness at Marine Beaches: A Prospective Cohort Study. *Environmental Health* **2010**, *9* (1), 66. https://doi.org/10.1186/1476-069X-9-66.

(33) Johnston, E. R.; Kim, M.; Hatt, J. K.; Phillips, J. R.; Yao, Q.; Song, Y.; Hazen, T. C.; Mayes, M. A.; Konstantinidis, K. T. Phosphate Addition Increases Tropical Forest Soil Respiration Primarily by Deconstraining Microbial Population Growth. *Soil Biology and Biochemistry* **2019**, *130*, 43–54. https://doi.org/10.1016/j.soilbio.2018.11.026.

(34) Ritalahti, K. M.; Amos, B. K.; Sung, Y.; Wu, Q.; Koenigsberg, S. S.; Löffler, F. E. Quantitative PCR Targeting 16S rRNA and Reductive Dehalogenase Genes Simultaneously Monitors Multiple Dehalococcoides Strains. *AEM* **2006**, *72* (4), 2765–2774. https://doi.org/10.1128/AEM.72.4.2765-2774.2006.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(35) 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. Metagenome-Based Comparisons of Decay Rates and Host-Specificity of Fecal Microbial Communities for Improved Microbial Source Tracking. bioRxiv 2021, 2021.06.17.448865. https://doi.org/10.1101/2021.06.17.448865.

(36) Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30* (15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

(37) Rodriguez-R, L. M.; Gunturu, S.; Tiedje, J. M.; Cole, J. R.; Konstantinidis, K. T. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. *mSystems* **2018**, *3* (3), e00039-18, /msystems/3/3/msys.00039-18.atom. https://doi.org/10.1128/mSystems.00039-18.

(38) Benoit, G.; Peterlongo, P.; Mariadassou, M.; Drezen, E.; Schbath, S.; Lavenier, D.; Lemaitre, C. Multiple Comparative Metagenomics Using Multiset *k* -Mer Counting. *PeerJ Computer Science* **2016**, *2*, e94. https://doi.org/10.7717/peerj-cs.94.

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

(40) Peng, Y.; Leung, H. C. M.; Yiu, S. M.; Chin, F. Y. L. IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth. *Bioinformatics* **2012**, *28* (11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174.

(41) Prjibelski, A.; Antipov, D.; Meleshko, D.; Lapidus, A.; Korobeynikov, A. Using SPAdes De Novo Assembler. *Current Protocols in Bioinformatics* **2020**, *70* (1). https://doi.org/10.1002/cpbi.102.

(42) Wu, Y.-W.; Simmons, B. A.; Singer, S. W. MaxBin 2.0: An Automated Binning Algorithm to Recover Genomes from Multiple Metagenomic Datasets. *Bioinformatics* **2016**, *32* (4), 605–607. https://doi.org/10.1093/bioinformatics/btv638.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(43) Kang, D. D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies. *PeerJ* **2019**, *7*. https://doi.org/10.7717/peerj.7359.

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

(45) Rodriguez-R, L. M.; Gunturu, S.; Harvey, W. T.; Rosselló-Mora, R.; Tiedje, J. M.; Cole, J. R.; Konstantinidis, K. T. The Microbial Genomes Atlas (MiGA) Webserver: Taxonomic and Gene Diversity Analysis of Archaea and Bacteria at the Whole Genome Level. *Nucleic Acids Res* **2018**, *46* (Web Server issue), W282–W288. https://doi.org/10.1093/nar/gky467.

(46) Weimann, A.; Mooren, K.; Frank, J.; Pope, P. B.; Bremges, A.; McHardy, A. C. From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. *mSystems* **2016**, *1* (6). https://doi.org/10.1128/mSystems.00101-16.

(47) Ruiz-Perez, C. A.; Conrad, R. E.; Konstantinidis, K. T. MicrobeAnnotator: A User-Friendly, Comprehensive Functional Annotation Pipeline for Microbial Genomes. *BMC Bioinformatics* **2021**, *22* (1), 11. https://doi.org/10.1186/s12859-020-03940-5.

(48) Jain, C.; Rodriguez-R, L. M.; Phillippy, A. M.; Konstantinidis, K. T.; Aluru, S. High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *Nature Communications* **2018**, *9* (1), 5114. https://doi.org/10.1038/s41467-018-07641-9.

(49) 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. The

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

PATRIC Bioinformatics Resource Center: Expanding Data and Analysis Capabilities. *Nucleic Acids Research* **2020**, *48* (D1), D606–D612. https://doi.org/10.1093/nar/gkz943.

(50) Boratyn, G. M.; Thierry-Mieg, J.; Thierry-Mieg, D.; Busby, B.; Madden, T. L. Magic-BLAST, an Accurate RNA-Seq Aligner for Long and Short Reads. *BMC Bioinformatics* **2019**, *20* (1), 405. https://doi.org/10.1186/s12859-019-2996-x.

(51) Liu, B.; Zheng, D.; Jin, Q.; Chen, L.; Yang, J. VFDB 2019: A Comparative Pathogenomic Platform with an Interactive Web Interface. Nucleic Acids Res 2019, 47 (D1), D687–D692. https://doi.org/10.1093/nar/gky1080.

(52) Buchfink, B.; Xie, C.; Huson, D. H. Fast and Sensitive Protein Alignment Using DIAMOND. *Nat Methods* **2015**, *12* (1), 59–60. https://doi.org/10.1038/nmeth.3176.

(53) 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; In Review, 2020. https://doi.org/10.21203/rs.3.rs-113339/v1.

(54) Orellana, L. H.; Rodriguez-R, L. M.; Konstantinidis, K. T. ROCker: Accurate Detection and Quantification of Target Genes in Short-Read Metagenomic Data Sets by Modeling Sliding-Window Bitscores. *Nucleic Acids Research* **2017**, *45* (3), e14–e14. https://doi.org/10.1093/nar/gkw900.

(55) Castro, J. C.; Rodriguez-R, L. M.; Harvey, W. T.; Weigand, M. R.; Hatt, J. K.; Carter, M. Q.; Konstantinidis, K. T. ImGLAD: Accurate Detection and Quantification of Target Organisms in Metagenomes. *PeerJ* **2018**, *6*. https://doi.org/10.7717/peerj.5882.

(56) Rodriguez-R, L. M.; Tsementzi, D.; Luo, C.; Konstantinidis, K. T. Iterative Subtractive Binning of Freshwater Chronoseries Metagenomes Identifies over 400 Novel Species and Their Ecologic

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Preferences. *Environ Microbiol* **2020**, *22* (8), 3394–3412. https://doi.org/10.1111/1462-2920.15112.

(57) Lander, E. S.; Waterman, M. S. Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis. *Genomics* **1988**, *2* (3), 231–239. https://doi.org/10.1016/0888-7543(88)90007-9.

(58) Nayfach, S.; Pollard, K. S. Average Genome Size Estimation Improves Comparative Metagenomics and Sheds Light on the Functional Ecology of the Human Microbiome. *Genome Biology* **2015**, *16* (1), 51. https://doi.org/10.1186/s13059-015-0611-7.

(59) Kopylova, E.; Noé, L.; Touzet, H. SortMeRNA: Fast and Accurate Filtering of Ribosomal RNAs in Metatranscriptomic Data. *Bioinformatics* **2012**, *28* (24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611.

(60) Wang, Q.; Garrity, G. M.; Tiedje, J. M.; Cole, J. R. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol* **2007**, *73* (16), 5261–5267. https://doi.org/10.1128/AEM.00062-07.

(61) Lin, H.; Peddada, S. D. Analysis of Compositions of Microbiomes with Bias Correction. *Nature Communications* **2020**, *11* (1), 3514. https://doi.org/10.1038/s41467-020-17041-7.

(62) Morton, J. T.; Marotz, C.; Washburne, A.; Silverman, J.; Zaramela, L. S.; Edlund, A.; Zengler, K.; Knight, R. Establishing Microbial Composition Measurement Standards with Reference Frames. *Nat Commun* **2019**, *10*. https://doi.org/10.1038/s41467-019-10656-5.

(63) Rodriguez-R, L. M.; Konstantinidis, K. T. Estimating Coverage in Metagenomic Data Sets and Why It Matters. *The ISME Journal* **2014**, *8* (11), 2349–2351. https://doi.org/10.1038/ismej.2014.76.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(64) 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. Critical Assessment of Metagenome Interpretation—a Benchmark of Metagenomics Software. Nature Methods 2017, 14 (11), 1063–1071. https://doi.org/10.1038/nmeth.4458.

(65) Ye, L.; Mei, R.; Liu, W.-T.; Ren, H.; Zhang, X.-X. Machine Learning-Aided Analyses of Thousands of Draft Genomes Reveal Specific Features of Activated Sludge Processes. *Microbiome* **2020**, *8* (1), 16. https://doi.org/10.1186/s40168-020-0794-3

(66) 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. New Insights from the Biogas Microbiome by Comprehensive Genome-Resolved Metagenomics of Nearly 1600 Species Originating from Multiple Anaerobic Digesters. *Biotechnology for Biofuels* **2020**, *13* (1), 25. https://doi.org/10.1186/s13068-020-01679-y.

(67) 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. A Unified

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Catalog of 204,938 Reference Genomes from the Human Gut Microbiome. *Nat Biotechnol* **2021**, *39* (1), 105–114. https://doi.org/10.1038/s41587-020-0603-3.

(68) 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. A Genomic Catalog of Earth's Microbiomes. *Nature Biotechnology* **2021**, *39* (4), 499–509. https://doi.org/10.1038/s41587-020-0718-6.

(69) Stewart, R. D.; Auffret, M. D.; Warr, A.; Walker, A. W.; Roehe, R.; Watson, M. Compendium of 4,941 Rumen Metagenome-Assembled Genomes for Rumen Microbiome Biology and Enzyme Discovery. Nature Biotechnology 2019, 37 (8), 953–961. https://doi.org/10.1038/s41587-019-0202-3.

(70) 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. Extensive Microbial Diversity within the Chicken Gut Microbiome Revealed by Metagenomics and Culture. PeerJ 2021, 9, e10941. https://doi.org/10.7717/peerj.10941.

(71) Chen, C.; Zhou, Y.; Fu, H.; Xiong, X.; Fang, S.; Jiang, H.; Wu, J.; Yang, H.; Gao, J.; Huang, L. Expanded Catalog of Microbial Genes and Metagenome-Assembled Genomes from the Pig Gut Microbiome. Nat Commun 2021, 12. https://doi.org/10.1038/s41467-021-21295-0.

(72) VandeWalle, J. L.; Goetz, G. W.; Huse, S. M.; Morrison, H. G.; Sogin, M. L.; Hoffmann, R. G.; Yan, K.; McLellan, S. L. Acinetobacter, Aeromonas and Trichococcus Populations Dominate the

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Microbial Community within Urban Sewer Infrastructure: Dominant Microbial Populations of Sewer Infrastructure. *Environmental Microbiology* **2012**, *14* (9), 2538–2552. https://doi.org/10.1111/j.1462-2920.2012.02757.x.

(73) Collection System Investigation Microbial Source Tracking (CSI-MST): Applying Molecular Markers to Identify Sewer Infrastructure Failures. Journal of Microbiological Methods 2020, 178, 106068. https://doi.org/10.1016/j.mimet.2020.106068.

(74) 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. Extended-Spectrum β-Lactamase and Carbapenemase Genes Are Substantially and Sequentially Reduced during Conveyance and Treatment of Urban Sewage. Environ. Sci. Technol. 2021, 55 (9), 5939–5949. https://doi.org/10.1021/acs.est.0c08548.

(75) Hultman, J.; Tamminen, M.; Pärnänen, K.; Cairns, J.; Karkman, A.; Virta, M. Host Range of Antibiotic Resistance Genes in Wastewater Treatment Plant Influent and Effluent. *FEMS Microbiology Ecology* **2018**, *94* (fiy038). https://doi.org/10.1093/femsec/fiy038.

(76) Wendl, M. C.; Kota, K.; Weinstock, G. M.; Mitreva, M. Coverage Theories for Metagenomic DNA Sequencing Based on a Generalization of Stevens' Theorem. J Math Biol 2013, 67 (5), 1141–1161. https://doi.org/10.1007/s00285-012-0586-x.

(77) Ebinger, A.; Fischer, S.; Höper, D. A Theoretical and Generalized Approach for the Assessment of the Sample-Specific Limit of Detection for Clinical Metagenomics. Computational and Structural Biotechnology Journal 2021, 19, 732–742. https://doi.org/10.1016/j.csbj.2020.12.040.

(78) 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. Drinking Water Microbiome Project: Is It Time? *Trends Microbiol* **2019**, *27* (8), 670–677. https://doi.org/10.1016/j.tim.2019.03.011.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

	Taxonomic Summary				Quality Summary					
Population	Confident Taxonomy (p<0.05)	Best match in MiGA TypeMat Database	Similarity (%)	Metric	Complete-ness (%)	Redund-ancy (%)	Length (Mbp)	N50 (bp)	CDs	GC (%)
01	Genus: *Arcobacter*	*Arcobacter cryaerophilus GCA 002992955*	92.8	ANI	87.7	1.9	1.38	7,214	1,519	28.77
03	Genus: *Acinetobacter*	*Acinetobacter johnsonii NZ CP065666*	96.5	ANI	78.3	0	1.99	6,506	2,157	41.90
04	Genus: *Aeromonas*	*Aeromonas caviae GCA 000819785*	93.5	ANI	56.6	0.9	2.99	6,625	3,099	61.78
13	Class: *Bacteroidia*	*Paludibacter propionicigenes WB4 NC 014734*	55.1	AAI	51.9	1.9	0.80	4,680	764	39.63
15	Species: *A. caviae*	*Aeromonas caviae GCA 000820265*	98.0	ANI	40.6	0	1.57	4,876	1,684	61.79
18	Genus: *Cloacibacterium*	*Cloacibacterium rupense GCA 014645495*	88.2	ANI	61.3	3.8	1.58	5,371	1,596	33.27
19	Family: *Campylobacteraceae*	*Arcobacter suis CECT 7833 NZ CP032100*	72.1	AAI	49.1	0	0.95	5,098	1,130	28.60
28	Order: *Neisseriales*	*Rivicola pingtungensis GCA 003201855*	67.3	AAI	75.5	0	1.13	5,350	1,176	56.67
29	Genus: *Moraxella*	*Moraxella osloensis GCA 001679175*	95.4	ANI	75.5	0	1.83	9,146	1,726	44.48
30	Species: *A. temperans*	*Acidovorax temperans GCA 006716905*	97.3	ANI	91.5	0.9	2.80	8,597	2,816	63.59
33	Genus: *Flavobacterium*	*Flavobacterium succinicans LMG 10402 GCA 000611675*	87.3	ANI	88.7	2.8	2.81	10,562	2,699	35.43
43	Species: *P. copri*	*Prevotella copri DSM 18205 GCA 009495405*	97.1	ANI	52.8	0	2.36	11,303	1,981	46.62
44	Species: *B. vulgatus*	*Bacteroides vulgatus ATCC 8482 NC 009614*	99.0	ANI	49.1	0	2.67	5,144	2,496	41.90
47	Family: *Aeromonadaceae*	*Tolumonas auensis DSM 9187 NC 012691*	83.5	ANI	98.1	1.9	2.67	16,590	2,612	47.97
49	Species: *R. pingtungensis*	*Rivicola pingtungensis GCA 003201855*	97.5	ANI	46.2	0.9	2.03	8,236	2,031	62.89

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

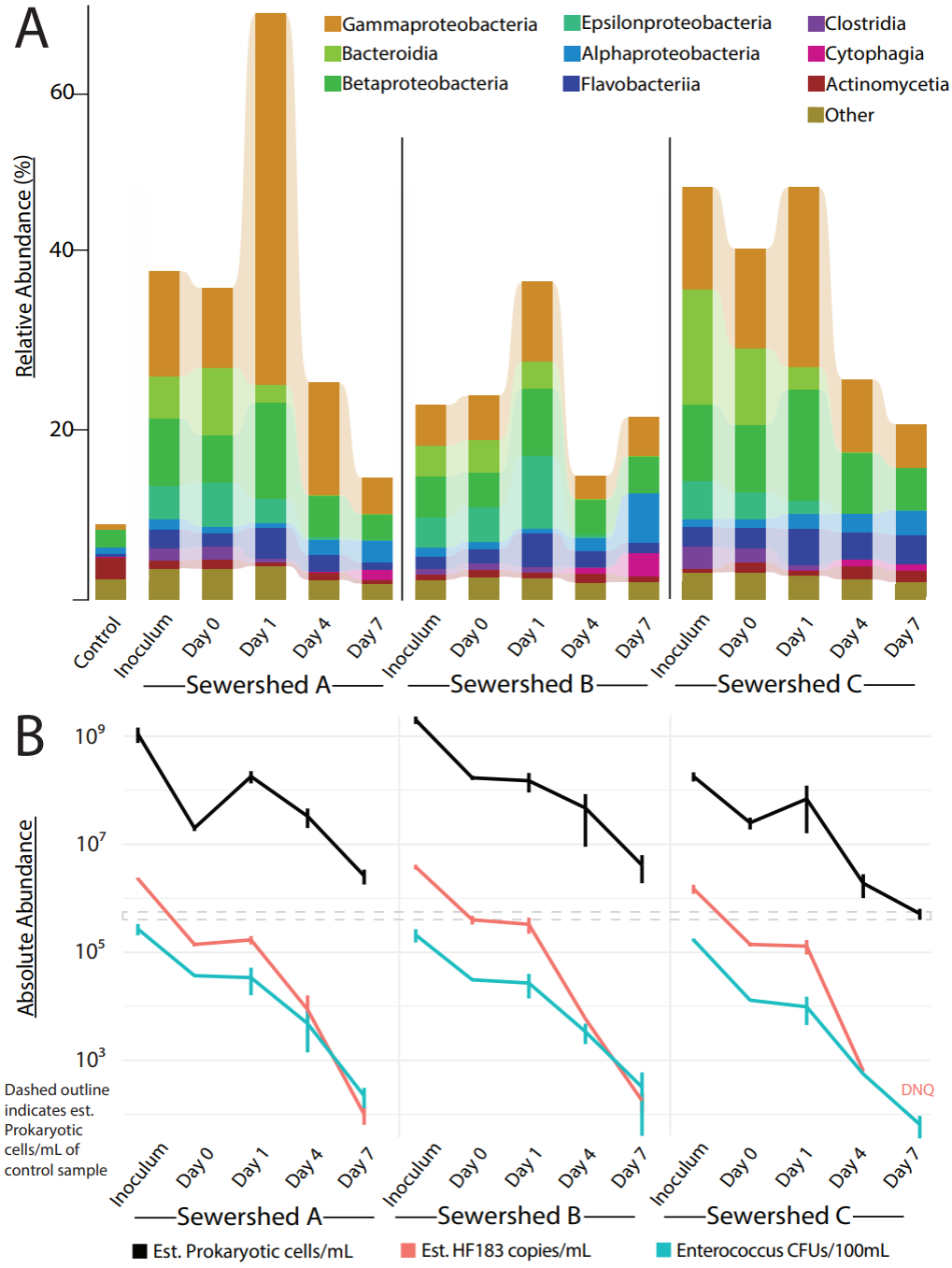


Figure 1. Panel A: Class level abundances across control, inoculum and timeseries for sewersheds A, B and C based on kmer classification by Kraken2 against a custom-built database of reference genomes. Total height of bars represents the percentage of kmers confidently classified to the corresponding 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

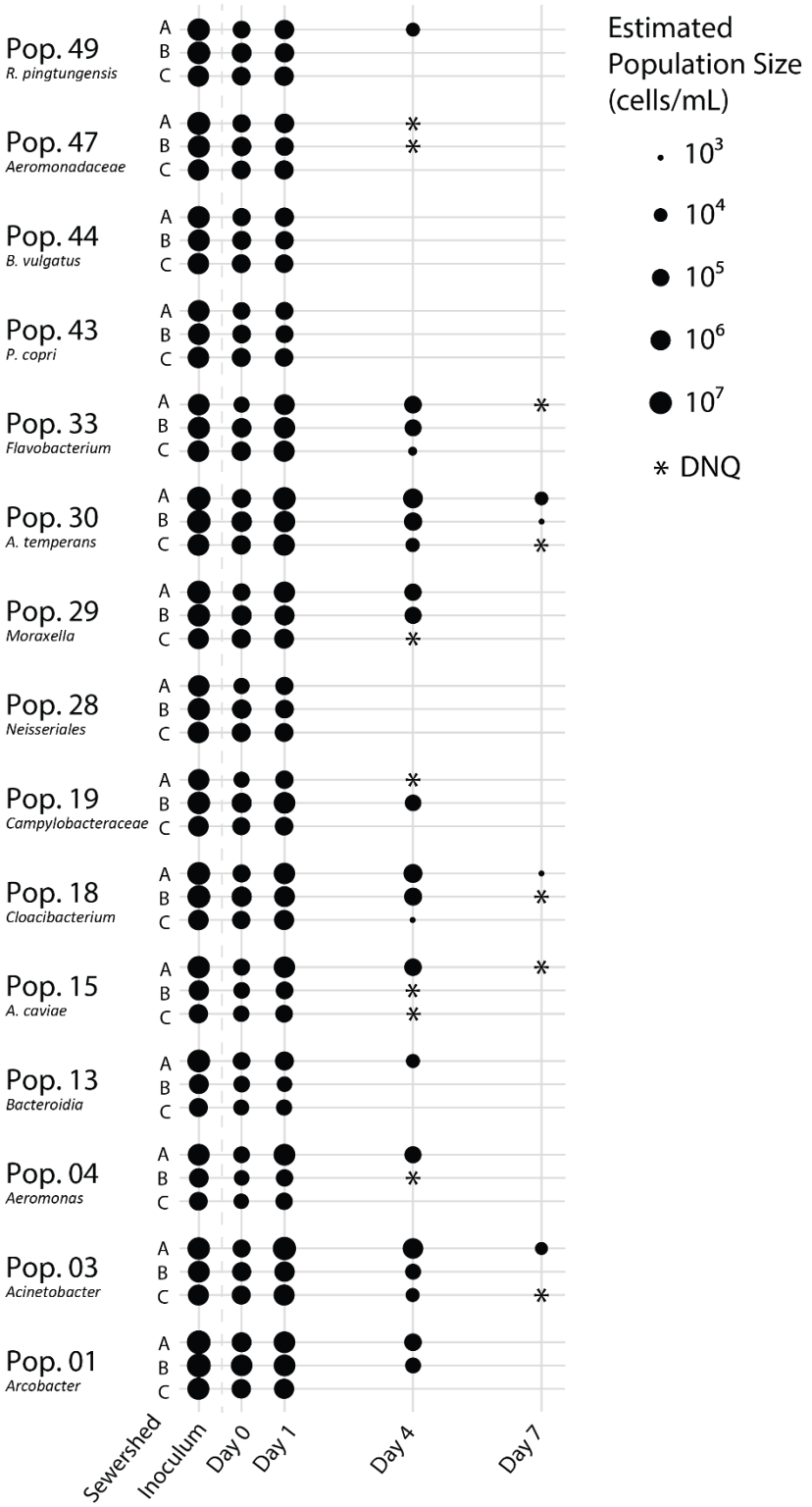


Figure 2. Estimated cell densities of sewage-associated populations across inoculum and timeseries samples. Cell densities (absolute abundances) were estimated as described in the Materials and Methods section.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

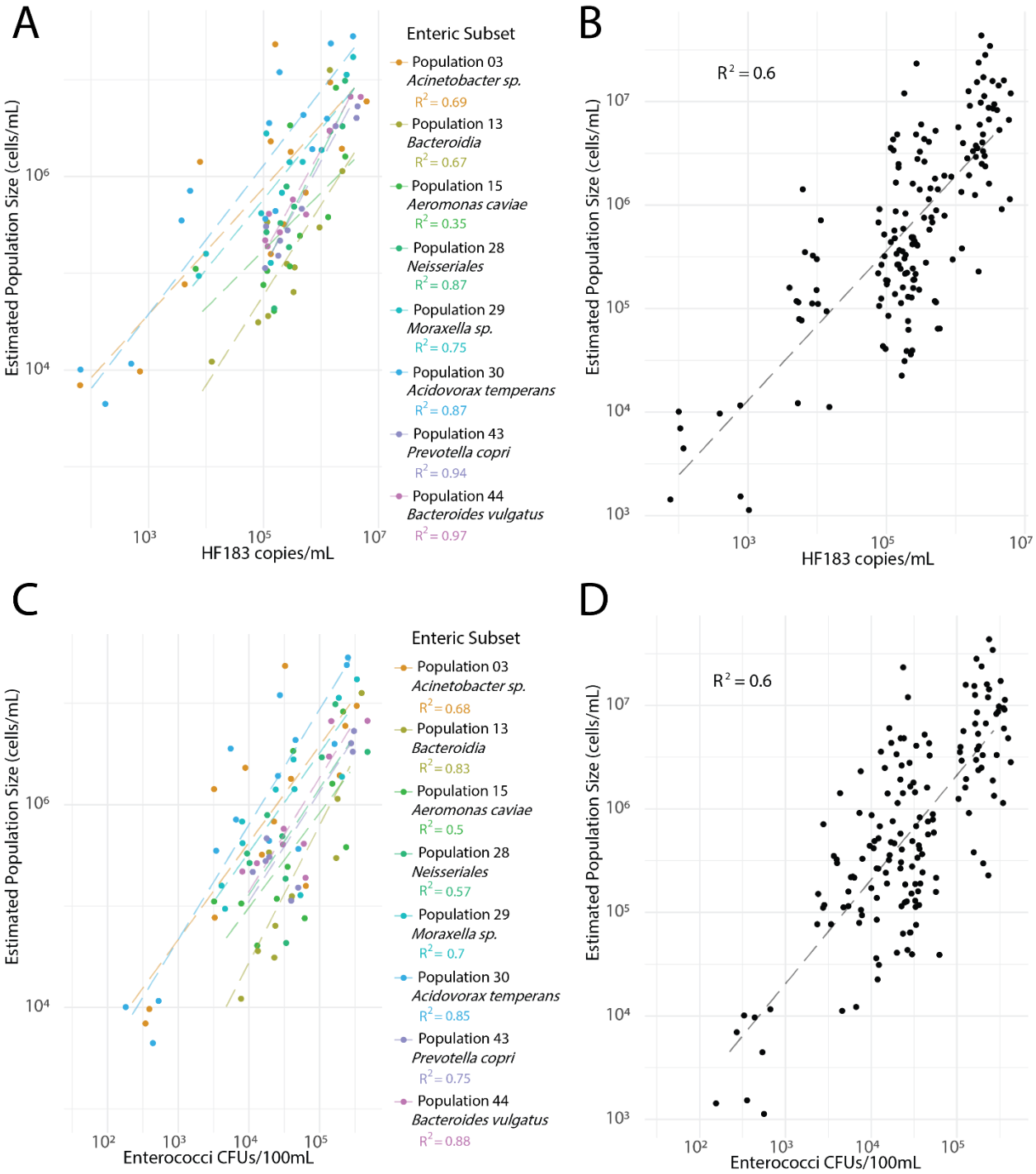


Figure 3. Log-log scatter plots of estimated population densities across inoculum and timeseries samples against HF183 and Enterococci concentrations. Lines of best fit are shown dashed with their associated coefficients. Panel A: HF183 copy number versus the concentration of sewage-associated populations 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).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

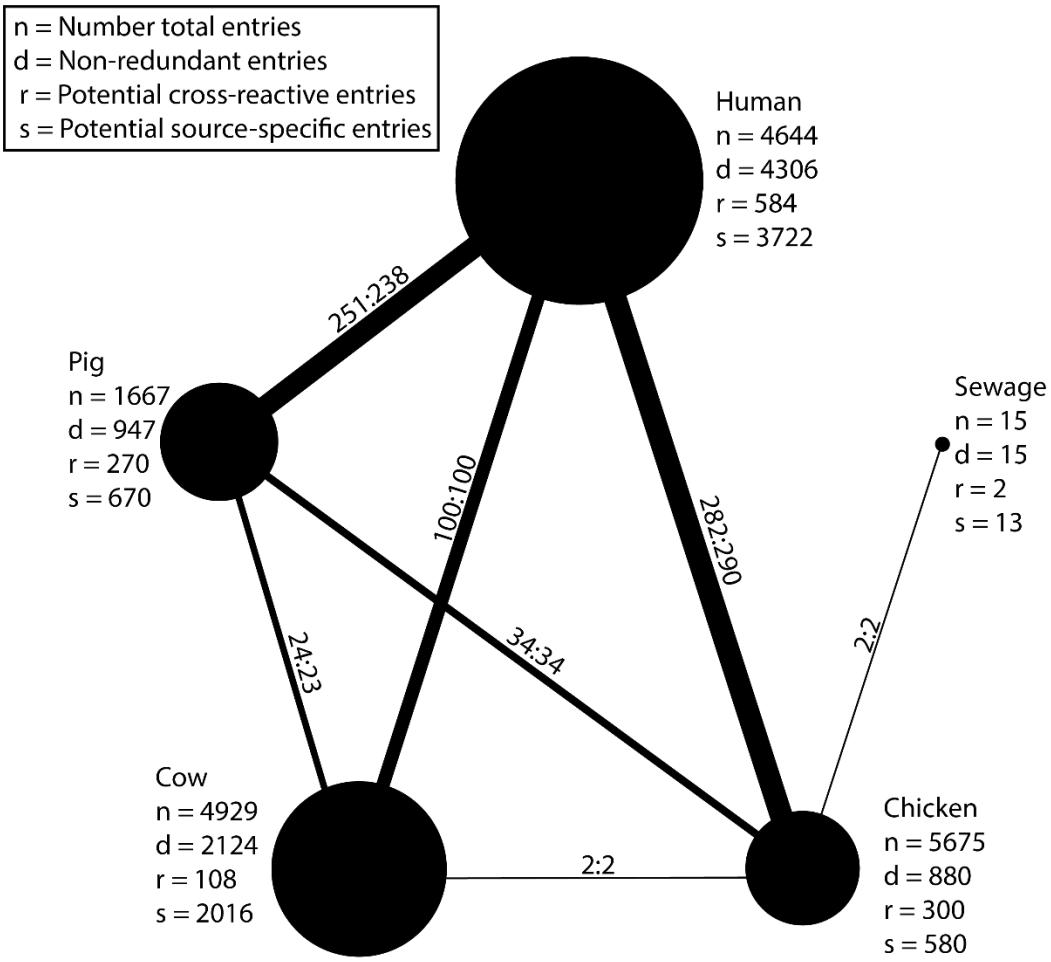
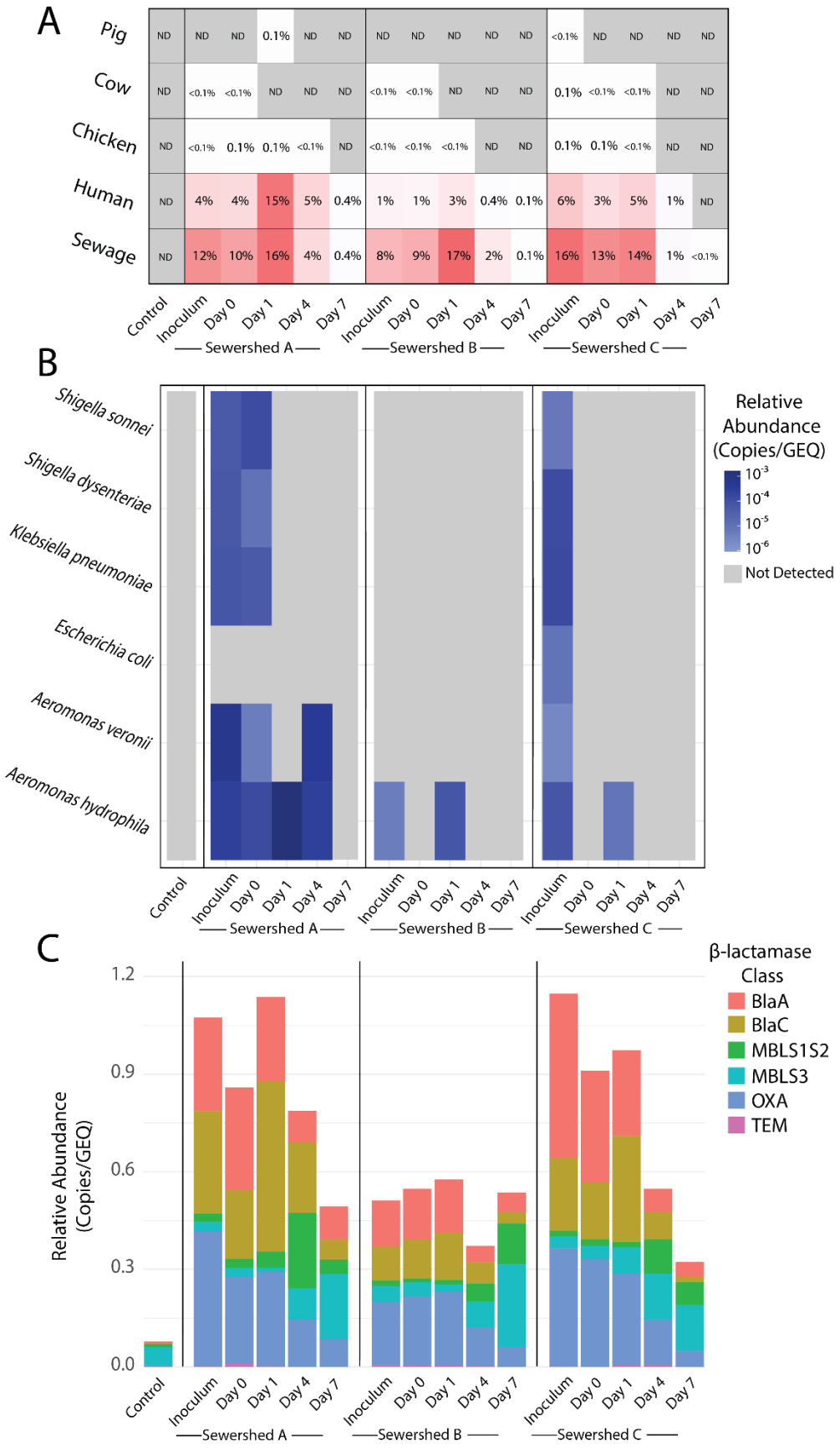


Figure 4. Overview of the curated genomic library for source attribution. Nodes represent sets of genomes recovered from public datasets of a given host microbiome, either "human", "pig", "cow", and "chicken". The "sewage" genomes shown here are those MAGs produced in this study. The radius of a node is proportional to the square root of the number of dereplicated genomes (d) remaining in the set following processing as described in the Methods section. Edges connecting nodes represent the amount of potentially cross-reactive genomes at the ANI ≥ 95% threshold level, and the line weights are drawn proportional to the square root of the largest number of these potentially cross-reactive matches (r). Ratios between nodes represent the number of matching genomes, with the value nearest a particular node representing the number of genomes from that dataset which matched across libraries.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.07.15.452399; this version posted July 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 5. Abundance patterns of Source Tracking 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 partitioning results based on reads mapped against MAGs curated for different fecal sources. Percentages represent estimates of the fraction of the prokaryotic population specifically belong to one of the fecal sources. **Panel B**: Virulence Factor (VF) gene abundance dynamics based on short reads mapping on experimentally verified VF reference nucleotide sequences (Figure key). **Panel B**: β-lactamase gene abundance dynamics across inoculum, timeseries and control metagenome based on Diamond --blastx 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.